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.

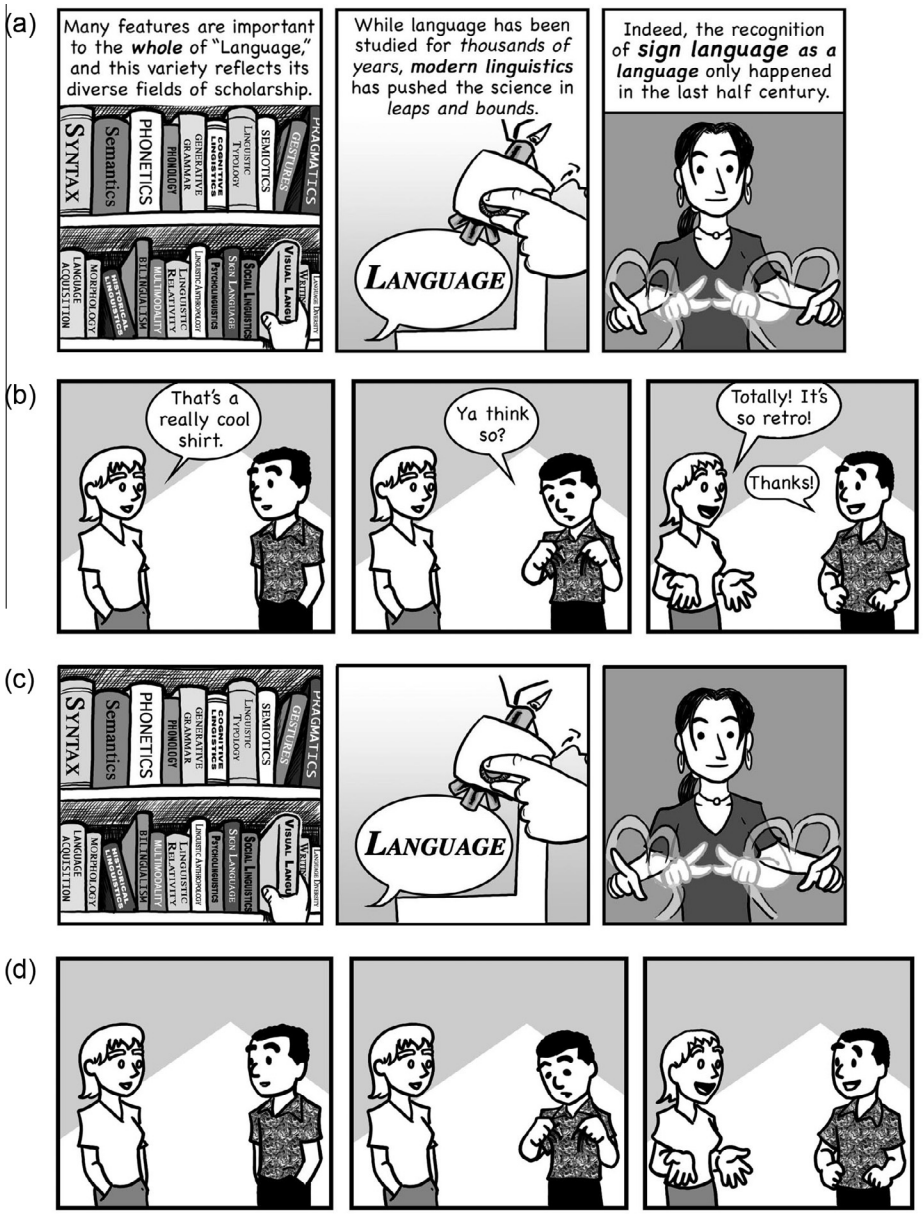

Fig. 1. Examples of multimodal relationships where text dominates in meaning. (a) Sequence where text controls all meaning. (b) Sequence where text controls all meaning, but visual sequence retains coherence. (c) Sequence in (a) with text omitted. (d) Sequence in (b) with text omitted.

panels play common functional roles. Panels 1 and 2 *together* set up the relationship between characters, and the next pair of panels starts the action. Third, because these panel pairs both show the same narrative state, yet each contain only a single character, we must infer their common spatial environment. We could depict panels 1/2 and 3/4 with a single panel containing both characters, rather than two separate panels, thus supporting that they group together. However, no such substitution could combine panels 2 and 3 into a single image, despite containing the same semantic contrast between characters. Fourth, visual narratives like this use recognizable sequencing patterns—"constructions" (e.g., Culicover & Jackendoff, 2005; Goldberg, 1995; Jackendoff, 2002)—which are stored in people's memories beyond panel-to-panel juxtapositions. This sequence uses an ***alternation*** between characters before converging, a pattern common in both comics and films (Bateman & Schmidt, 2012; Bordwell & Thompson, 1997; Cohn, 2013c). All of these phenomena warrant a system that extends beyond purely semantic panel relations.

In contrast, the theory of ***Visual Narrative Grammar*** (VNG) argues that sequential image understanding is guided by a narrative structure that assigns categorical roles to panels and orders them using hierarchic constituents, beyond the linear semantic relations between images (Cohn, 2003, 2013c; Cohn et al., 2014; Cohn, Paczynski, et al., 2012). This organization is analogous to the way that syntactic structure gives words grammatical categories that are ordered in a hierarchic constituent structure, but, importantly, VNG as a "grammar" organizes meaning at a discourse level. VNG therefore bears surface similarity to previous "grammatical" approaches to discourse (e.g., Clark, 1996; Hinds, 1976; Mandler & Johnson, 1977; Rumelhart, 1975), but differs from these precedents in that: (1) it is based on contemporary models of *construction grammar* (Culicover & Jackendoff, 2005; Jackendoff, 2002), not a phrase structure grammar (e.g., Chomsky, 1965); (2) it makes an unambiguous separation between equal contributions of the grammar and meaning (Cohn, 2013c; Cohn, Paczynski, et al., 2012); (3) it proposes additional modifiers (conjunction, Refiners,

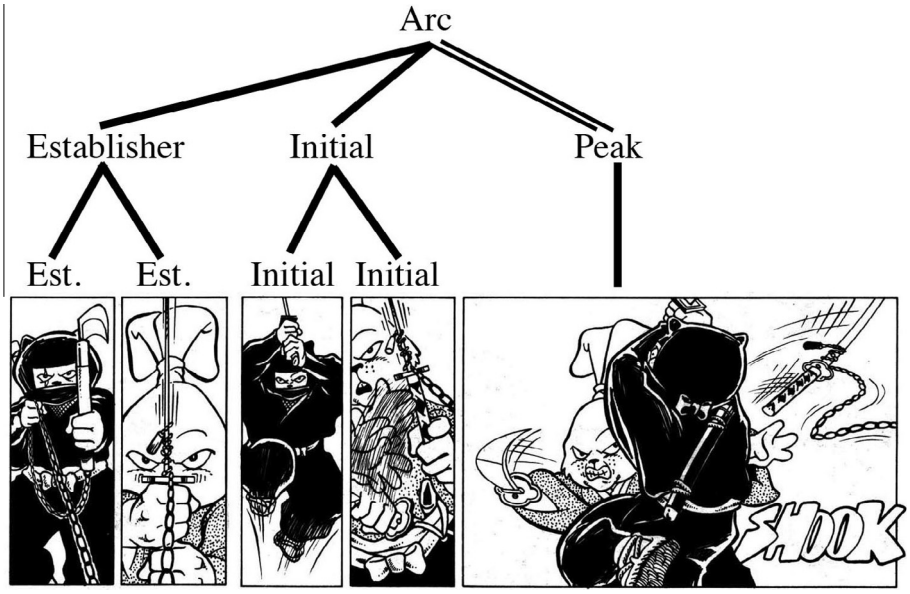

Fig. 2. Visual narrative. *Usagi Yojimbo* art © 1987 Stan Sakai.

etc.) that extend beyond basic canonical schemas (Cohn, 2013c, 2015); and (4) has been supported by behavioral and neurocognitive research showing similarities with the processing of syntactic structure at a sentence level (Cohn, 2014b; Cohn, Paczynski, et al., 2012; Cohn et al., 2014).

Let's briefly consider how VNG would analyze Fig. 2. The first two panels act as Establishers, which function to set up the interaction, here between the ninja and samurai. The second pair are Initials, which begin the characters' actions—the ninja jumps (panel 3) and the samurai draws his sword (panel 4). The sequence culminates in the Peak, where the samurai cuts down the ninja. At the maximal level, we see a canonical narrative arc—a set up (Establisher), initiation (Initial), and culmination (Peak). However, each component expands into its own constituent, modified using "conjunction," where the constituent comprises multiple types of the same category (Cohn, 2013c, 2015).

By dividing up the scene using conjunction, the overall spatial environment of the Establisher and the Initial must be inferred (mapped in VNG to the superordinate nodes). This provides one illustration of how the narrative functions to package semantic information. As stated, a single panel containing both ninja and samurai could suffice, instead of two panels.[2] This would convey the same *meaning*, while using a slightly different narrative *structure*. In this way, narrative structure functions to package semantic information at a discourse level, analogous to the way that syntactic structure functionally packages semantics at a sentence level (Culicover & Jackendoff, 2005; Jackendoff, 2002). Thus, narrative structure is a "macro-syntax" for ordering the *semantic* relationships often specified by models of discourse coherence (e.g., Asher & Lascarides, 2003; Halliday & Hasan, 1976; Kehler, 2002; Zwaan & Radvansky, 1998). This semantic information is incorporated into a *situation model* in memory for the comprehension of the sequence (Zwaan & Radvansky, 1998), while the narrative structure comprises the *textbase* by which that meaning is conveyed (van Dijk & Kintsch, 1983). Just as empirical work has borne out the separation of syntax and semantics in sentences (e.g., Marslen-Wilson & Tyler, 1980; Osterhout & Nicol, 1999; Van Petten & Kutas, 1991), experimentation has supported a separation of structure (narrative) and meaning (semantics) for visual narrative sequences (Cohn, Paczynski, et al., 2012). This functional viewpoint of "packaging information" is important, since it also characterizes the challenge of multimodal interactions: how do we understand (and produce) meaningful expressions expressed in text, images, or both?

VNG now allows us to differentiate the multimodal interactions in Fig. 1a and b. The image sequence in Fig. 1a does not use a narrative structure—nothing binds the images into a coherent structure. However, the sequence in Fig. 1b does use a narrative structure (albeit a very simple one). Previous experimental research has demonstrated that text-less incongruous sequences like Fig. 1c are comprehended significantly worse than those with a coherent structure (Cohn, Paczynski, et al., 2012; Cohn & Wittenberg, 2015; Gernsbacher, Varner, & Faust, 1990; Osaka, Yaoi, Minamoto, & Osaka, 2014). Yet, the inclusion of text in Fig. 1a appears to mitigate this incongruity. This means that the multimodal interaction somehow alters the comprehension of the visual sequence. It also implies that the multimodal interaction involved in Fig. 1a is manifestly different than that of Fig. 1b, which retains some congruity whether in a multimodal interaction or not.

If the felicity of these sequences varies based on multimodal interactions, it contrasts the assumptions that a uniform system governs both monomodal and multimodal visual narrative sequences (Bateman & Wildfeuer, 2014; McCloud, 1993; Painter et al., 2012). Furthermore, unlike approaches that focus on the *semantic* contributions of different modalities in visual narratives, the primary question here concerns the relative *structure* of different modalities in the expression of semantics. How might our expanded model of sequential image structure account for these multimodal differences?

2. The parallel architecture

On its own, verbal language uses three primary components: A ***modality*** (phonology), ***meaning*** (conceptual structure), and ***grammar*** (syntactic structure). While some combination of these parts occur in most all linguistic models, Jackendoff's (2002)

[2] Such a substitution acts as a diagnostic test for narrative conjunction: if a single panel with multiple characters can substitute for multiple panels each with a single character, then those panels are engaged in conjunction. In Fig. 2, this substitution would work for panels 1/2 and 3/4, showing they are conjoined, but not 2/3, which provides evidence against them forming a constituent (and for the major boundary falling between them).

parallel architecture argues for an equal contribution of each of these structures. The "visual language" of sequential images used in visual narratives also involves three similar components (Cohn, 2013b, 2013c): A modality (graphic structure), meaning (conceptual structure), and a grammar (narrative structure). Altogether, these components can unite into a single, holistic parallel architecture, as shown in Fig. 3a. It also includes a "bodily" structure[3] for the phonology of sign language and gesture, i.e., primarily for the hands and face (see also Table 1). Various tiers and substructures within these components are excluded for simplicity (ex. Phonology would consist of syllabic, segmental, and prosodic tiers), as are interfaces to necessary external structures (ex. perceptual systems, motor systems, etc.). Jackendoff's (2002) model can accommodate multimodality well because each structure is given equal weight, and therefore does not need to give primacy to one particular modality (such as verbal or visual) or to one particular structure (such as either semantics or grammar).

