Visual/Verbal Collaboration in Print

Complementary Differences, Necessary Ties, and an Untapped Rhetorical Opportunity

Susan M. Hagan
Carnegie Mellon University, Pittsburgh, Pennsylvania

Those who focus on the study of visual information continue to search for effective ways to conceptualize that inquiry. However, many visual examples are better categorized as visual/verbal collaboration, complicating analysis. When analysis is based on the assumption that visual and verbal modalities perform in similar ways, important complementary differences are overlooked. Therefore, this investigation presents a series of observations from a perspective rooted in difference, which leads to the argument that visual/verbal messages develop when cohesive and perceptual relationships form between image and text, resulting in four types of loose to tight visual/verbal collaboration. Examples of each can clarify, contradict, or challenge common understanding for a particular audience. Finally, a perspective in difference uncovers another kind of image/text collaboration, which instead of relying solely on actual images and text, depends on a weave of actual with imagined text and images, leading to an untapped rhetorical opportunity.

Keywords: communication design; interdisciplinary; integrative; graphic; visual rhetoric; visual studies

The study of visual rhetoric is made more intriguing and more difficult because the artifacts examined are better categorized as visual/verbal collaboration. Given that each modality does bring the other to mind, a perspective based on the similarities between the two might seem to be the most useful approach. In fact, that approach has produced useful results. Aspects of parallelism have been identified (Fahnestock, 2003), along with the impact of social convention on interpretation regardless of the modality used (Kostelnick & Hassett, 2003; Kress & van Leeuwen, 1996, 2001),
Additionally, investigations in educational psychology have shown that both imagined and concrete images can be used as textual enhancements capable of improving learning outcomes (Carney & Levin, 2002; Levin, Anglin, & Carney, 1987). However, similarity perspectives have also resulted in analyses that must be challenged. For example, perspectives that support reading images (Kress & van Leeuwen, 1996), as well as the claim that “the visual component of a text is an independently organized and structured message—connected with the verbal text, but in no way dependent on it: and similarly the other way around”¹ (Kress & van Leeuwen, 1996, p. 17), might seem compelling given one set of examples but become unsatisfying given others.

Image reading perspectives become troubling in light of cognitive studies that include Paivio’s (1986) seminal work on the structural and functional differences between visual and verbal information, data on eye scans of images (Buswell, 1935; Norton & Stark, 1971), and explanations concerning how the eye and mind work together to capture information (Solso, 1994). Their work suggests that a grammar of images might be hoped for but not realized because of the particular way that the brain processes sequential information versus information that is simultaneously available for processing. Those processing differences support Arnheim’s (1974) observation that the eye and mind tend toward the easiest thing to do. The easiest thing to do with text, relatively speaking, is to maintain its sequence. The easiest thing to do when looking at images, which the eye sees by way of active looking from feature to feature, is to follow one’s own history and interests. This work will show why differences in looking can lead to important differences in interpretation.

Furthermore, arguments for independent image and text messages are challenged by examples that can only be described as having dependant connections between text and image (Atwan, 2005; Barthes, 1977; Birdsell & Groarke, 1996; Blair, 2004; Bonsiepe, 1965; Ehses, 1989; Faigley, George, Palchik, & Selfe, 2004; Meggs, 1992). Atwan (2005) noticed “convergences” between image and text, while Faigley et al. (2004) reported the feeling of being “caught off-guard by the unexpected meaning which results from placing each word and object together” (p. 12).

At the beginning of this investigation, I initially focused on unexpected meaning possibilities, but ongoing analysis led to a more basic interest in how these two modalities manage to collaborate at all. To explain the interconnected structure and function of visual/verbal collaboration, I used aspects of Arnheim’s (1974) work on perception in art, as well as Halliday...
and Hasan’s (1976) analysis of cohesion in English. By adapting their work to this project, observations emerged that particular choices produce four loose to tight types of visual/verbal collaboration. In time, I became convinced that the mechanism of collaboration should be examined alongside the meaning outcomes produced by that collaboration. Furthermore, I became encouraged that a perspective based on difference not only aids in the analysis of visual/verbal collaboration but also leads to important opportunities related to the weave of imagined images and text with the concrete images and text on the page, especially when applied to arguments concerning foreign communities or concepts that audiences need to integrate into existing contexts in either public or academic settings. As Schroeder, Fox, and Bizzell (2002) argue, new approaches to academic discourse are needed when the old approaches are not up to the task. They noted that “these new discourse forms develop because they enable kinds of rigorous academic work that simply cannot be done within the traditional discourse” (p. x). An example, shown later in this text, illustrates how visual/verbal collaboration can challenge assumptions that might be held by newcomers to a field of study. From another perspective rooted in textual enhancement, Mayer (2002) noted that images aid explanatory text when those images are placed in close proximity to the relevant words. My work argues that stronger ties are available, which invite the reader to move from text to image more effectively than proximity alone allows. However, before examining those claims, the next section provides a detailed look at the assumptions and data that informed this study.