Note that this approach incorporates both visual-graphic and bodily communication into a single architecture along with verbal language, rather than viewing them as auxiliary systems. That is, the proposed model treats verbal, signed, and visual-graphic communication as ternary parts of a *single, holistic communicative system for conceptual expression*, out of which monomodal and multimodal expressions depend on which parts of the system are engaged. Different behaviors (speech, writing, signing, gesturing, drawing, etc.) arise as emergent interactions between these component structures. Some manifested monomodal expressions are depicted in Fig. 3b–h, and in Table 1, and we now turn to detailing these and other emergent behaviors.

2.1. Interfaces including grammars

Jackendoff's (2002) original parallel architecture for speech would use the Phonology–Syntax–Conceptual Structure interaction (Fig. 3g, Table 1g). As he described, lexical items permeate the interfaces between structures, existing as elements with cross-listed features from each component. For example, the "lexical entry" for the word *star* includes its Phonology (/star/), Syntax ([Noun; singular, count]) and semantics ([$_{\text{Object}}$ TYPE: STAR]). This notion of a lexical item can apply to all modalities (Cohn, 2013b), whether they are systematic (i.e., stored in memory)—i.e., words, constructions, signed words, gestural emblems, or conventionalized images—or non-systematic (i.e., not entrenched in memory)—i.e., novel words, gesticulations, images, etc. In addition, every component in the parallel architecture uses combinatorial principles, not just syntax (cf., Chomsky, 1965; Hauser, Chomsky, & Fitch, 2002). Thus, when referring to the "grammar" of a form throughout, its sense will be restricted to syntax and narrative, though all components may use combinatorial structure (elaborated below).

A similar mapping between structures occurs for sign language, which uses a Bodily–Syntax–Conceptual Structure interaction (Fig. 3h, Table 1h)—essentially the same interaction as speech with only a change in the modality. As per "visual language theory" (Cohn, 2013b), grammatical sequential images employ the Graphic–Narrative–Conceptual Structure interaction (Fig. 3e, Table 1i), since images have no need for Syntax (elaborated below), and rely solely on the narrative structures as a primary grammatical structure. In contrast, discourse-level verbal expressions may be mediated with a Syntax–Narrative interface, since meaningful expressions beyond the sentence level must also have an organizing system (i.e., here called "narrative"), though discourse-level semantics draws on the same principles as sentence-level Conceptual Structure (e.g., Jackendoff, 1990; Langacker, 2001). That is, narrative structure when applied to the verbal (or bodily) domain often orders information already packaged in a syntax, though in the visual domain it can bypass Syntax to serve as the primary grammar for sequential images (elaborated below). Thus, "Narrative" is included in the notation for both verbal (Table 1g) and signed languages (Table 1h), though it is omitted from the parallel architecture in Fig. 3g and h for simplicity (along with other interfaces).

If we link Graphic Structure to Phonology, it results in writing—the depiction of sounds graphically (to varying degrees)—as in Fig. 3f, Table 1l. A Graphic–Bodily Structure interface would thus be necessary for the writing of signed languages (e.g., Sutton, 1995), and also for drawings in which the physical articulation may be meaningful (Cohn, 2012; Green, 2014). A Graphic–Syntax mapping would also be necessary for constructions like *I* ♥ *New York*, where ♥ acts as a verb, but does not correspond to the alphabetic orthographic system (and indeed, has a phonology of "heart" that contrasts from the constructional phonology of "love"). Even more challenging is the creative use of this *Subject–ImageVerb–Object* construction, like *I* ♠ *poker*, where the explicit semantics (and phonology) may be less clear (though prevalent on t-shirts).

2.2. Interfaces excluding grammars

Meaningful expressions can also appear without a grammatical structure. Jackendoff (2002) argues that non-syntactic language appears in the verbal modality (Fig. 3c, Table 1d) in words like *abracadabra*, *kaboom!*, and *gadzooks!*, which cannot be inserted into sentences aside from quotations. These words are often related to exclamations (*ugh!, yikes!*) or to onomatopoeia (*bang!, pow!, kablamo!*). Non-syntactic expressions can also occur by pragmatic choice. For example, to the question *What type of wine would you like?* one might answer *White,* a non-syntactic utterance, or with *I'd like white wine*, a full sentence context (i.e., syntactic). In the parallel architecture, such syntactically-impoverished utterances involve no "hidden" syntax, yet demand inference of the unstated meanings. This type of non-syntactic expression emerges because of conversational choice and/or pragmatic constraints (e.g., Grice, 1967), rather than limitations imposed by intrinsic lexical features (as in exclamations or onomatopoeia). I will refer to these phenomena across modalities as ***atactic*** expressions (i.e., "non-ordered")—self-standing expressions that lack or forego grammatical structures (be it syntax or narrative).

Meaningful bodily motions use the Bodily–Conceptual linkage, without syntax (Fig. 3d, Table 1e), both for novel gesticulations, which are not entrenched in memory, or conventional gestural emblems (ex. thumbs up, ok, middle-finger, etc.), which are stored in long-term memory. Both types appear as isolates without combining in grammatical sequences. Note that the Phonology–Bodily interface accounts for the mapping of beat gestures to prosody (McNeill, 1992), and can interface with Syntax for coordination and substitution of gestures with sentence structures (Clark, 1996; Fricke, 2013), even though the gestures themselves do not have syntax (as in sign language). However, these interfaces start creeping into multimodal interactions, discussed later on.

Finally, individual images would use a Graphic–Conceptual interface with no Narrative (Fig. 3b, Table 1f). Internally-complex atactic examples would be individual drawings or paintings, while more simple examples would be the visual signs used to indicate bathrooms, computer icons, emoji, and many traditional single unit graphics throughout history (Dreyfuss, 1984; Liungman, 1991). Some have argued that complex individual images contain their own internal "syntax" (Engelhardt, 2007; Kress & van

[3] The sign language literature typically just refers to this as "phonology" (though once coined as "cherology"). I use "bodily structure" simply to avoid confusion between "verbal phonology" and "signed phonology," while staking no position on terminology.

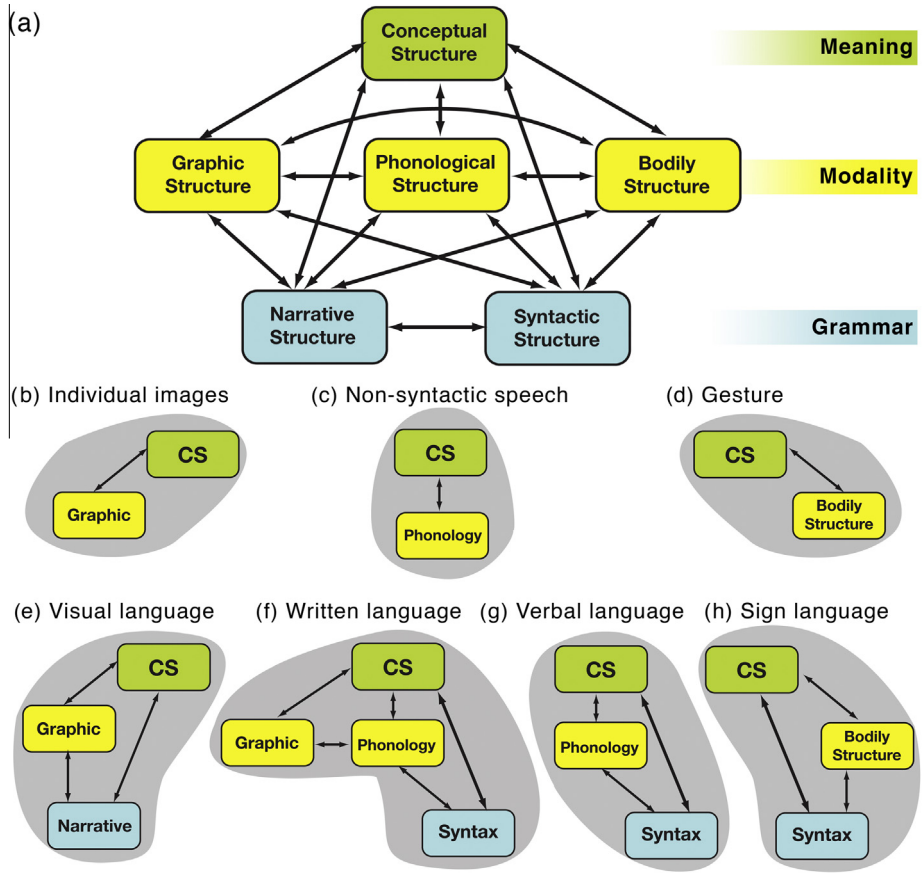

Fig. 3. The parallel architecture, expanded to allow for multimodal interactions (a), along with several types of monomodal expressions, as manifested within this model. These include behaviors without grammatical structures (b–d) and those with grammatical structures (e–h), simplified for clarity.

Leeuwen, 1996). However, such approaches typically describe *semantic* relations, and thus reflect combinatorial properties of conceptual structure and/or graphic structure of images (Cohn, 2012, 2013b; Marr, 1982; Willats, 1997; Willats, 2005), rather than a system that goes beyond aspects of perceptual processing.[4] Furthermore, individuals who do not learn the vocabulary of a drawing system will retain a "resilient" ability (Cohn, 2012; Goldin-Meadow, 2003b) to draw basic scenes, though exposure is necessary to develop the ability to draw coherent narrative sequences (Wilson, 2015; Wilson & Wilson, 1987). This is parallel with the "resilient" ability to convey meaning manually (as in gestures), despite not learning the vocabulary and grammar of sign language (Goldin-Meadow, 2003b).

While atactic expressions may lack a syntactic structure, they may still interface with narrative structures. A direct Phonology–Narrative–Conceptual Structure interaction appears in an utterance like *skid-crash-hospital* (Pinker, 1994), which conveys a narrative arc without syntax, as in Table 1j (see also Chwilla & Kolk, 2005). Similar non-syntactic discourse expressions no doubt appear in the bodily domain as well (Table 1k). Thus, narrative is not contingent upon syntax (in any modality), though they do interface.

2.2.1. Characteristics of grammars

An essential point for this discussion is: How can we tell whether a "grammar" is present in a modality? Because the parallel architecture posits combinatorial structure appearing in all components, when using the term "grammar" throughout, it refers to the system that functionally packages unitized conceptual information (i.e., syntax or narrative). Though a grammar should require a sequence of units, not all sequences of units use a fully complex grammar (Cohn, in preparation; Jackendoff & Wittenberg, 2014). For example, lists (be they verbal or visual) may not have grammatical structures, though they may bind semantically associated units in specific pragmatic contexts. Complex grammars—whether syntax or narrative—have several traits, including the formation of constituents, distributionally defined grammatical roles applied to both units and constituents, and a separation of these structures from semantics. Such structures enable the ability to distinguish structural ambiguities, to resolve distance dependencies, to create different surface presentations of the same meaning, and to create a single surface presentation capable of conveying multiple meanings.

In the absence of knowing a system's characteristics (i.e., one does not know how to recognize the structures), we can turn to diagnostic tests, which can also frame experimental designs. Syntax-specific diagnostics have long been used in linguistic research (e.g., Cheng & Corver, 2013), and similar tests have been outlined for VNG (Cohn, 2013c, 2014a). For example, because a grammar often creates order for a sequence, a "movement test" that reorders units or groups of units should render some rearrangements infelicitous (such as moving units across constituent boundaries) while others as acceptable (such as moving whole constituents or units in complementary distribution). Acceptability of all orders (i.e., all units can be moved into any order) should

[4] I leave it an open question whether narrative structures, as detailed in VNG, can appear within a single image. Future works will explore these issues in further detail.

Table 1
Various interactions between structures which manifest as types of monomodal (Autonomous) expressions. X = Presence of structure for an expression; ( ) = optionality.

Conceptual structure		Expressive modalities			Grammatical structures		Emergent structure
		Verbal-auditory	Visual-bodily	Visual-graphic	Syntax	Narrative	
Aconceptual expressions							
a		X					Non-meaningful verbalization
b			X				Non-meaningful gesticulation
c				X			Non-meaningful images
Atactic expressions							
d	X	X					Individual words
e	X		X				Individual gestures
f	X			X			Individual images
Linguistic expressions							
g	X	X			X	X	Verbal language
h	X		X		X	X	Sign language
i	X			X		X	Visual language
j	X	X				X	Non-syntactic verbal discourse
k	X		X			X	Non-syntactic bodily discourse
l	X	X		X	(X)	(X)	Written verbal language
m	X		X	X	(X)	(X)	Written sign language

suggest a non-grammatical sequence because it implies no sense of order.[5]

Returning to the example of lists: even though units may be grouped by *semantic associative* relations (like grouping categories of fruits, vegetables, and frozen goods on a shopping list), the units within those groups and the groups themselves could have free ordering. In the visual domain, experimental evidence has supported that sequences bound by semantic associative relations alone use different cognitive resources from those with a narrative grammar (Cohn, Paczynski, et al., 2012). Movement is one among several modality-general diagnostic tests, such as deletion or substitution of units, and can complement modality-specific tests to form a broad suite of diagnostics for assessing grammaticality and experimentally testing such distinctions.[6]

2.3. Interfaces excluding semantics

So far we have focused on the mappings that involve a modality's expression of meaning, with or without grammar, but we can also exclude conceptual structure. Jackendoff (2002) argues that some words use both Phonology and Syntax, but no semantics, like the *it* in *It's raining,* which serves as a dummy subject, or the *do* in do-support (*I **did**n't like him*), which primarily just carries Tense. A Phonology–Syntax link without conceptual structure also appears in "Jabberwocky" sentences (Canseco-Gonzalez et al., 1997; Münte, Matzke, & Johannes, 1997), though such sentences often use function words (prepositions, determiners) and these pseudo-words often rely on morphemes suggestive of meaning (such as Lewis Carroll's "slithy" as *lithe* plus *slimy*) and/or enabling syntactic relations. Whether visuals alone could use a Graphic–Narrative combination without meaning is debatable, but has been argued for "abstract comics" (Baetens, 2011; Molotiu, 2009).

Jackendoff (2002) also observes that Phonology alone—with no connection to meaning or Syntax—can yield "vocables" found in music like *sha-la-la-la-la,* or the nonsense refrains from nursery rhymes like *eenie-meenie-minie-moe* (Table 1a). Non-meaningful gesticulations would be similar for the bodily modality (Table 1b). Finally, graphic structure alone, without meaning or grammar, leads to "scribbles" or "abstract art" where lines, shapes, and/or colors play with the properties of the visual domain, but lack conceptual meaning (Table 1c).

As should be evident throughout, all possible mappings between component parts of the parallel architecture manifest as identifiable expressions. We might broadly characterize the modality and the grammars as means for *presenting* the meaning contained in conceptual structure. Across contexts, this meaning would be what is incorporated into a mental model for their comprehension (Brouwer, Fitz, & Hoeks, 2012; Zwaan & Radvansky, 1998), or would be what is drawn from for production. So far, all of these interactions within the parallel architecture result in monomodal or "Autonomous" expressions. Multimodal expressions thus manifest when more than one behavior emerges at the same time—i.e., between the forms detailed thus far—and such interactions will essentially be byproducts of the makeup of this architecture.

3. Multimodal interactions

Based on this architecture, this approach posits three major categories of manifested relations between modalities (see Table 2): *Autonomous*, *Dominant*, and *Assertive*. We will address each interaction individually, but it is worth remembering that, as McCloud observed, multimodal expression generally does not remain fixed to one interactive state, but shifts between interactions.

Multimodal interactions will be expressed in two ways: diagrammed with a parallel architecture (expanding on Fig. 3) or detailed through a table (expanding on Table 1 with Table 2). In diagrammatic form, only interacting structures will be depicted, with non-relevant structures omitted for simplicity. A solid line (———) will indicate the "***semantically dominant***" modality, *the one that carries more semantic weight in a multimodal expression*, while a dashed line (- - - -) will indicate a non-semantically dominant modality. Labeled groupings of cognitive structures will illustrate the emergent behavior (i.e., writing, visual language, etc.). Again, the semantically dominant modality will use a solid background while the secondary modality will use a dashed outline. For simplicity, these diagrams will focus on the primary structures involved, though this should not be taken to imply that

[5] Note that movement tasks alone may not work in all cases. For example, rearranging the units of languages with free word-order might "fail" a movement test because of the structural principles intrinsic to that culturally-specific type of grammar. Thus, movement diagnostics are not a guaranteed method of assessing grammar, and should be viewed as part of a suite of diagnostic tests.

[6] Given that diagnostic tests are a "quick and dirty" experiment using one's own intuitions (Culicover & Jackendoff, 2010), they require fluency in the language being analyzed. Fluency restrictions also extend to manipulations of visual language(s) (see Cohn, 2013b; Cohn, Paczynski, et al., 2012; Nakazawa, 2005; Wilson & Wilson, 1987). Testing for grammar in a system outside of one's fluency marks a longstanding issue in anthropological linguistics.

omitted structures or interfaces could not play a role in these interactions, so long as the defining relationships maintain.

Table 2 provides a second notation for these multimodal interactions. While Table 1 notated the presence or absence of structures in a monomodal expression, complications arise in the multimodal interactions in Table 2, where modalities may contribute to the overall semantic whole in different ways. Here, notation may differ based on semantic dominance ("D" indicates the dominant modality versus "X" for a non-dominant modality) or whether multiple modalities share equally in semantic dominance ("X" and "Y"). Optionality is marked with parentheses ( ), used often when the classification does not hinge on the presence or absence of a structure. For example, in Table 1l, "written language" does not depend on the presence or absence of syntax or narrative, though they could be involved. Similarly, the "verbal–visual" interactions in Table 2 could equally apply to written and spoken forms of verbal language, and thus includes parentheses in the "visual-graphic" column for writing. For simplicity, we primarily focus here on interactions with written language ("verbal-graphic") alone.

Both the figures and tables depict the "presence of a structure" as a binary distinction (i.e., present or not). In actuality, this contribution may be a percentage, especially for the contributions of conceptual structure in semantic dominance, given that a modality's contribution to meaning may have gradations. However, for the sake of simplicity in establishing the framework, simple binary oppositions will suffice, and exploring these more graded contributions can be left to future work.

3.1. Autonomy

The most basic expression is when a modality appears on its own, monomodally. For verbalized communication devoid of visual (or bodily) features, this includes radio or phone conversations, while in written form it includes bare written text, as in most books or some Internet chat interfaces. These productions can be considered ***Autonomous***—since they feature only a single modality, whether as fully grammatical or atactic expressions. Bare text (like this paragraph) or verbalized speech in absence of gestures would be *Verb(al)-Autonomous* linguistic production. Likewise, "silent" or "textless" comics would be *Vis(ual)-Autonomous*, since the visual language appears without any writing. Autonomous expression also extends to the bodily domain, when a sign language or gestures alone are used for communication. Within the parallel architecture, these Autonomous expressions would be depicted by the basic groupings for each individual modality in Fig. 3b–g.

Multimodal interactions balance different structures. Thus, testing multimodal examples against Autonomous expressions should allow us to assess the semantic contribution of each modality and/or the presence of structure in that modality alone. Throughout, deletion and/or substitution of a modality will be used as diagnostic tools to investigate and test the contributions to a multimodal interaction. These contrasts should also provide ways to manipulate sequences in empirical experimentation.

3.2. Dominance

3.2.1. Asymmetric visual/verbal Dominance

We now turn to multimodal interactions. Consider Fig. 4a, where an action scene is supplemented by onomatopoeia. These atactic sound effects enrich the visual sequence, mostly detailing the manner of motion for the actions (or lack thereof). Here, the text provides supplementary information to the essential, semantically dominant visuals. We can test this relationship by deleting either of the expressions: The visuals could communicate the primary message without the onomatopoeia, but the opposite would not be the case. The onomatopoeia *Ssshhh. . .–Swoosh–Fwap!* without images do not carry a comparable meaning as the multimodal interaction (or even comparable to just the visuals alone). Thus, the visuals "semantically dominate" the expression.

This multimodal interaction is of ***Dominance***, where a single Modality (visual-graphic) uses a Grammar (narrative) and controls the meaning (Semantic Dominance), while the other modality (verbal-graphic) plays a supportive role semantically, with no grammatical (syntactic) structures. These relationships are laid out in Table 2c, and Fig. 4b. The Conceptual–Graphic–Narrative linkages represent the visual sequence, while the Conceptual–Phonological link conveys the verbal "atactic" expressions—the non-syntactic onomatopoeia. *Vis-Dominance* can occur with onomatopoeia, as in Fig. 4a, but also technically when text belongs as part of the *intrinsic* properties of the image (Cohn, 2013a), such as the word "Stop" on a street-sign in the depicted "visual world."

Inversely, *Verb-Dominance* uses fully grammatical language with atactic images (Fig. 4c, Table 2b). These interactions appear prevalently, such as when text combines with purely illustrative images—as in newspapers, magazines, academic papers (like this one), or textbooks with the standard "See Fig. 1" index. Verb-Dominance also appears in everyday text-based conversation, where atactic emoticons or emoji combine with fully grammatical writing (Schnoebelen, 2012). A spoken Verb-Dominant interaction might occur when a lecturer speaks while drawing images on a blackboard or uses an all imagistic slideshow—essentially "panels" unfurling temporally on screen rather than physically on a page. The visual modality here would lack grammatical structures (i.e., a coherent and structured meaningful sequence), yet it would still offer semantic information to the overall communicative act. Just a reminder, this interaction between grammatical text and non-grammatical (likely single) images may have many types of surface semantic relationships—such as the images elaborating, supplementing, or illustrating the text (Martinec & Salway, 2005; McCloud, 1993; Painter et al., 2012; Royce, 2007)—but the Dominant interaction simply characterizes the underlying contributions of structure beneath such varied surface relations.

Consider also Fig. 5a. Though the images appear in sequence, no real narrative structures motivate the graphics, since no alteration or variation precipitates connections across panels. Indeed, each panel shows the exact same image. In some sense, the characters here are meaningful placeholders for the content of the text. This Verb-Dominant relationship allows the visuals to maintain a sequence but have no structure. The text conveys nearly all the meaning and narrative, while the images contextualize them (i.e., who is speaking). Support for this relationship comes from deleting the text (where nothing then would happen), or by substituting an entirely different dialogue with the same graphics, as in Fig. 5b (where most all the meaning then changes).

In some respects, this continuous repetition of a single panel is a bimodal "sleight of hand," disguising the Verb-Dominance over the visuals by making the contribution of the images *seem* larger than it actually is. The sequential representation merely accentuates the exchange by partitioning the dialogue into separate panel units, which act as graphic "units of attention" throughout a sequence (Cohn, 2007). However, we can collapse this whole situation into a single image (Fig. 5c), which conveys the same basic semantics as the sequence. This difference in the formatting and arrangement of how the sequence's *text* is broken up and presented. The direction of attention using panels has changed—the "pacing"—however, the overall *meaning* remains the same (i.e., who is speaking, in what order, and about what). This type of Verb-Dominance is often employed in the single-panel comics found in newspapers and magazines when single images provide a visual component to a textual message or joke.

Pushing the compositional alteration further, Verb-Dominance persists when arranging the visuals like an Internet chat

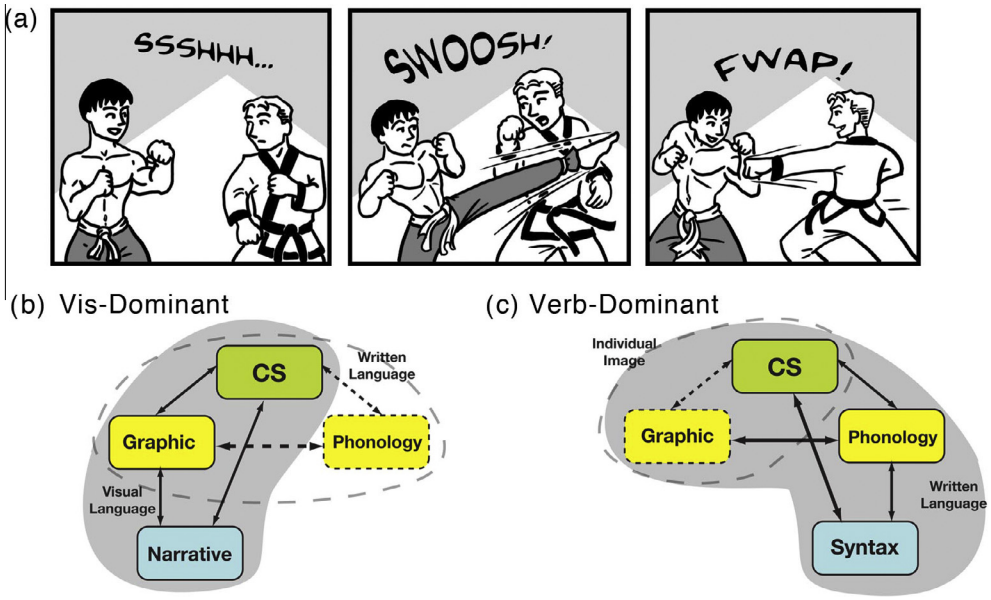

Fig. 4. (a) Vis-Dominant sequence where the visual sequence uses the narrative grammar, while supplemented by non-syntactic sound effects. (b) Vis-Dominant and (c) Verb-Dominant interactions within the parallel architecture (see also Table 2).