**Background**

My work is based on three foundational assumptions. First, image and text are potentially equal contributors to meaning, even though specific examples might emphasize one or the other modality. Second, visual and verbal information contain structural and functional elements (Paivio, 1986) that produce complementary differences in what each modality concretely communicates. Finally, these differences affect what the audience imagines and what will be communicated in a more concrete fashion when either modality is presented without the other. Images, on one hand, specialize in concrete scenes while often inviting imagined text. Text, on the other, specializes in concrete statements, questions, and demands (Olson, 1994) while inviting imagined images through textual elements such as scenic description and metaphor. Imagining suggests broader individual
interpretations based on the histories and interests of the individuals who do the imagining, whereas concrete elements suggest relatively more stable interpretations. Olson (1994) noted that concrete images produce broad textual interpretations, even for expert “readers,” whereas actual text produces more stable interpretations within cultures that have cultivated shared syntax along with genre expectations. The cognitive psychologist Robert Solso (1994) arrived at similar conclusions. However, when it comes to showing actual scenes, text is no match for the stability of images. Therefore, these two modalities are complementary concerning the concrete and imagined elements that they provide.

The data for this investigation were drawn from 130 examples consisting of 109 professional examples and 21 student examples. Although many examples have similar structures and functions, they were selected to examine a broad array of differences. I sorted these examples into categories that reflected my objectives. My first sort sought to identify the presence or absence of unexpected meaning. However, that method produced two clear problems. First, I had to assume a universal audience that would agree on expected versus unexpected meaning. Second, structurally disparate examples, which shared similar meaning effects, had to be placed in similar categories, whereas structurally similar examples had to be placed in distinct categories. I later chose to sort these examples in terms of their perceptual relationships, whether or not they produced unexpected meaning. I looked for overlaps, proximate placement, or alignment (Arnheim, 1974) between text and image. However, perceptual relationships alone failed to account for a functional relationship between text and image. I soon began to notice that just as images are difficult to analyze as if they were texts, visual/verbal collaboration could not be satisfactorily explained solely in terms of perceptual relationships. Cohesive relationships also seemed to play a part in visual/verbal collaboration. As Halliday and Hasan (1976) noted, cohesive relationships in text form across sentences and cannot be explained in terms of structural considerations. I found similar relationships in the visual/verbal examples. Identifying those cohesive relationships completed the analytical shift from a focus on unexpected meaning to the perceptual and cohesive ties that form loose to tight relationships between visual and verbal information.

Perceptual ties in visual/verbal collaboration include shared location, similar location, alignment, and overlaps. Shared location is not one of Arnheim’s (1974) categories, because it is particular to visual/verbal collaboration in which typographic shape and textual syntax share the same space. Furthermore, although I use Arnheim’s work on perception to show
how similarities in location, size, and alignment help tie visual and verbal elements together in collaboration, I disagree that perceptual structures reveal shared textual interpretations. As mentioned earlier, I find Solso’s (1994) argument concerning the impact of individual histories and interests more compelling.

If perceptual ties allow an audience to notice a connection between two modalities, cohesive ties clarify the nature of that connection. Eight types of cohesive ties have been identified in this work. Cohesive ties are based on, but do not mimic, Halliday and Hasan’s (1976) work, because visual/verbal interaction does not mimic textual sequence. The cohesive ties include blends, exophoric ties, substitution ties, repetition ties, referencing ties, referencing images, overlap collocation, and collocation ties. These types are described in more detail later in this article. However, it is important to note that just as shared location is unique to visual/verbal collaboration, blends are similarly unique because the meaning relationships between typographic shape and textual content are outside of the scope of single-modality interests.

Given particular perceptual and cohesive relationships, this work suggests four basic types of visual/verbal collaboration. They are typographic interplay, interplay in parallel, interplay in sequence, and interweaving. Typographic interplay is a tight type of collaboration that identifies the relationship between typographic shape and its potential meaning combined with textual syntax and its content. On one hand, it can be found in examples as simple as the page you are now reading, in which social conventions concerning typographic shape interact with academic content. On the other hand, typographic interplay can also use shape in ways that challenge or contradict the ordinary meaning of text for a particular audience. An example is presented later.

If typographic interplay is the tightest type of collaboration, interplay in parallel is the loosest. Loose collaboration reflects overall thematic relationships between image and text, whereas tight types of collaboration link specific visual information to specific words. Interplay in parallel is often found in magazine articles, which invite audiences to look or read at their own time and pace. A tighter type of collaboration is produced by interplay in sequence, because it invites an over-time and directed invitation to look and then read in a particular order. This type is often found in art history texts, in which looking is critical to understanding. However, I argue that the most common sequences, such as “see Figure X,” do not reflect the potential inherent in the form. Finally, interweaving is almost as tight as typographic interplay. But unlike typographic interplay, examples of interweaving usually contain few words. Those words maintain a tight relationship to
limited visual information, which collaborate to communicate big picture ideas quickly. Examples such as posters and book covers often, but not always, convey this big-picture information, depending on the potential meaning relationships between text and image.

Each of these types can potentially produce cross-modal meaning that clarifies, contradicts, or challenges ordinary meaning. Cross-modal meaning is defined here as shared understanding gained by an audience that must both look and read. In a very basic sense, all text on the page produces at least a low-level version of cross-modal meaning, because social conventions (Kostelnick & Hassett, 2003), as well as physical characteristics, influence the interpretation of typographic elements, which clarify the text as a student paper, a journal article, or a newspaper clipping. Cross-modal meaning is communicated when concrete visual and verbal elements collaborate to challenge, contradict, or clarify the ordinary, or associative, meaning of isolated modalities. Ordinary meaning refers to the denotative and connotative understandings common to a particular audience. Associative meaning refers to individually held meaning, such as that constructed when a scenic description of a city such as Paris conjures up a street-level view of the Latin Quarter for one person, while another sees the view from her balcony in the fifth arrondissement. When text and image invite associative meaning, the example lies outside of the scope of this work.

When all its parts are in place, this work aids both analysis and production. It provides heuristics for the identification of four loose to tight interactions between visual and verbal elements. Those types can be identified prior to considering the three categories of cross-modal meaning that any single example might suggest. Furthermore, this work provides a space to reflect on the need to clarify, contradict, or challenge the ordinary or associative meaning of text or image, as well as when it might be useful to invite the audience to imagine scene or text. Finally, although this study offers a method of identifying types and considering the cross-modal meaning within each example, it does not identify heuristics for successful collaboration in any but the most general ways. In the next section, I present a brief overview of past work concerning the interaction of text and image, which in some ways forecasts my project.

**Previous Explorations**

As mentioned earlier, cross-modal examples have been studied for decades. In the 1960s, theorists began to consider the visual/verbal nature
of much “visual” communication. Barthes (1977) famously did so by introducing the anchor and relay in the 1960s. About the same time, the design theorist Gui Bonsiepe (1965) attempted to combine rhetorical with semiotic concepts to explain the visual/verbal mix. In the 1980s, the design theorist Hanno Ehres (1989) applied Bonsiepe’s work to poster design, while the visual studies expert W. J. T. Mitchell (1986) argued that texts produce imagined images and images produce imagined text. During the 1990s, the social semioticians Gunther Kress and Theo van Leeuwen (1996) noted that “some things can be ‘said’ only visually, others only verbally” (p. 2), while the design theorist Philip Meggs (1992) argued for visual/verbal synergy, which he found in certain advertising examples. About that same time, the theorists David Birdsell and Leo Groarke (1996) observed “an amalgam of the verbal and the visual” (p. 2), in which visual information plays a critical role in argument. That conversation continues in this century with Kress and van Leeuwen’s (2001) Multimodal Discourse, a framework that stresses the idea of modality similarity through the filter of common semiotic principles that the authors believe “operate within and across different modes” (p. 2). Interestingly, the authors also state, “For us there is now a slight question about the fundamental issue of whether grammars of distinct modes are quite so uncontentiously ‘there’ as our own efforts in relation to images, for instance, suggest” (p. 124).

From a rhetorical perspective, Faigley et al. (2004) noticed unexpected meaning, Atwan (2005) saw convergence, and Blair (2004) found that images contribute to logical arguments in conjunction with text. Goggins (2004) provided a metaview: “The bifurcation of word and image—of visual and verbal rhetoric...threatens to render invisible a whole host of other kinds of theoretical practices, objects and participants because they do not appear on the dichotomized radar screen” (p. 106).

These perspectives form the first steps in the still new study of visual/verbal collaboration. These theorists have provided evidence not only that the collaboration exists but also that visual information contributes to meaning. The following example helps explain why each modality has the ability to clarify, contradict, or challenge the ordinary or associative meaning of the other.

**An Example of the Complementary Differences That Lead to Cross-Modal Meaning**

Consider this image of a teddy bear (Figure 1). Take a moment to reflect on the story that image might bring to mind. Put that story and image aside...
for a moment while you read the words “go bear hunting.” Taken literally, the demand to hunt might produce a mind’s-eye view of an unreasonably large animal with big teeth and a fierce appetite starring down the barrel of your gun. Although that imagined image might be exhilarating, frightening, or repugnant, your view will change considerably if the concrete image of the teddy bear finds a location near the words “go bear hunting” (Figure 2).

The visual proximity between the text “go bear hunting” and the image of the bear builds perceptual ties (Hagan, 2003) that lead, in this case, to visual grouping through proximity. However, proximity does not, by itself, produce meaning bonds. Instead, proximity allows a cohesive tie (Hagan, 2003) to develop. That cohesive tie binds the named bear to the image of the bear. The image, next to that text, transforms the default meaning of the word bear, within the context of “bear hunting,” to a new literal meaning in the presence of a specific teddy bear’s image. The collaboration of perceptual and cohesive ties challenges common associations concerning “bear
hunting” by introducing situated possibilities related to shopping. Furthermore, those same ties narrow imagined stories about this teddy bear to one that concerns acquisition.

This example not only illustrates a tight relationship between image and text but also illustrates the concept of cross-modal meaning. In this case, cross-modal meaning challenges the default meaning of the simple command to “go bear hunting,” a command that is far more slippery than the ordinary meaning of those words might suggest. As Hopper and Closs Traugott (2003) pointed out, “in real-world language a single set of units and constructions must serve a much larger set of functions, owing to memory and parsing limitations” (p. 42). Because both verbal and visual communication elements serve a larger set of functions, visual/verbal collaboration can redirect a typical construction, challenging the mind’s-eye view of the scenic text “bear hunting” as well as the inner speech that an image might ordinarily encourage an individual to compose. Inner speech

Figure 2
An Example of Visual/Verbal Collaboration In Which the Image Affects the Denotative Meaning of the Text While the Text Constrains the Associative Meanings of the Image

is the story that is not printed on the page. Instead, it is suggested by the image.

The rest of this work considers in more detail the complementary differences (Hagan, 2003) between visual and verbal modalities, the cohesive and perceptual ties (Hagan, 2003) that bind those modalities, and the ways in which those ties interact to produce four types of visual/verbal collaboration. Examples of these types illustrate how the information in each modality produces cross-modal meaning that clarifies, contradicts, or challenges ordinary or associative meaning within the context of the intended audience. Finally, the work points to an opportunity to use both imagined and concrete elements as part of a larger strategy.