(Fig. 5d). Again, this sequence shows the same semantic components as in Fig. 5a–c, only changed in format. The visuals still provide a context for the text, and each image could just as well be replaced by "Woman" and "Man" to round out the "screenplay/Internet chat" format (though, this would result in Autonomy, not Dominance). These varying representations may affect a reader differently on some level of attention or emotion, and they certainly alter the pacing of how meaning is *presented* (McCloud, 1993; Wang, Hu, Hengeveld, & Rauterberg, 2014), and likely result in different processing in terms of memory, information load, etc. However, the primary propositional and compositional semantics garnered from the expressions remains the same throughout (i.e., Man, Woman, and the content of the dialogue), due to the consistency of Verb-Dominance over non-grammatical visuals.

Persistent semantics with different surface forms in sentences has long been debated in syntax–semantics relations (e.g., Chomsky, 1970; Goldberg, 1995; Jackendoff, 1990, 1991). While many manipulations to syntax also change meaning, others do not. For example, semantics do not seem to change in adjective/adverb pairs like *It's surprising that Bob came* versus *Surprisingly, Bob came*, or certain cases of the dative alternation (Jackendoff, 1990). Such manipulations reflect the functional basis of syntax to *package* meaningful information, as in the relation between semantics and narrative structure. Also like changes to syntax, not all alterations of framing visual sequences maintain the same semantics, as in this Verb-Dominant example, as will be evident in discussions of "absorption" below.

Beyond illustrating Verb-Dominance, this example points out an important aspect of the intrinsic properties of visual narrative. Though images appear in sequence, it *does not* mean that they necessarily exhibit narrative structures in the sense of VNG. Sequence alone does not make a visual narrative—such structures are motivated by structural relationships maintained between the visuals (Cohn, 2013c), and extend beyond semantic associative relationships alone (Cohn, Paczynski, et al., 2012). Recall Fig. 1a, where the relationship between images did not have coherent structure on its own. Here, the text motivates the meaning (and narrative), and the images provide an illustrative role, while retaining no narrative connection between each other (which becomes clear when omitting the text). This variety of Verb-Dominant interaction often emerges in non-fiction "comics" and many illustrated books (Cohn, 2013a).

3.2.2. Co-speech gesture dominance

As a parallel to text–image interactions, let's consider how gestures relate to speech. Though gestures can express meaning autonomously, they do not use a grammar (i.e., they are "atactic" with no syntax). Manual expressions that do have a combinatorial system go beyond gestures in full sign languages (note: gestures may also accompany sign language), though combinatorial qualities may emerge when speakers must rely on the manual modality alone (Goldin-Meadow, 2003b; Goldin-Meadow & Feldman, 1977; Goldin-Meadow, So, Ôzyûrek, & Mylander, 2008). Co-speech gestures often enhance, parallel, supplement, or enrich the meaning of the fully grammatical verbal language (Clark, 1996; Goldin-Meadow, 2003a; McNeill, 1992). These might include "concurrent gestures" that accompany speech, as in *I caught a huge fish!* with a wide stretch of the arms, or "component gestures" which substitute for words, as in *The fish was just <gesturing widely>!* (Clark, 1996). Such mappings require an interface of Bodily Structures with Syntax, despite the lack of syntactic structures in the manual modality itself. In both cases, the syntactic speech carries the primary semantic weight while the single gestures lend support. This relationship is also retained in "catchments," where recurring aspects of form and/or meaning gesturally repeat a persistent visuospatial imagery across a larger discourse (McNeill, 2000a; McNeill et al., 2001), somewhat comparable to the recurrence of non-grammatical sequential images across a Verb-Dominant text–image relationship, like Fig. 5a and b.

We can characterize co-speech gestures as another type of Verb-Dominant (Table 2a) or potentially Co-Dominant interaction (discussed below), where the verbal-auditory modality contains the primary meaning in a grammatical expression, accompanied by non-grammatical meaning in the visual-bodily modality (again, grammatical manual expression would be sign language). This also occurs in catchments, since these repeated gestures lack a grammatical structure (syntax or narrative), but do retain physical or semantic elements across discourse units, just as semantic elements may persist across non-grammatical image units in visual sequences (Cohn, Paczynski, et al., 2012; Saraceni, 2001).

Fig. 5. Verb-Dominant examples that maintain similar semantics. (a) Visual sequence with no grammar. (b) Visual sequence with no grammar and contextual dialogue. (c) Visual sequence collapsed into one image. (d) Chat styled dialogue.

Speech–gesture combinations create "composite signals" (Clark, 1996) that act as singular semantic units rather than parallel independent meanings, originating from a common "growth point" of conceptual information (McNeill, 1992). This notion can be extended to many text–image relations as well (Cohn, 2013a), where bundled text and image form a coherent message within a single panel that progresses to subsequent units (e.g., McCloud, 1993). In both cases, modalities may make a *semantic correspondence* in relating to common conceptual information (McNeill, 1992), but the visual-graphic domain does not use *temporal correspondence* as in co-speech gesture (i.e., association from occurring at the same time) unless it is produced in real time interactions (e.g., Green, 2014). Rather, text and image link through *spatial correspondence* in written form.

3.2.3. Co-Dominance

The multimodal interactions thus far have placed one modality over another. We might also posit a relationship that balances semantic dominance, but where at least one modality lacks grammatical structure, as in Fig. 6a. Here, the visuals are more than just illustrative—they must exist in order to convey the whole of the meaning. This ***Co-Dominant*** relationship distributes semantic dominance across modalities, yet one (or both) of those modalities still lacks grammatical structure (Table 2f), as in other Dominant interactions. This type of interaction often occurs in "single panel comics" (such as *Family Circus* or *The Far Side*) and many advertisements, which only use a single image (no grammar) combined with text, yet both image and text must be necessary for the meaning. Another example where both modalities share meaning, yet both lack grammar, might be the Internet meme where an image shows a mishap with only the word *FAIL* written on it. In all these cases, neither portion could be deleted and have the same message, while one or both lacks grammar.

A sequential case of Co-Dominance might occur in the *Sinfest* strip in Fig. 6c. Alone, the images have little relationship to each other—they are fairly incongruous if the text were omitted. Rather, they connect through a superordinate semantic field (Saraceni, 2001) defined by the text, with each image showing something that the character "can't quit": 1. The woman, 2. various vices, 3. books/knowledge, 4. himself. Both modalities are necessary for the meaning, but only the verbal form uses a grammar (syntax), here distributed across the panels (note also that each text–image pairing alone does not make a coherent unit—this is only achieved after reaching the end of the strip).

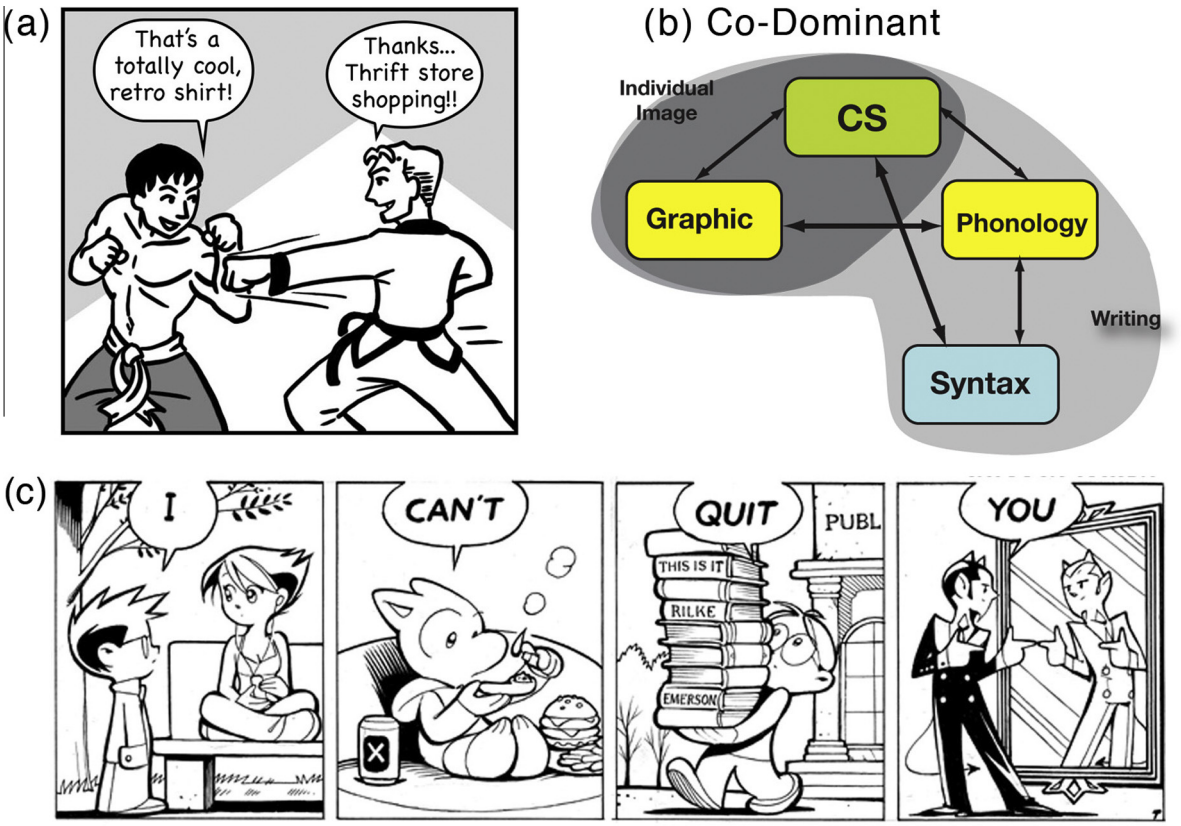

Fig. 6. (a) Co-Dominance where both modalities share equally in meaning, but only the text uses grammatical structures. (b) Co-Dominance in the parallel architecture. (c) Co-Dominance in an image sequence with no narrative grammar that retains equal semantic weight with text. *Sinfest* is © 2008 Tatsuya Ishida.

3.2.4. Substitution

It is worth exploring another manifestation of Dominant relationships, when one modality "substitutes" into the other. Non-grammatical forms of one modality might replace elements in another modality, as in component gestures, where manual expressions replace words in speech (*The fish was just <gesturing widely>!*). Here, meaning is expressed manually (possibly as a novel, unsystematic gesticulation) instead of phonologically (as a stored, systematic lexical item), while still maintaining the contextual syntactic structure (Fricke, 2013). Similar substitution occurs in *code-switching* between languages, particularly relevant for "bimodal bilinguals" who substitute sign language lexical items into spoken grammar, though less often speech into sign (Emmorey, Borinstein, Thompson, & Gollan, 2008). This contrasts from the more frequent *code-blends* where expressions occur simultaneously in both modalities—again more for sign alongside speech (Emmorey et al., 2008; Pyers & Emmorey, 2008)—like non-substitutive Dominant relations. The parallel architecture facilitates this relationship easily because a single grammatical structure can convey an overall conceptual meaning while presenting it in different modalities.

Component gestures may appear more natural as multimodal relationships than, for instance, the parallel Verb-Dominant situation substituting visual elements into the syntax of a sentence. However, these visual substitutions do appear in the aforementioned *I ♥ New York*-type constructions, in children's books, and increasingly in the use of emoticons in digital communication. Substitution also appears in the "rebus" in Fig. 7a. Though the visual elements lead to awkward felicity of the sentence, the syntactic structures of the verbal modality maintain throughout the sentence.

Naturalistic substitution of the verbal into the visual does occur in visual narratives, as in Fig. 7b where the onomatopoeia replaces the visual Peak panel of the gun firing. This Vis-Dominant substitution appears conventionally in comics, sometimes accompanying "action stars," which show a star-shaped "flash" blown up to the size of a full panel (Cohn, 2013b). Substituting the sound effect for a panel depicting the event works on a semantic level because of the metonymic link between the object/event (gun firing) and its emergent sound (*Bang!*). This also works at a narrative level: Just as action stars can substitute for a Peak (Cohn, 2013c; Cohn & Wittenberg, 2015)—as would be possible here—so too can onomatopoeia.[7] This would not work semantically if the referencing sound effect had been *Blargh!* or *Smooch!*, which have no relation to a gun firing, though it may maintain the narrative structure as a Peak. Such mismatches would presumably incur semantic processing costs, as found with incongruous substitutions of images into sentences (Ganis, Kutas, & Sereno, 1996; Nigam, Hoffman, & Simons, 1992). Nevertheless, verbal substitutions play a role in the otherwise visual narrative grammar (as a Peak), structurally analogous to the way that a component gesture or sign language code-switch might substitute for a spoken word as a noun or verb in the syntax of a spoken sentence (Clark, 1996; Emmorey et al., 2008). Again, the grammar in these substituted units is maintained since such constructions belong to their own independent structure (syntax or narrative) within the parallel architecture, but manifest in a non-dominant modality.