The Complementary Differences

Visual/verbal collaboration is possible because certain similarities do exist between images and text, allowing the two modalities to form meaning relationships. But visual/verbal collaboration is most useful because images and text contain complementary differences that produce synergistic effects. Complementary differences can be identified using the structural and functional systems of dual coding theory, developed by the cognitive psychologist Allan Paivio (1986). In his seminal work, Paivio described the differences between verbal structures, which are linearly organized for processing, and visual structures, which are “simultaneously available” (italics added) for processing” (p. 60) but not simultaneously processed. These visual and verbal elements function in one of three ways: independently when one system operates at a time without involving the other, simultaneously when systems operate in parallel with no connections, and in a connected fashion when one system activates the other. The propositions presented here focuses on loose to tight connections in which one system activates the other in both imagined and concrete ways.

When considering the complementary differences between text and image, it is important to question the idea of an image grammar (Kress & van Leeuwen, 1996), posited as a socially organized method of reading images that a motivated community might develop either from an implicit or an explicit method of reading. Although Olson (1994) illustrated why image grammars have been historically problematic, arguing that even expert image readers routinely produce varied interpretations of the same image in different situations (see Olson, 1994, for a description), cognitive theory (Paivio, 1986; Solso, 1994) brings more clarity to Olson’s challenge.
From a cognitive perspective, the limitations of vision demand that viewers jump from spot to spot to see an image (or read a text). Individuals cannot see everything within that visible spectrum clearly at any one instant in time (Solso, 1994). In fact, only the objects in the very center of vision communicate clearly. Move just one or two degrees outside of that center, and the world begins to blur. That blurring happens because the fovea in the eye, the area of sharpest sight is the size of a pinhead.

Foveal vision does not provide a panoramic view, but the fovea “more than any other structure is what we see the outside world with” (Solso, 1994, p. 22). Individuals might believe that they see the whole scene all at once, but they actually must focus and refocus to accomplish that goal. The speed with which they do this makes it seem as if sight takes place all at once. Individuals compensate for foveal vision (the inability to see all at once) by making saccadic eye movements. “These movements, quick jumps from one feature to the next, yield a vision characterized by constant movement” (Solso, 1994, p. 25). Because individuals cannot see all at once, they see actively. The physical inability to see all at once is one way meaning emerges differently from visual versus textual information. Even though the end result of bit-by-bit spatial looking provides each individual with a sense of the whole, parts of that whole are ignored by some but seen by others (Buswell, 1935).

Still, it must be noted that reading text also occurs bit by bit, leading to a sense of the whole. Some aspects of the text are remembered, whereas others are not. Therefore, why would image interpretation differ significantly from the interpretation of text? The important consequence of saccadic jumps in an image, which is simultaneously available for processing, as opposed to text, relates to increased personal motivation (Solso, 1994). Differences in motivation produce dissimilar patterns of jumping around an image, which lead to personal gestalt understanding of that experience. For example, a painting of a kitchen shows a refrigerator with a bowl of yarn on top of that appliance, a cat seated on the floor, and an antique oak kitchen table behind the cat and next to the refrigerator. Two different viewers might both look at the cat first. But where will each viewer look next? That lightening quick decision is part of a four-step process.

To begin the process of interpretation, light that hits the retina is transformed into neural information that makes its way to the optic cortex, where early perception takes place. Once that ray of light is reflected on the retina and becomes neural information, billions of neurons take that tiny bit of stimulation to multiple areas of the cerebral cortex, where feature analysis first takes place (Solso, 1994). More important for our purposes, that
feature takes on meaning. It is the meaning an individual finds in a feature that activates the motor cortex to mobilize the eye for movement: the next saccadic jump in the visual dance. The next jump, and the jump after that, are influenced by personal history, interests, and training, along with social conventions, perceptual contrast, and perceptual groupings (Solso, 1994).

In that painting of the kitchen referred to earlier, one viewer’s personal interests might encourage eye movements from the cat to the bowl of yarn, to build a story of playful interaction. However, the other viewer’s history might link the cat to the kitchen table, to remember a story of inevitable destruction. Personal histories, interests, and training lead to saccadic jumps that are idiosyncratic (Buswell, 1935; Solso, 1994), producing personal rather than similar interpretations. In fact, eye-scan studies on images (Buswell, 1935; Norton & Stark, 1971) have shown that individual scan paths do not mimic one another, even if individuals do eventually alight on some shared areas of an image.

Furthermore, a shared process of linking image elements is not likely, because it is not the simplest thing to do when viewing images. As Arnheim (1974) noted, humans perceive objects and object relationships in the simplest way possible. However, although limiting the number of objects in an image will limit what can be perceived and connected (Gombrich, 1982), image editing cannot offer guarantees concerning how individuals with different histories and interests will interpret even limited information unless that information has become culturally constrained. In short, the sequence of looking at image elements, and the inner speech those elements and relationships produce, even within gestalt understanding, not only calls image grammars into question but also calls arguments for image structures as producers of shared meaning (Arnheim, 1974) into question. For that reason, the propositions presented here separate the perceptual from the cohesive.

Of course, typography is a visual structure and for that reason can be processed in ways that seem simultaneously available, but significant differences still exist. For example, if one reads the last few lines of an editorial before deciding to commit to the entire argument, he or she might have done so because the whole text is available at once. Still, that text is not simultaneously available to the extent that images tend to be. This becomes evident if one alters the order of the following words from “the cat lies on the table and destroys the yarn,” to “the cat lies on the yarn and destroys the table,” which produces a substantially different effect on the owner of the table. As Olson (1994) pointed out, the combination of syntax and cultural agreement has resulted in a structure that helps audiences share textual understanding in ways that images cannot, as yet, imitate. Additionally,
text’s structure as a sequenced grouping whose linearity is the simplest thing to perceive (Arnheim, 1974) encourages similar linear saccadic jumps (Solso, 1994). This linearity has helped produce a socially developed syntax: the core difference, and the complementary advantage, words have over images. Text has the ability to construct specific statements, questions, and demands, a function that images cannot mimic (Olson, 1994).

Images also have a core difference and complementary advantage over text. Images have the superior ability to communicate spatial relationships, a function that text cannot duplicate. The words simply do not exist that allow us to concisely and convincingly communicate those connections (Jackendoff & Landau, 1995). Images communicate spatial relationships using elements that depict the world as realistic, distorted, simplified, or nonobjective (Arntson, 1993). The example above shows how those depictions differ (Figure 3). Realistic shapes echo the three-dimensional world. Distorted shapes echo a world that has been pushed and pulled in unnatural ways.

Figure 3
Four Ways of Expressing Visual Information That Move From Reproductions of the Natural World to Objects Not Like Those Found in the World

Realistic  Distorted

Simplified  Nonobjective
Simplified shapes show only the information necessary to name the object. Nonobjective shapes represent geometric or unnamable objects.

After all that, it also must be stated that in some situations, audiences do share, to some degree, the stories that an image communicates or the imagery that a text produces. As Radley (1990) pointed out, objects can be transformed from “ordinary to special purposes” (pp. 46-59), leading to a shared frame of reference. For example, photographs of objects such as the Twin Towers or the Great Wall of China have been transformed into stories of conflict and grim resolve. Even though one individual might construct a statement while another interprets the image as communicating a demand, it is important to note that shared elements sometimes do remain embedded. Similarly, scenic description can produce shared image elements as long as the audience also shares a frame of reference. For example, “The young romantics are doing the mirror thing. Sitting across the table from each other, both have their hands in prayer folds as they look right into the other’s college romance eyes.”

Although the image is short on details (tall, short, blonde, brunette), most of us have seen or felt that moment. However, if obvious commonalities exist, why don’t images contain a grammar? Why can’t scenic description communicate similar views across individuals? Whether or not an image communicates a somewhat shared text or a text communicates a somewhat shared image might depend on the weight of cultural understanding versus the weight of the individual’s history and interests. When culture overrides individual history, shared interpretations take precedence. However, when cultural forces have less weight, individual interpretation takes precedence. In a recent interview (Reitstaetter, Rheindorf, & van Leeuwen, 2005) van Leeuwen referred to this phenomenon in terms of domains, comparing the placement of a stop sign on the street with its placement in an art gallery. The first domain is heavily regulated, whereas the second is not. This observation points to the possibility that the key to modality use lies in understanding when individual imagining will impede ideas that the rhetor wishes to regulate and when imagining will produce a context for a more important point. Sometimes a blonde romantic is just as good as a brunette romantic—sometimes not.

The Cohesive and Perceptual Ties

The importance of the imagined and the concrete can only emerge from a set of propositions that provides a rationale for tying two concrete modalities together. In visual/verbal collaboration, that interaction is a two-part
process based on the work of Arnheim (1974) and Halliday and Hasan (1976). That process begins at the moment of noticing when perceptual similarities, coupled with the human tendency to make the simplest connections possible, encourage perceptual groupings (Arnheim, 1974), which humans notice in different sequences (Buswell, 1935; Norton & Stark, 1971). For example, typography as a visual element depends on proximity to hold letters together as words, while the alignment of words on a baseline encourages the perception of sentences.

In visual/verbal collaboration, perceptual ties produce one part of a two-part interconnection. Although the perceptual ties identified here (Table 1) are based on Arnheim’s (1974) work, differences arise because those perceptual groupings form across modalities. For example, in pictures (as well as in visual/verbal collaboration), elements often group because they exist in a similar location. But in visual/verbal collaboration, shared location is also possible. Shared location cements text with typography. Additionally, although contrast is not the method used to group elements, in visual/verbal collaboration, contrast breadcrumbs provide a way back to the text after inviting an individual to look out at an image. Contrast breadcrumbs, such as bolded text, are sometimes found in art history books, in which looking is critical to understanding.

Other perceptual ties will be familiar to students of Arnheim’s (1974) work. Those ties are proximity, which is the close location of image and text; similar shape, which emerges when a the shape of an image echoes the outside edges of a block of text; alignment, in which image and text are connected by an invisible line, such as a baseline; and finally overlaps, in which images lay over text or visa versa.