[7] Because action stars can substitute for Peaks, they can serve as one of the several diagnostic tests for Peak panels in a sequence (Cohn, 2013c; Cohn & Wittenberg, 2015). However, it is unclear if all onomatopoeia could be used as an additional "substitution diagnostic" in this way. My intuition would be to say that this would be overly broad, but this is an empirical question.

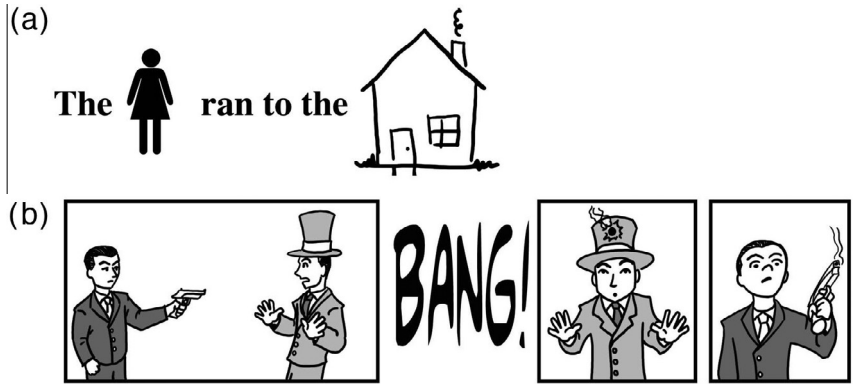

Fig. 7. Sequences where one modality is substituted into the "grammar" of another modality. (a) Substitution of images into a verbal sentence. (b) Substitution of a word into a visual sequence.

3.3. Assertion

3.3.1. Asymmetric visual/verbal Assertion

In Dominant relations, only one modality used grammatical structures, but another interactive type balances grammar in *multiple* modalities. To illustrate, let's combine the dialogue from the Verb-Dominant examples with visuals that have distinct sequential features. In Figs. 8a and 1b, slight changes across the panels allow the images to convey their own narrative structure (a fairly simple Establisher–Initial–Peak sequence). Yet, the images still do not motivate the meaning, as evident when omitting the text (Figs. 8b and 1d). Sequential visuals now supplement the dominant text to give more expressive meaning beyond just placeholders for the dialogue (as in Fig. 5).

This type of interaction is characterized by grammatical structure appearing in both modalities (syntax, narrative) while the meaning remains controlled by only one—an ***Assertive*** interaction (Fig. 9, Table 2d). As stated, deletion of a modality can test for the semantic dominance of one modality over another. Figs. 8b and 1d omit the text from the *Verb-Assertive* (Table 2d) examples in Figs. 8a and 1b, respectively. These Vis-Autonomous sequences convey little of the sense of the multimodal meaning, with no indication of marriage or friends in Fig. 8b, and no mention of why the shirt is important in Fig. 1d. Visually, the *only* difference between these two sequences is the single figure of the man in the second panel—though when combined with text, they seem like very different exchanges. This can largely be attributed to the Assertion of the verbal over the visuals, with the words guiding the semantics.

Once again, because the text motivates the meaning, these panels can collapse into a single image (Fig. 8c). In this case, the posture of the characters at least lends gestural vigor beyond Fig. 5c, yet the overall semantics of both remain quite similar. Both images become Verb-Dominant, but the point of origin is different. In Fig. 5c, the resulting single image had collapsed across panels without narrative grammar, while in Fig. 8c collapsing across panels loses the narrative grammar (changing from a Verb-Assertive to Verb-Dominant relationship). Like before, the pacing here changes, and though a "narrative" does occur, it is not within the *visuals*.

In some sense, the semantics of the visuals in Fig. 8c have been "absorbed" by the text, resulting in the loss of its narrative structures (unlike the more invariant semantics with the Verb-Dominant framing in Fig. 5). ***Absorption*** is the idea that structure in one modality may be lost by conveying that same information in another modality. In these cases, the same basic meaning is conveyed, but the structure changes because of its form. While this process could potentially go in either direction, apparent absorption of visual meaning into the verbal form occurs quite frequently in visual narratives. For example, captions reading *Meanwhile, across town...* or *In the warehouse...* while showing people within a location could replace an image of that place, i.e., a warehouse, prior to the reset of the sequence (Cohn, 2013b). In these cases, expressing the location in text rather than images may allow for a more economical message (less panels), but may lead the content of each domain to convey different aspects of meaning, and thereby require the integration of novel information across multiple sources (e.g., Fauconnier & Turner, 2002).

A related process occurs in the substitutive example in Fig. 7b, since the Peak is represented by text instead of an image. Other examples of absorption are discussed at length by Stainbrook (2003, 2015) who describes how surface cohesive relations between images and text contribute toward a holistic mental model of a multimodal discourse (e.g., van Dijk & Kintsch, 1983), and comparable notions are echoed in Painter et al.'s (2012) discussion of text or images "committing" different amounts to a global meaning. From a production perspective, absorption merely reflects possibilities for distributing meaning into different modalities from a common conceptual structure. This is consistent with McNeill's (1992) notion of a growth point as the conceptual origin for both speech and gesture, only here the structure (i.e., grammar) may change depending on which modality expresses that meaning. Such cases may provide a good starting place for exploring the more fine-grained interactions between modalities' structures in terms of grammar–semantics interactions.

The reverse of this interaction, *Vis-Assertion* (Table 2e), occurs when the visuals have a narrative structure and guide the meaning of a sequence, beyond that in the text. In Fig. 10a, the dialogue is mostly redundant with the visuals, enriching the primary meaning in the images by conveying the overall attitude and tone of the fighters' relationship. Yet, the text contributes minimally to the overall gist of the sequence. Without the dialogue (Fig. 10b), the visuals still convey the information necessary to understand the sequence. If the visuals were deleted, the sequence would make far less sense.

3.3.2. Co-Assertion

The most complicated interactions arise when both modalities use grammatical structures and both modalities semantically contribute in non-negligible ways. Consider Fig. 11a, where we now combine the components from each of the previous Assertive examples.

In this *Co-Assertive* interaction (Fig. 12, Table 2g), both the visual and verbal modalities are necessary to form a semantic

Fig. 8. Example sequence of Verb-Assertion (a) and contrasting manipulations (b,c). (a) Verb-Assertion with no meaningful connection to the images. (b) Sequence omitting the text of (a). (c) Collapsing of image sequence into single image.

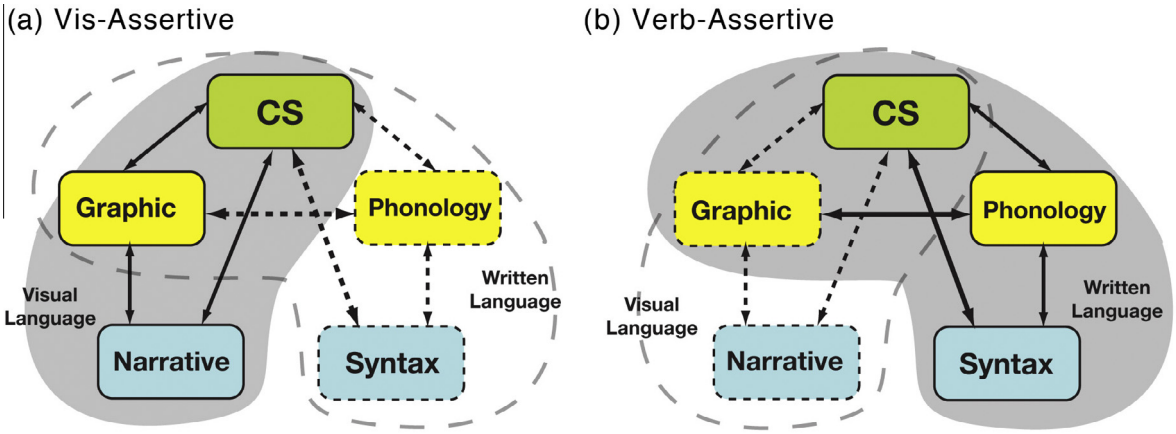

Fig. 9. Assertive relationships of text and image.

whole greater than either of their parts. The visuals only contain semantics related to sparring, while the text discusses fashion. Together, this interaction informs us about the relationship between the characters and the nature of the combative exchange (their banter in Fig. 11a implies less opposition as the dialogue in Fig. 10a). This "emergent" semantics of the multimodal combination could not be found within the depicted meaning of each individual modality. Such emergent meaning likely involves

Fig. 10. Vis-Assertive sequence (a) and comparable Vis-Autonomous sequence resulting from the omission of text (b).

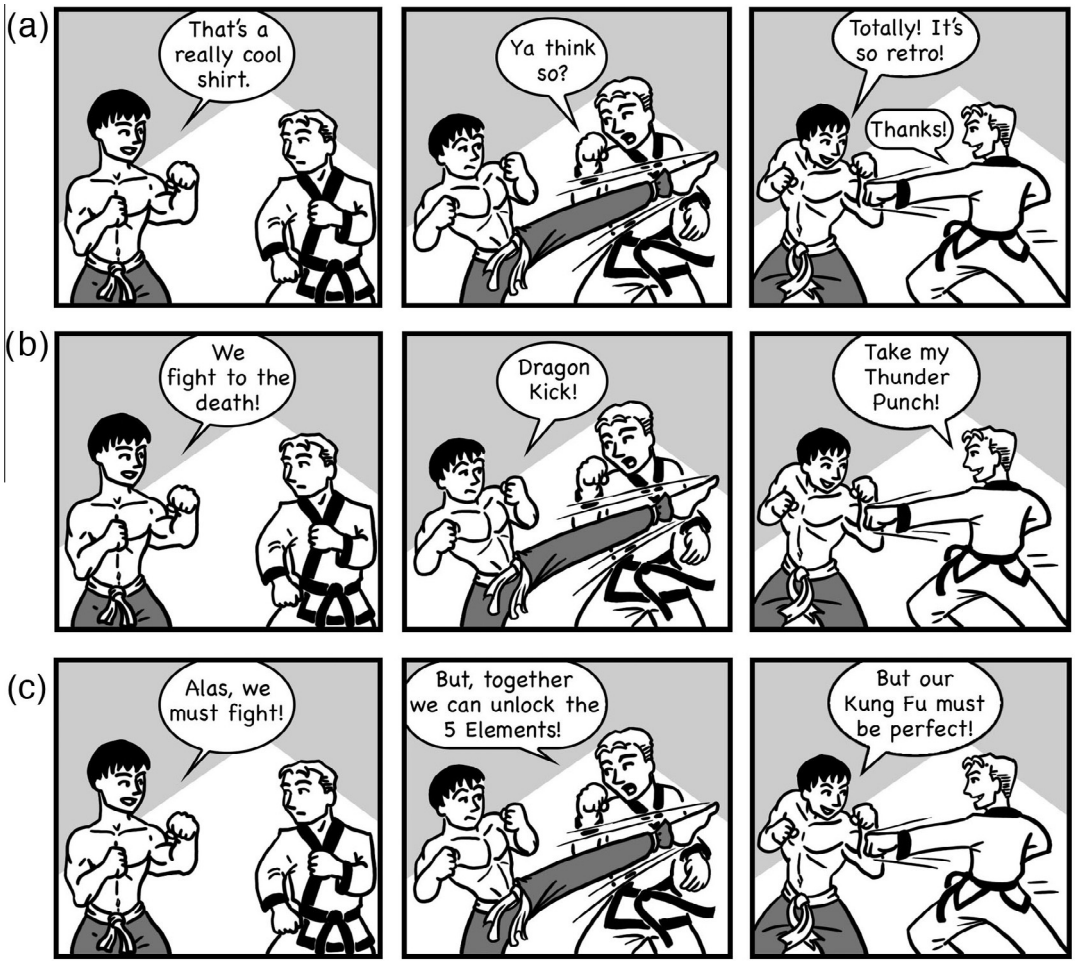

Fig. 11. Co-Assertive sequences with varying surface text–image interactions. (a) Parallel text–image surface structure. (b) Additive surface structure. (c) Interdependent surface structure.

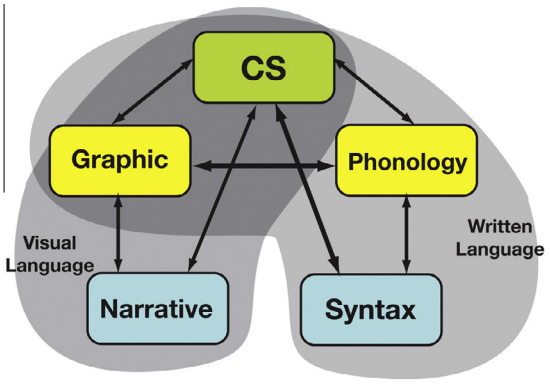

Fig. 12. Co-Assertive interaction between text and image.