### Table 1

<table>
<thead>
<tr>
<th>Name</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Shared location</td>
<td>Cements text with typography</td>
</tr>
<tr>
<td>2. Contrast breadcrumb</td>
<td>Maps text to image and image back to text</td>
</tr>
<tr>
<td>3. Proximity</td>
<td>Close location of text and image</td>
</tr>
<tr>
<td>4. Similar shape</td>
<td>Shape of image echoes type block</td>
</tr>
<tr>
<td>5. Alignment</td>
<td>An invisible connection, such as a baseline</td>
</tr>
<tr>
<td>6. Overlaps</td>
<td>One modality lies on top of the other</td>
</tr>
</tbody>
</table>

Note: Perceptual ties are based on Arnheim’s (1974) work on perceptual groupings.
However, if perceptual ties were the only ties, cross-modal meaning, as shared understanding, would not likely emerge. Although Arnheim (1974) argued that perceptual structures produce shared meaning, this work argues that perceptual ties lead to shared meaning because the modalities also contain potential denotative or connotative relationships that help produce cohesive ties like those found in “bear hunting.” Halliday and Hasan’s (1976) work on cohesion in English is instrumental in identifying those meaning links. Their seminal work argues that texts emerge when ideas form meaning relationships across sentences. The authors focused on those relationships rather than on elements within the sentence, arguing that cohesion within the sentence “could be explained simply as a function of its structure” (p. 7). It is the ability for ideas to bond outside of their internal structures that becomes pivotal to this work, because it shows that visual/verbal collaboration, like textuality, emerges from something in addition to its structure.

Textuality takes the form of two types of reference: those that are situational (exophoric) and those that are textual (endophoric). Exophoric or situational references occur when the meaning relationships are not explicit in the text but can be inferred from the situation. Textual references occur when meaning relationships jump forward to future sentences (cataphoric) or backward to previous sentences (anaphoric). Other relationships form because of word repetition, word substitution, or collocation (close proximity) between related words. Cohesion emerges when denotative and connotative meaning relationships form a text. The following examples illustrate these types.

He is a nice guy. Everyone says so.

In this case, a situational relationship occurs. Although who “he” is does not become explicitly apparent to the reader, the people within the situation agree on those connotations, allowing textuality to emerge.

A pica ruler is on the desk. It has a measurement system used by typographers.

In this textual (endophoric) relationship, the word *it* refers (anaphorically) to the particular “ruler” denoted in the earlier sentence.

A cataphoric reference might read as follows.

These are burnt. Cookies never turn out for me in this oven.

In these two sentences, what is burnt in the first sentence refers to the unfortunate cookies in the next sentence.
Substitution inserts a word that is a synonym in the context of the text.

Sally is still writing her book. She told me the other day the monster has taken over her life.

In this case, “the monster” substitutes for the word *book* in the earlier sentence.

Finally, collocation concerns the close proximity of one word to another related word, which leads to a meaning connection between the two. One example would be

It’s another football Sunday. I wonder if the black and gold will win.

Because of its close proximity to the word *football*, the “black and gold” no longer represents colors but instead a team. This example is not substitution because “football Sunday” references an event, not a team.

Visual/verbal collaboration would not be possible without cohesive ties, just as texts in English would not be possible without cohesion. Barthes’s (1977) work on denotative and connotative relationships suggests cohesion. However, his limited study notes only that text anchors the polysemous image, whereas this work suggests that image and text contain situational relationships, reference each other forward and back, substitute for each other, repeat, and produce collocative bonds. The names and nature of the cohesive ties can be seen in Table 2.

A core difference between text and visual/verbal collaboration produces types of cohesion that are unique to visual/verbal collaboration. These

---

**Table 2**

*Cohesive Ties: Names and Definitions of the Methods Used to Create Relationships Between the Meaning Potential Held in Both Text and Image*

<table>
<thead>
<tr>
<th>Name</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Exophoric ties</td>
<td>Connections are loosely implied</td>
</tr>
<tr>
<td>2. Blends</td>
<td>Content links typographic shape and text</td>
</tr>
<tr>
<td>3. Substitution ties</td>
<td>Image replaces a noun</td>
</tr>
<tr>
<td>4. Repetition ties</td>
<td>Image repeats a noun</td>
</tr>
<tr>
<td>5. Referencing ties</td>
<td>References from text out to the image</td>
</tr>
<tr>
<td>6. Referencing images</td>
<td>Elements that refer from an image back to text</td>
</tr>
<tr>
<td>7. Overlap collocation</td>
<td>One modality overlaps the related other</td>
</tr>
<tr>
<td>8. Collocation ties</td>
<td>Content links to closely placed image and text</td>
</tr>
</tbody>
</table>

---
particular differences emerge because of the spatial/linear nature of visual/verbal collaboration, in which a greater number of perceptual relationships are available. Examples follow that illustrate why existing cohesive categories do not map seamlessly onto visual/verbal collaboration, even though the basic concepts embodied by those meaning relationships do. In visual/verbal collaboration, eight types of cohesive ties have been identified thus far. They are blends, exophoric ties, substitution ties, repetition ties, referencing ties, referencing images, overlap collocation, and collocation ties.