"conceptual blends" (Fauconnier & Turner, 2002) that arise from negotiating the semantics in both the visual and verbal domains to create an additional meaning unexpressed overtly in either. Such blends are common in multimodal interactions both within and outside of visual narratives (Cohn, 2010a; Forceville, 2015; Forceville & Urios-Aparisi, 2009).

Co-Assertion allows for each modality to equally contribute its own conceptual representation to a broader composite whole. This can either mean that (1) a balance is struck between the two forms, (2) that they intersect semantically, or (3) that they each convey their own meaning independently, yet coalesce to form an emergent meaning. McCloud describes these dually semantic relationships under his Parallel (Fig. 11a), Duo-Specific, Additive (Fig. 11b), and Interdependent (Fig. 11c) text–image relationships. These categories thus provide surface semantic distinctions for the underlying Co-Assertive interaction between structure and meaning.

4. The balance of structure and meaning

This proposed model of multimodality in the parallel architecture characterizes the interaction of the semantic dominance of a modality with the presence of grammatical structures (syntax, narrative). These underlying structures form the basis of multimodal interactions, which on the surface might be characterized by the types of surface semantic relationships described throughout the literature on multimodality (e.g., Forceville & Urios-Aparisi, 2009; Horn, 1998; Kress, 2009; Martinec & Salway, 2005; McCloud, 1993; Painter et al., 2012; Royce, 2007). In addition, although this work has emphasized bimodal interactions, it should also apply to trimodal interactions involving images, speech, and bodily expressions. Multimodal interactions would be most complex when all three modalities use grammatical structures, as in the storytelling practices of Aboriginals in Central Australia, who use spoken discourse along with a visual language drawn in the sand, and an auxiliary sign language (Green, 2014; Wilkins, 1997/2015). While such interactions would become noticeably more complex, they are hypothesized as retaining the same types of interactions described throughout.

A prediction of this framework is that semantic dominance correlates with a greater presence of grammatical structure. We might thus hypothesize that *as a modality is relied upon more to convey meaning, its structural complexity increases*. Such complexity may be particularly true for modalities other than verbal language, which in practice seems to be relied upon more than visual-graphic or visual-bodily expressions. We saw this in reverse in the "absorption" of sequential images into a single image (in Figs. 5c and 8c). Because the verbal was able to "absorb" the meaning of the whole, the narrative structures of the visuals were lost.

Some research supports the possibility of this tradeoff between semantic weight and grammatical complexity. For example, gestures have minimal combinatorial structure, and are usually subservient to the semantically and grammatical dominant speech (McNeill, 1992). However, when non-signing adults are forced to rely solely on manual communication, conventional and consistent patterns arise spontaneously that becomes closer to the structures found in sign languages (Goldin-Meadow, 2006; Goldin-Meadow, McNeill, & Singleton, 1996). Similar complexity arises when deaf children are born to hearing parents, and must create "homesign" systems of manual expression, which again follow consistent combinatorial patterns (Goldin-Meadow, 2003b; Goldin-Meadow & Feldman, 1977). Of course, sign languages alone are far more complex than any gestures that accompany speech or appear in isolation. Increasing complexity in the absence of speech also comes from the aforementioned Central Australian sand narratives. Wilkins (1997/2015) noted that the visual modality of sand narratives increase in complexity when spoken and sign language are used less.

Multimodal support for this tradeoff also occurs in bimodal bilingualism between signed and spoken languages. Here, one modality's syntax typically remains constant while either

Table 2
Various interactions between structures which manifest as types of multimodal interactions. X = Presence of structure for an expression; D = Presence of structure that is semantically dominant; X,Y = Presence of structures that share semantic dominance; ( ) = optionality.

	Conceptual structure	Expressive modalities		Grammatical structures			Emergent interaction
		Verbal-auditory	Visual-bodily	Visual-graphic	Syntax	Narrative	
a	D	D			D	(D)	*Verbal-Dominant*
	X		X				Co-speech gesture
b	D	D		(D)	D	(D)	*Verbal-Dominant*
	X			X			Verbal over visual
c	D			D		D	*Visual-Dominant*
	X	X		X			Visual over verbal
d	D	D		(D)	D	(D)	*Verbal-Assertive*
	X			X		X	Verbal over visual
e	D			D		D	*Visual-Assertive*
	X	X		(X)	X	(X)	Visual over verbal
f	X	X		(X)	(X)		*Co-Dominant*
	Y			(Y)			Equal semantics, visual no grammar
g	X	X		(X)	X	X	*Co-Assertive*
	Y			Y		Y	Equal semantics, both with grammar

substituting lexical items from another modality (code-switching) or expressing them simultaneously (code-blending) (Emmorey et al., 2008)—i.e., a Dominant relation between sign and speech. When both modalities retain grammar (Assertion) the non-dominant modality's syntax often alters based on the structures of the dominant one, such as when spoken English takes on characteristics of the dominant American Sign Language (Emmorey et al., 2008). This model makes a similar prediction that the *narrative* grammar would be preferenced in the semantically dominant modality, but could switch as meaning shifts more weight to a different modality.

5. Future directions

This approach has proposed three broad categories of expressions arising from different interactions of the underlying parallel architecture's contributions of grammar and meaning:

(1) *Autonomous* – where only one modality is present.
(2) *Dominant* – where multiple modalities are present, but only one uses a grammar.
(3) *Assertive* – where multiple modalities are present, and all use a grammar.

For each of these categories, semantic dominance can be asymmetrically weighted to one modality (Verbal-, Visual-), typically the one with grammar, or shared across multiple modalities (Co-). While broad, these categories can provide a descriptive tool for characterizing the nature of the interactions between structures. As such, this approach can be used, among other things, for analysis of corpora, for experimental design on multimodal comprehension, and for describing the general properties of multimodal interactions in contexts such as education.

5.1. Corpus analyses

This framework offers an analytical tool for studying the properties of multimodal interactions within different contexts, particularly visual narratives. With a growing focus on corpus analysis of the visual languages used in comics (e.g., Cohn, 2011; Cohn, Taylor-Weiner, & Grossman, 2012; Forceville, 2011; Forceville, Veale, & Feyaerts, 2010; Guérin et al., 2013), this framework can be used to investigate multimodal interactions to address questions like: Do various genres or cultures use multimodal interactions in different ways? Have multimodal interactions changed over time? Do different multimodal interactions correlate with growing complexities in structures?

Coders analyzing various works might follow a decision tree as presented in Fig. 13. First, one asks, *Is a modality present?* If only one is present, it is an Autonomous relationship (which could be broken down into grammatical and atactic types, if desired). If multiple are present, then ask about semantic dominance: *Does one control the meaning?* The deciding factor here is whether one modality can retain the gist of the multimodal utterance in the absence of the other modality. This can be answered again as either "only one" or "multiple," and is tested using a deletion test (as demonstrated throughout). In both cases, the final disambiguating question relates to grammar: *Do both modalities have grammar?* This is determined by assessing the relative relations between units in each of their modalities (see also modality-specific diagnostic tests). For cases where one modality is semantically dominant, the presence of grammar in both yields an Assertive relationship while the lack of grammar in one modality yields a Dominant relationship. The particular nature of these interactions (Visual-, Verbal-) is determined by which modality may or may not be semantically dominant. For cases where semantic dominance is balanced, the presence of grammar in both yields a Co-Assertive relationship, where the lack of grammar in one (or both) leads to Co-Dominance. Note also that the last two questions are reversible: the order of determining grammar and semantic dominance does not matter.

Let's apply this workflow to Fig. 1a and b as an example. First, we ask: *how many modalities are present?* Both Fig. 1a and b use text and image, and thus we can rule out Autonomy and move to the next question about semantic dominance: *Does one modality control meaning?* Here, a deletion test (as in Fig. 1c and d) shows that the text is more semantically dominant since most of the overall meaning is lost without it. This gives us a "Verbal-" prefix moving into the final question: *Do both modalities have grammar?* Since we know that the text uses a grammar, we can now look to the relations between images. In Fig. 1a, the images have few connections besides a general semantic field about language, rendering a "no" answer and thus a Verb-Dominant interaction. Fig. 1b uses at least some contiguous connections between images, rendering a "yes" answer and thus a Verb-Assertive interaction.

We analyzed the reliability of this workflow using a sample drawn from an ongoing corpus study of American superhero comics from 1940 through 2014 (16 comics, 1744 total panels analyzed). Two independent coders categorized one panel at a time while progressing through each book using this workflow. We found an inter-rater reliability of ICC(2,2) = .930 using Cronbach's alpha (Shrout & Fleiss, 1979) for participants' classifications of panels into the outlined types of multimodal interactions, suggesting that this methodology can provide a reasonably effective method of coding multimodal interactions.

These proposed distinctions can also be important for coding other facets of visual communication, beyond the multimodal categories themselves. Because this approach argues that narrative structure may contribute differently depending on the multimodal relationship, corpus analysis of narrative structure and sequential image semantics may benefit from also recording multimodal interactions. For example, throughout a multimodal visual narrative, all sequential images may not use narrative grammar, as such structures may be diminished or transferred to the verbal domain in Verb-Dominant or Verb-Assertive interactions. Such observations also extend to theoretical and experimental investigation of the narrative structures of visual sequences, which would be maximal for Vis-Autonomous and Vis-Dominant sequences, where semantics of the verbal form would be least impactful. On the whole though, this approach makes the case that monomodal and multimodal expressions—particularly visual narrative sequences—cannot be treated uniformly in their analyses (Bateman & Wildfeuer, 2014; McCloud, 1993; Painter et al., 2012).

5.2. Psychological experimentation

Experimentation can also benefit from this framework. Given that theoretical architectures for wordless visual narratives have thus far been successful in framing experimental designs for sequential image comprehension (Cohn, 2014b; Cohn, Paczynski, et al., 2012; Cohn & Wittenberg, 2015; Cohn et al., 2014), we can use this framework to design experiments testing the processing and comprehension of multimodal interactions (and to test the veracity of this model itself). This is directly parallel to the decades of examples of psycholinguistic research that have directly used diagnostics and observations from theoretical linguistics to inspire experimental designs.

The diagnostic tests used throughout (e.g., deletion, framing alterations, substitution, etc.) offer an explicit way to design experimental stimuli that is not provided in other approaches to text–image multimodality. For example, certain Verb-Dominant

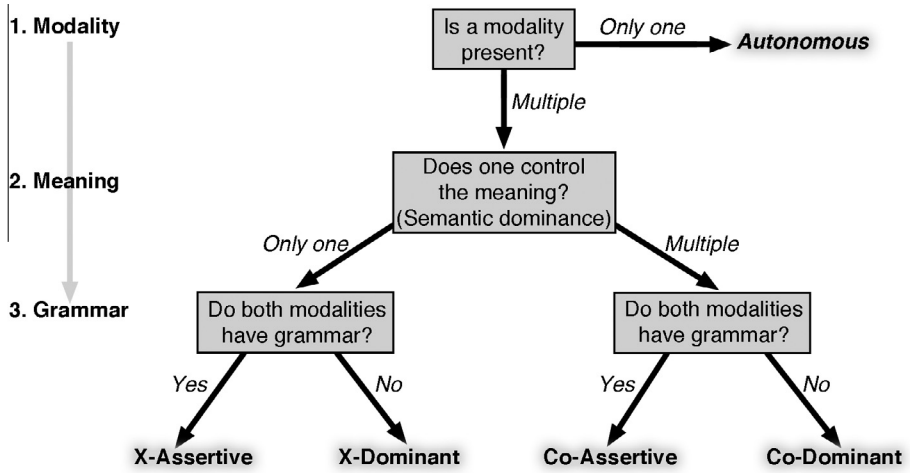

Fig. 13. Step-by-step method for analysis of multimodal interactions.

sequences were hypothesized to be easier to understand with text than without text, as in Fig. 1a, which, as an Autonomous sequence (Fig. 1c), should be incongruous. We would expect easier processing (faster reading or response times, attenuated ERP responses, etc.) for Autonomous sequences that use a narrative grammar than do not use one, as found in prior research (Cohn, Paczynski, et al., 2012; Cohn & Wittenberg, 2015; Gernsbacher et al., 1990; Osaka et al., 2014), but perhaps for this difference to be mitigated in a multimodal context. Such results would support the idea that including text can change the processing of visual sequences, and would contrast predictions that a uniform system governs both monomodal and multimodal visual narrative sequences (Bateman & Wildfeuer, 2014; McCloud, 1993; Painter et al., 2012).