Blends are unique to visual/verbal collaboration. They form when the potential meaning in typographic shape and the meaning of textual content form a relationship that produces meaning associations built from cultural, physical, or contextual relationships. For example, meaning associations emerge in the typographic/text collaboration in this journal because of readability needs that drive typographic choice, cultural expectations that further constrain that choice, content interests that drive textual choice, and stylistic expectations that further constrain that text. The next tie, the exophoric tie, appears when connections between images and text are loosely implied rather than particularly referenced. These loose relationships emerge between the collection of images on the page and the overall text, inviting the audience to look and to read at their own time and pace. They are common in popular magazines. A third tie, the substitution tie, is found when an image replaces an earlier noun or verb, or visa versa. For example, if the second occurrence of the word bear is replaced by an image of a bear in “The bear is for sale. To buy the bear, see the cashier,” a substitution tie replaces textual repetition. Repetition ties in visual/verbal collaboration are different from substitution ties because textual meaning is immediately repeated by image meaning. For example, textual substitution occurs in the sentences “I just painted my living room walls. I love all that blue.” In those sentences, the word blue substitutes for the word walls, producing a meaning relationship. The word blue was not present in the first sentence and does not need to be there, because meaning builds sequentially. However, if a color picture of a specific blue wall is placed above that same text, cross-modal repetition emerges from spatial/linear interaction.

When one wants to invite the audience out of the text and into a specific image (common in art history texts), referencing ties such as demonstrative pronouns, suggestions, or commands can refer to that image. These referencing ties can be used in combination with referencing images. Referencing images use visual elements that provide an effective return to the text after viewing a table or image. When arrows or lines provide a path to and from the text, they reference the text. A seventh tie, called overlap collocation,
emerges when one modality lies on top of the other. Placing the phrase “football Sunday” over the colors black and gold communicates an upcoming Pittsburgh Steelers game to a knowledgeable audience. Finally, collocation ties emerge when images and text, with potential meaning relationships, are placed close together. For example, if two photographs of the same teenager are displayed, one with the phrase “Rhodes scholar” and the other with the phrase “convicted felon,” the potential meaning relationships in each example produce cross-modal collocation.

**Four Types of Visual/Verbal Collaboration**

When perceptual and cohesive ties combine in identifiable ways, the four types of visual/verbal collaboration shown on the next page become apparent (Figure 4). In any example within those four types, these ties work together with the complementary differences between image and text to build cross-modal meaning. This section identifies the four types of visual/verbal collaboration and then uses examples to show how cross-modal meaning emerges in particular situations. The four types, as mentioned earlier, are typographic interplay, interplay in parallel, interplay in sequence, and interweaving. *Interplay* is a critical term because the word *typography* sometimes indicates the study of shape divorced from meaning. Müller-Brockmann, the father of the grid, championed this idea. He believed that typography and grid elements should present content in a neutral fashion (Schweimer-Scheddin, 1995). *Typographic interplay*, on the other hand, identifies the meaning potential between typographic shape and text. Typographic interplay emerges when the cohesive tie, the blend interacts with the perceptual tie. Shared location results in the tightest relationship between word and image. *Blends* and *shared location* might seem to be synonyms, but they indicate the interaction between structure and function present in all types of visual/verbal collaboration.

As others have noticed (Drucker, 1994; Eckkrammer, 2004; van Leeuwen, 2005), typography and words interact. Although Drucker (1994) posited that text can be “unmarked,” as well as “marked,” by its typography, this work argues that text and typography are always marked, blending in ways that are either culturally meaningful or contextually meaningful or lean toward personal associations. No text is unmarked, even though the experience of reading can reduce the effect of the typography over time. Van Leeuwen (2005) echoed this sentiment, noting that connotative meaning in typography strengthens cultural expectations, and communicates
metaphorically (as when the word anger is **ANGRY**). However, that perspective does not account for typography/text relationships that challenge or contradict collective understanding, nor does it specifically isolate typographic interplay from types of interplay with images.

In specific instances of typographic interplay, the cohesive tie (blend) and perceptual tie (shared location) produce meaning that clarifies, contradicts, or challenges default understanding. In Figure 5, blending typographic

---

**Figure 4**

Four Types of Visual/Verbal Collaboration Along With the Cohesive Ties and Perceptual Ties Used to Create Them

<table>
<thead>
<tr>
<th>Types</th>
<th>Perceptual Ties</th>
<th>Cohesive Ties</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Typographic Interplay</strong></td>
<td>shared location</td>
<td>blend</td>
</tr>
<tr>
<td>(Without image)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Interplay in Parallel</strong></td>
<td>similar location</td>
<td>exophoric tie</td>
</tr>
<tr>
<td>(Loose Link)</td>
<td>similar shape</td>
<td></td>
</tr>
<tr>
<td></td>
<td>alignment</td>
<td></td>
</tr>
<tr>
<td></td>
<td>overlap</td>
<td></td>
</tr>
<tr>
<td><strong>Interplay in Sequence</strong></td>
<td>contrast breadcrumb</td>
<td>substitution tie</td>
</tr>
<tr>
<td>(Text to...)</td>
<td>similar location</td>
<td>referencing image</td>
</tr>
<tr>
<td></td>
<td>similar shape</td>
<td>repetition tie</td>
</tr>
<tr>
<td></td>
<td>alignment</td>
<td>collocation tie</td>
</tr>
<tr>
<td></td>
<td>overlap</td>
<td>referencing tie</td>
</tr>
<tr>
<td></td>
<td>collocation</td>
<td></td>
</tr>
<tr>
<td><strong>Interweaving</strong></td>
<td>similar location</td>
<td>overlap collocation</td>
</tr>
<tr>
<td>&amp; read</td>
<td>similar shape</td>
<td>collocation tie</td>
</tr>
<tr>
<td></td>
<td>alignment</td>
<td></td>
</tr>
<tr>
<td></td>
<td>overlap</td>
<td></td>
</tr>
</tbody>
</table>

Perceptual and cohesive ties can mix in multiple ways, not just in the order found here.
shape with text in a shared location embeds a message within a message in a way that either clarifies or challenges meaning, depending on the audience. The words “the fate of the earth” read normally. However, a change in typographic shape allows an immediate conclusion to emerge from the headline text. That one-word conclusion embedded in the text is hate. A combination of blend and shared location embed hate in the fate of the earth.