In general, these distinctions should help frame all experiments on multimodal relations, whether about structure, semantics, or their application in other domains. For example, an experiment that compares surface semantic connections between modalities using both Verb-Dominant and Verb-Assertive stimuli may introduce an additional confound because of the relative contributions of narrative structure across images. In contrast, a study investigating the *visual* narrative structures may be confounded by the inclusion of Verb-Dominant and Verb-Assertive sequences, where the visuals may have a diminished structure. Characterizing such potential pitfalls allows for more controlled experimental designs, no matter the area under analysis.

This framework can also potentially benefit several existing lines of research, particularly multimodal research on discourse. For example, particular interest has been given to studying the segmentation of events in visual narratives (Magliano & Zacks, 2011; Zacks, 2014)—i.e., the structure and characteristics of breaks between events—and this framework allows a way to show the dynamic contributions of both visual and verbal information on such processes. This model suggests that visual narratives may differ in their segmentation of events depending on the nature of the multimodal interaction (for example, segmentation should be diminished or absent in Verb-Dominant examples), and can aid in framing such research. In addition, this framework can benefit the focus on inference in visual narratives (Cohn & Kutas, 2015; Cohn & Wittenberg, 2015; Magliano, Dijkstra, & Zwaan, 1996; McCloud, 1993), by characterizing how inference generation is supported by individual and multiple modalities. For instance, inferred information in a Vis-Autonomous sequence may be provided overtly with text in a Vis-Assertive sequence without changing the visual sequence at all. Such examples offer a way to investigate issues of inference across modalities, as well as to start addressing the "absorption" of structure between modalities.

More basic questions are also posed by this model, including: what are the costs or benefits of meaning coming from different sources, and does it differ depending on the presence of grammar? What are the costs or benefits of emergent multimodal semantics? Given that visual narrative and verbal syntax have been argued to draw on common cognitive resources (Cohn et al., 2014), would "grammar" appearing concurrently in various domains lead to costs or facilitation of combinatorial processing, and under what conditions?

5.3. Learning and literacy

Finally, focus on multimodality has been especially apparent in educational contexts. A growing literature has shown that formatting educational material with the visual language of "comics" has proven to be an effective learning tool (Kogo & Kogo, 1998; Nakazawa, 2005, 2015; Nalu & Bliss, 2011; Short, Randolph-Seng, & McKenny, 2013), and "illustrated children's books" have long been accepted as useful tools for learning to read (Meek, 1988; Painter et al., 2012). Meanwhile, Mayer (2005, 2009) has provided a wealth of studies showing the benefits of both individual and sequential images combined with text for learning. While foundational, these prior works have made little distinctions between the different types of multimodal interactions. Are some types of text–image relationships more beneficial to reading comprehension (broadly construed) or education at different levels? This framework provides a way to explore whether certain text–image relations may be more or less advantageous to the educational benefit of learners. In particular, it can help in identifying optimal relations between domains to promote learning, and places where integration of multiple domains may tax cognitive load (e.g., Ayres & Sweller, 2005). Such endeavors echo the research on co-speech gesture showing "mismatches" between gesture and speech both aid learning and reveal readiness to learn (Goldin-Meadow, 2003a; Hutchins & Nomura, 2011). By viewing dynamic multimodal text–image relations in a similar framework, a dialogue and integration can perhaps be formed between these complimentary lines of research.

In light of the proposed model, it is worth noting that full "literacy" for multimodal interactions would cover both the verbal *and* the visual modalities, as opposed to merely using the visual language to bootstrap the understanding of writing or learning

(e.g., Cimermanová, 2015; Mayer, 2009), despite its importance. To this point, this overall parallel architecture emphasizes both the structures involved in behaviors, but also the *interfaces* linking these structures. This is especially important when considering development, both for monomodal and multimodal information, since fluency in a structure requires an interface between structures, not just the structure itself. For example, just because one learns a spoken language that uses syntactic structures, it does not imbue fluency in sign language, which must require lexical items that establish their own interface between syntax, meaning, and bodily structures.

We can imagine a similar relationship for narrative structures. Thus, despite the widespread agreement that comparable structures underlie narrative/discourse across modalities (Cohn, 2013c; Gernsbacher, 1990; Magliano, Radvansky, & Copeland, 2007), just because one develops proficiency in the narrative structures of verbal language, it does not grant fluency in the visual language (or vice versa). Indeed, comprehension and production of visual narrative sequences requires exposure to and practice with cultural visual languages (Nakazawa, 2005, 2015; Wilson, 2015; Wilson & Wilson, 1987), and lack of such experience results in deficiencies in comprehending sequential images (Byram & Garforth, 1980; Fussell & Haaland, 1978). Even within a population of self-described "comic readers," brain responses to manipulations of sequential images differ between levels of proficiency (Cohn & Kutas, 2015; Cohn & Maher, 2015; Cohn, Paczynski, et al., 2012). Development of visual narrative fluency may thus differ based on a learner's exposure to certain types of multimodal relations. If a learner only experiences Verb-Dominant interactions, where the visuals have no grammatical structures, it may not be enough to provide full fluency of Vis-Autonomous or Vis-Assertive sequences. This framework at least provides a way to examine such issues. Concerns regarding fluency are also important to remember for researchers who may design experimental tasks falsely assuming the universality of sequential images, and/or believing that it requires no fluency to analyze such phenomena to begin with.

Finally, the emphasis on interfaces in this model may have additional consequences in clinical contexts. For example, some clinical populations have difficulty understanding verbal discourse. This approach would posit that, without further experimentation, it is unclear whether such deficits belong to the structures themselves or to the interfaces between structures. It may thus be possible to access narrative structures via another interface, like visuals, be it in monomodal or multimodal contexts. Thus, this model gives a basis for clinical populations to potentially benefit from visual language in comprehending verbal language, and provides a framework by which to study these open questions.

6. Conclusion

This approach has proposed a theoretical model that integrates the three primary modalities of human expression into a single parallel architecture. This framework thereby characterizes broad scale multimodal relationships based on interactions between underlying components in the model. While the categories here are discrete, as stated previously, linguistic and communicative performance likely weaves through these interactive types based on the degree to which underlying structures are engaged. Insofar as research on multimodality aims to extend beyond just looking at semantic relationships, we must aim for a model that can account for the varying contributions of each structure in multimodal production and the cognitive processes that allow for such integration.

To these broader goals, this framework provides only a start, and remains limited in that it describes multimodality at a level of interacting "structures." It does not detail how the components *within* those structures interact, which would be crucially important for describing the actual substance of multimodal interactions. Prior works examining the form and semantics of multimodality provide a good start at this endeavor (e.g., Kress, 2009; Martinec & Salway, 2005; Royce, 2007; Stainbrook, 2015), though future work will need to extend beyond these surface observations to describe the component structures underlying those interactions. This increased level of detail would require extensive formalization of both a semantic system *and* the structural components by which those meanings are conveyed (phonology, graphic structure, morphology, syntax, narrative, etc.). In other words, characterizing the component parts of the entire architecture of human language, in verbal, bodily, and graphic forms. Such complexity is exactly why multimodal interactions are both challenging and interesting, and make a worthy goal for future research.

Acknowledgements

Ray Jackendoff, Eva Wittenberg, Stephanie Gottwald, Ariel Goldberg, Kaitlin Pederson, and Ryan Taylor are thanked for their helpful discussion and comments on early drafts of this paper.

References

Asher, N., & Lascarides, A. (2003). *Logics of conversation*. Cambridge: Cambridge University Press.
Ayres, P., & Sweller, J. (2005). The split-attention principle in multimedia learning. In R. E. Mayer (Ed.). *The Cambridge handbook of multimedia learning* (Vol. 2, pp. 135–146). Cambridge, UK: Cambridge University Press.
Baetens, J. (2011). Abstraction in comics. *SubStance, 40*(1), 94–113.
Bateman, J. A. (2014). *Text and image: A critical introduction to the visual/verbal divide*. New York: Routledge.
Bateman, J. A., & Schmidt, K.-H. (2012). *Multimodal film analysis: How films mean*. New York: Routledge.
Bateman, J. A., & Wildfeuer, J. (2014). A multimodal discourse theory of visual narrative. *Journal of Pragmatics, 74*, 180–208. http://dx.doi.org/10.1016/j.pragma.2014.10.001.
Bordwell, D., & Thompson, K. (1997). *Film art: An introduction* (5th ed.). New York, NY: McGraw-Hill.
Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about Semantic Illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research, 1446*, 127–143. http://dx.doi.org/10.1016/j.brainres.2012.01.055.
Byram, M. L., & Garforth, C. (1980). Research and testing non-formal education materials: A multi-media extension project in Botswana. *Educational Broadcasting International, 13*(4), 190–194.
Canseco-Gonzalez, E., Swinney, D. A., Love, T., Walenski, M., Ahrens, K., & Neville, H. (1997). Processing grammatical information using Jabberwocky sentences: An ERP study. *Paper presented at the cognitive neuroscience society, fourth annual meeting, Boston, MA*.
Cheng, L. L.-S., & Corver, N. (2013). *Diagnosing syntax*. Oxford University Press.
Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.
Chomsky, N. (1970). Remarks on nominalization. In R. Jacobs & P. Rosenbaum (Eds.), *Readings in English transformational grammar* (pp. 184–221). Waltham, MA: Ginn.
Chwilla, D. J., & Kolk, H. H. J. (2005). Accessing world knowledge: Evidence from N400 and reaction time priming. *Cognitive Brain Research, 25*(3), 589–606. http://dx.doi.org/10.1016/j.cogbrainres.2005.08.011.
Cimermanová, I. (2015). Using comics with novice EFL readers to develop reading literacy. *Procedia – Social and Behavioral Sciences, 174*, 2452–2459. http://dx.doi.org/10.1016/j.sbspro.2015.01.916.
Clark, H. H. (1996). *Using language*. Cambridge, UK: Cambridge University Press.
Cohn, N. (2003). *Early writings on visual language*. Carlsbad, CA: Emaki Productions.
Cohn, N. (2007). A visual lexicon. *Public Journal of Semiotics, 1*(1), 53–84. <http://www.semiotics.ca>.
Cohn, N. (2010a). Extra! Extra! Semantics in comics!: The conceptual structure of *Chicago Tribune* advertisements. *Journal of Pragmatics, 42*(11), 3138–3146. http://dx.doi.org/10.1016/j.pragma.2010.04.016.
Cohn, N. (2010b). The limits of time and transitions: Challenges to theories of sequential image comprehension. *Studies in Comics, 1*(1), 127–147.
Cohn, N. (2011). A different kind of cultural frame: An analysis of panels in American comics and Japanese manga. *Image & Narrative, 12*(1), 120–134.
Cohn, N. (2012). Explaining "I can't draw": Parallels between the structure and development of language and drawing. *Human Development, 55*(4), 167–192. http://dx.doi.org/10.1159/000341842.
Cohn, N. (2013a). Beyond speech balloons and thought bubbles: The integration of text and image. *Semiotica, 2013*(197), 35–63. http://dx.doi.org/10.1515/sem-2013-0079.

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

Cohn, N. (2013c). Visual narrative structure. *Cognitive Science, 37*(3), 413–452. http://dx.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. *Frontiers in Psychology, 5*, 1–9. http://dx.doi.org/10.3389/fpsyg.2014.00680.

Cohn, N. (2014b). You're a good structure, Charlie Brown: The distribution of narrative categories in comic strips. *Cognitive Science, 38*(7), 1317–1359. http://dx.doi.org/10.1111/cogs.12116.

Cohn, N. (2015). What can you draw without narrative? A hierarchy of grammatical complexity for sequential images (in preparation).

Cohn, N. (2015). Narrative conjunction's junction function: The interface of narrative grammar and semantics in sequential images. *Journal of Pragmatics*. http://dx.doi.org/10.1016/j.pragma.2015.09.001 (in press).

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. http://dx.doi.org/10.1016/j.neuropsychologia.2015.08.026.

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

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

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

Cohn, N., Taylor-Weiner, A., & Grossman, S. (2012). Framing attention in Japanese and American comics: Cross-cultural differences in attentional structure. *Frontiers in Psychology – Cultural Psychology, 3*, 1–12. http://dx.doi.org/10.3389/fpsyg.2012.00349.

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

Culicover, P. W., & Jackendoff, R. (2005). *Simpler Syntax*. Oxford: Oxford University Press.

Culicover, P. W., & Jackendoff, R. (2010). Quantitative methods alone are not enough: Response to Gibson and Fedorenko. *Trends in Cognitive Sciences, 14*(6), 234–235. <http://www.sciencedirect.com/science/article/pii/S1364661310000707>.

Dreyfuss, H. (1984). *Symbol sourcebook: An authoritative guide to international graphic symbols*. John Wiley & Sons.

Emmorey, K., Borinstein, H. B., Thompson, R., & Gollan, T. H. (2008). Bimodal bilingualism. *Bilingualism: Language and Cognition, 11*(01), 43–61. http://dx.doi.org/10.1017/S1366728907003203.

Engelhardt, Y. (2007). Syntactic structures in graphics. *Computational Visualistics and Picture Morphology, 5*, 23–35. <http://yuriweb.com/engelhardt-graphic-syntax.pdf>.

Fauconnier, G., & Turner, M. (2002). *The way we think: Conceptual blending and the mind's hidden complexities*. New York: Basic Books.

Forceville, C. (2011). Pictorial runes in Tintin and the Picaros. *Journal of Pragmatics, 43*, 875–890.

Forceville, C. (2015). Conceptual metaphor theory, blending theory, and other cognitivist perspectives on comics. In N. Cohn (Ed.), *The visual narrative reader: Interdisciplinary approaches to the structure, comprehension, and development of comics and sequential images*. London: Bloomsbury.

Forceville, C., & Urios-Aparisi, E. (2009). *Multimodal metaphor*. New York: Mouton De Gruyter.

Forceville, C., Veale, T., & Feyaerts, K. (2010). Balloonics: The visuals of balloons in comics. In J. Goggin & D. Hassler-Forest (Eds.), *The rise and reason of comics and graphic literature: Critical essays on the form*. Jefferson: McFarland & Company Inc..

Fricke, E. (2013). Towards a unified grammar of gesture and speech: A multimodal approach. *Body–language–communication. An international handbook on multimodality in human interaction* (pp. 733–754).

Fussell, D., & Haaland, A. (1978). Communicating with pictures in Nepal: Results of practical study used in visual education. *Educational Broadcasting International, 11*(1), 25–31.

Ganis, G., Kutas, M., & Sereno, M. I. (1996). The search for "common sense": An electrophysiological study of the comprehension of words and pictures in reading. *Journal of Cognitive Neuroscience, 8*, 89–106.

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

Gernsbacher, M. A., Varner, K. R., & Faust, M. (1990). Investigating differences in general comprehension skill. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 16*, 430–445.

Gibson, E., & Fedorenko, E. (2010). Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences, 14*(6), 233–234.

Goldberg, A. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago, IL: University of Chicago Press.

Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends in Cognitive Sciences, 3*(11), 419–429. http://dx.doi.org/10.1016/S1364-6613(99)01397-2.

Goldin-Meadow, S. (2003a). *Hearing gesture: How our hands help us think*. Cambridge: Harvard University Press.

Goldin-Meadow, S. (2003b). *The resilience of language: What gesture creation in deaf children can tell us about how all children learn language*. New York and Hove: Psychology Press.

Goldin-Meadow, S. (2006). Talking and thinking with our hands. *Current Directions in Psychological Science, 15*(1), 34–39.

Goldin-Meadow, S., & Feldman, H. (1977). The development of language-like communication without a language model. *Science, New Series, 197*(4301), 401–403.

Goldin-Meadow, S., McNeill, D., & Singleton, J. (1996). Silence is liberating: Removing the handcuffs on grammatical expression in the manual modality. *Psychological Review, 103*(1), 34–55.

Goldin-Meadow, S., So, W. C., Ôzyûrek, A., & Mylander, C. (2008). The natural order of events: How speakers of different languages represent events nonverbally. *Proceedings of the National Academy of Sciences, 105*(27), 9163–9168.

Green, J. (2014). *Drawn from the ground: Sound, sign and inscription in Central Australian sand stories*. Cambridge, UK: Cambridge University Press.

Grice, H. P. (1967). William James lectures: Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *1975: Syntax and semantics 3* (pp. 41–58). New York, NY: Academic Press.

Guérin, C., Rigaud, C., Mercier, A., Ammar-Boudjelal, F., Bertet, K., Bouju, A., . . . Revel, A. (2013). eBDtheque: A representative database of comics. *Paper presented at the Document Analysis and Recognition (ICDAR), 2013 12th international conference on*.

Hagan, S. M. (2007). Visual/verbal collaboration in print: Complementary differences, necessary ties, and an untapped rhetorical opportunity. *Written Communication, 24*(1), 49–83.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in english*. London: Longman.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science, 298*, 1569–1579.

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

Hobbs, J.R. (1985). *On the coherence and structure of discourse*. Stanford, CA: CSLI Technical Report 85-37.

Horn, R. (1998). *Visual language: Global communications for the 21st century*. Bainbridge Island, WA: MacroVU Inc..

Hutchins, E., & Nomura, S. (2011). Collaborative construction of multimodal utterances. *Embodied interaction* (pp. 29–43).

Jackendoff, R. (1990). *Semantic structures*. Cambridge, MA: MIT Press.

Jackendoff, R. (1991). Parts and boundaries. *Cognition, 41*, 9–45.

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford: Oxford University Press.

Jackendoff, R., & Wittenberg, E. (2014). What you can say without syntax: A hierarchy of grammatical complexity. In F. Newmeyer & L. Preston (Eds.), *Measuring linguistic complexity*. Oxford: Oxford University Press.

Kehler, A. (2002). *Coherence, reference, and the theory of grammar*. Stanford: CSLI Publications.

Kogo, T., & Kogo, C. (1998). The effects of comic-based presentation of instructional materials on comprehension and retention. *Japanese Journal of Education Technology, 22*, 87–94.

Kress, G. (2009). *Multimodality: A social semiotic approach to contemporary communication*. New York: Routledge.

Kress, G., & van Leeuwen, T. (1996). *Reading images: The grammar of visual design*. London: Routledge.

Kress, G., & van Leeuwen, T. (2001). *Multimodal discourse: The modes and media of contemporary communication*. London: Oxford Press.

Langacker, R. W. (2001). Discourse in cognitive grammar. *Cognitive Linguistics, 12*(2), 143–188. http://dx.doi.org/10.1515/cogl.12.2.143.

Liungman, C. G. (1991). *Dictionary of symbols*. Santa Barbara, CA: ABC-CLIO Inc..

Magliano, J. P., Dijkstra, K., & Zwaan, R. A. (1996). Generating predictive inferences while viewing a movie. *Discourse Processes, 22*, 199–224.

Magliano, J. P., Radvansky, G. A., & Copeland, D. E. (2007). Beyond language comprehension: Situation models as a form of autobiographical memory. In F. Schmalhofer & C. A. Perfetti (Eds.), *Higher level language processes in the brain: Inference and comprehension processes*. Manwah, NJ: Lawrence Earlbaum.

Magliano, J. P., & Zacks, J. M. (2011). The impact of continuity editing in narrative film on event segmentation. *Cognitive Science, 35*(8), 1489–1517. http://dx.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. *Cognitive Psychology, 9*, 111–151.

Marr, D. (1982). *Vision*. San Francisco, CA: Freeman.

Marslen-Wilson, W. D., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition, 8*, 1–71.

Martinec, R., & Salway, A. (2005). A system for image–text relations in new (and old) media. *Visual Communication, 4*(3), 337–371. http://dx.doi.org/10.1177/1470357205055928.

Mayer, R. E. (2005). *The Cambridge handbook of multimedia learning*. Cambridge University Press.

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge University Press.

McCloud, S. (1993). *Understanding comics: The invisible art*. New York, NY: Harper Collins.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago, IL: University of Chicago Press.

McNeill, D. (2000b). *Language and gesture*. Cambridge: Cambridge University Press.

McNeill, D. (2000a). Catchments and contexts: Non-modular factors in speech and gesture production. In D. McNeill (Ed.), *Language and gesture* (pp. 312–328). Cambridge: Cambridge University Press.

McNeill, D., Quek, F., McCullough, K. E., Duncan, S., Furuyama, N., Bryll, R., ... Ansari, R. (2001). Catchments, prosody and discourse. *Gesture, 1*(1), 9–33. http://dx.doi.org/10.1075/gest.1.1.03mcn.

Meek, M. (1988). *How texts teach what readers need to learn*. South Woodchester, UK: Thimble Press.

Mitchell, W. J. T. (1986). *Iconology: Image, text, ideology*. Chicago, IL: University of Chicago Press.

Molotiu, A. (2009). *Abstract comics: The anthology: 1967–2009*. Seattle, WA: Fantagraphics Books.

Münte, T. F., Matzke, M., & Johannes, S. (1997). Brain activity associated with syntactic incongruencies in words and psuedo-words. *Journal of Cognitive Neuroscience, 9*, 318–329.

Nakazawa, J. (2015). Manga literacy and manga comprehension in Japanese children. In N. Cohn (Ed.), *The visual narrative reader* (pp. 157–184). London: Bloomsbury.

Nakazawa, J. (2005). Development of manga (comic book) literacy in children. In D. W. Shwalb, J. Nakazawa, & B. J. Shwalb (Eds.), *Applied developmental psychology: Theory, practice, and research from Japan* (pp. 23–42). Greenwich, CT: Information Age Publishing.

Nalu, A., & Bliss, J. P. (2011). Comics as a cognitive training medium for expert decision making. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 55*(1), 2123–2127.

Nigam, A., Hoffman, J., & Simons, R. (1992). N400 to semantically anomalous pictures and words. *Journal of Cognitive Neuroscience, 4*(1), 15–22. http://dx.doi.org/10.1162/jocn.1992.4.1.15.

Osaka, M., Yaoi, K., Minamoto, T., & Osaka, N. (2014). Serial changes of humor comprehension for four-frame comic manga: An fMRI study. *Scientific Reports, 4*. http://dx.doi.org/10.1038/srep05828.

Osterhout, L., & Nicol, J. L. (1999). On the distinctiveness, independence, and time course of the brain responses to syntactic and semantic anomalies. *Language and Cognitive Processes, 14*(3), 283–317.

Painter, C., Martin, J. R., & Unsworth, L. (2012). *Reading visual narratives: Image analysis of children's picture books*. London: Equi-nox.

Pinker, S. (1994). *The language instinct: How the mind creates language*. New York: HarperCollins.

Pyers, J. E., & Emmorey, K. (2008). The face of bimodal bilingualism: Grammatical markers in American Sign Language are produced when bilinguals speak to English monolinguals. *Psychological Science, 19*(6), 531–535. http://dx.doi.org/10.1111/j.1467-9280.2008.02119.x.

Royce, T. D. (1998). Synergy on the page: Exploring intersemiotic complementarity in page-based multimodal text. *JASFL Occasional Papers, 1*(1), 25–49.

Royce, T. D. (2007). Intersemiotic complementarity: A framework for multimodal discourse analysis. In T. D. Royce & W. L. Bowcher (Eds.), *New directions in the analysis of multimodal discourse* (pp. 63).

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

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

Schnoebelen, T. J. (2012). *Emotions are relational: Positioning and the use of affective linguistic resources* (Doctoral Dissertation). Palo Alto, CA: Stanford University.

Short, J. C., Randolph-Seng, B., & McKenny, A. F. (2013). Graphic presentation: An empirical examination of the graphic novel approach to communicate business concepts. *Business Communication Quarterly, 76*(3), 273–303. http://dx.doi.org/10.1177/1080569913482574.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420.

Stainbrook, E. J. (2003). *Reading comics: A theoretical analysis of textuality and discourse in the comics medium* (Doctoral Dissertation). Indiana, PA: Indiana University of Pennsylvania.

Stainbrook, E. J. (2015). A little cohesion between friends; Or, we're just exploring our textuality: Reconciling cohesion in written language and visual language. In N. Cohn (Ed.), *The visual narrative reader* (pp. 129–154). London: Bloomsbury.

Sutton, V. (1995). *Lessons in sign writing*. La Jolla, CA: The Deaf Action Committee.

van Dijk, T., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

Van Petten, C., & Kutas, M. (1991). Influences of semantic and syntactic context on open- and closed-class words. *Memory and Cognition, 19*, 95–112.

Wang, X., Hu, J., Hengeveld, B. J., & Rauterberg, G. W. M. (2014). Can time perception be affected by interactive comics? *Paper presented at the 22nd International Conference on Computers in Education, Japan*.

Wilkins, D. P. (1997/2015). Alternative representations of space: Arrernte narratives in sand. In N. Cohn (Ed.), *The visual narrative reader* (pp. 252–281). London: Bloomsbury.

Willats, J. (1997). *Art and representation: New principles in the analysis of pictures*. Princeton: Princeton University Press.

Willats, J. (2005). *Making sense of children's drawings*. Mahwah, NJ: Lawrence Erlbaum.

Wilson, B. (2015). What happened and what happened next: Kids' visual narratives across cultures. In N. Cohn (Ed.), *The visual narrative reader* (pp. 185–227). London: Bloomsbury.

Wilson, B., & Wilson, M. (1987). Pictorial composition and narrative structure: Themes and creation of meaning in the drawings of Egyptian and Japanese children. *Visual Arts Research, 13*(2), 10–21.

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

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*(2), 162–185.