Because all of the types contain typography, typographic interplay is contained within the other three types of visual/verbal collaboration. Those types, which add nontypographic image elements (from realistic images to nonobjective shapes) are interplay in parallel, interplay in sequence, and interweaving. Interplay in parallel uses the following perceptual ties: proximity, alignment, similar shape, and overlaps. However, only one cohesive tie marks this type. The exophoric tie produces nonspecific connections between the text and the images that an audience infers are related. Text provides statements, questions, and demands concerning representations, such as profiles, histories, or arguments, without referring to any specific
images that are also present on the page. The visual information provides concrete scenes associated with the overall text, producing an exophoric tie. For example, if a text makes statements about a college while images of a classroom are in close proximity to those statements, the audience infers a relationship between the two that is not explicitly spelled out. This parallel arrangement leaves the audience free to look or read at their own time and pace.

If interplay in parallel downplays specific modality interactions, interplay in sequence explicitly develops them. Although the perceptual ties are similar to those found in examples of interplay in parallel (proximity, alignment, similar shape, and overlaps), this type emerges when images and text reference each other at particular moments during the reading and viewing experience. A variety of cohesive ties might be present, but they always develop a sequence that is either author invited or audience directed. Cohesive ties, including referencing ties, repetition ties, substitution ties, collocation ties, and referencing images, are common in art history books, travel guides, and multipanel comics. These cohesive ties invite the reader to look out from the text at particular moments. Although a substitution tie sometimes inserts an image in place of a noun or verb, which could be considered a synonym, the use of textual conventions such as “(Figure X)” can also parenthetically substitute for the image. Boldface type, “(Figure X),” is a contrast breadcrumb to help the eye move back to the proper place in the text after viewing the image. Unlike contrast breadcrumbs, which act as perceptual way finders, lines and arrows, here known as referencing images, draw a connection from text to image that encourages a sequentially cohesive relationship. Figure 6 on the next page contains a referencing tie in combination with a referencing image. The text states, “This is Iwakura Lain,” producing a referencing tie by linking the demonstrative pronoun this to the image of the girl, while the line under the text connects to the illustration, providing a referencing image back to the text.

Although that example produced an author-invited sequence that introduces the central conflict of Lain’s life, other situations need to support audience-directed sequences. Audience-directed sequences are common in genres such as travel guides that clarify understanding by allowing an audience to engage in an ongoing and repeatable process of inquiry concerning individual experience. Because author-invited sequences, which build a particular story over time, contain the same cohesive and perceptual ties as audience-directed sequences, cultural expectations often provide the final clues that allow an audience to know which of these sequences has been developed. For example, in travel guides, repeated numbers often tie word
and image together cohesively. In those cases, the numbers that mark locations on the map repeat as numbered descriptions of those locations. This repetition allows the audience to choose a self-directed sequence of exploration within a limited set of options.

Multipanel cartoons lie somewhere between author-invited and audience-directed sequences. Although their narrative form and bounding boxes suggest a cohesive linear progression, the visual/verbal collaboration within each box invites the audience to access image or text using their own sequence. The bounding boxes, a cohesive tie particular to this genre, suggest a sequence based on social convention that helps tell the story properly.

Interweaving uses most of the perceptual ties that mark interplay in parallel and interplay in sequence but introduces a new type of cohesive tie, overlap collocation. That tie, along with collocation ties and repetition ties,
creates specific relationships between text and image in which sequence is not an overriding factor because of the limited amount of information present. For that reason, interweaving invites the audience to process big-picture ideas quickly. These broad-based concepts are often meant to communicate directly and deeply (Arnheim, 1974), as well as concisely. As “bear hunting” illustrated, proximity to a limited amount of text can allow repetition ties to clearly emerge. Like other types, interweaving invites cross-modal meaning. In the following example described here, visual elements contradict the ordinary meaning of the text in a way that makes it ironic. Imagine a holiday package with a handwritten tag that contains the words “Merry Christmas.” Those words traditionally bring to mind a family holiday with gifts and maybe a few carols. However, what if those words were tied to visual information that invited the audience to rethink default meaning?

In this case, the Christmas tag overlaps a package that is wrapped as one might with ribbon, but this particular package is wrapped with barbed wire. This description of interweaving was designed by Rand (1940) for the December 1940 issue of Direction Magazine. “Merry Christmas,” held captive in a barbed-wire package, communicates a horrific irony manifested in a presentation of contradictions, contradictions that challenge the idea of merriment in the face of the tragedies of World War II. The wish to be “merry” was less likely when so many men and women could not enjoy the simple pleasures of extended family gatherings—and in fact never would again. The claim that Christmas is merry was not to be taken seriously, because the perceptually and cohesively tied evidence showed otherwise. Those ties allow words to direct the meaning of the image, while the image challenges the ordinary meaning of the text. The image in isolation would have many options for interpretation. The package could fit a number of national and personal holidays. The text in isolation has ordinary meaning for mid-20th-century Christians. Because meaning is slippery, and can be altered within the context of the situation, a persuasive effect emerges with fresh clarity.

**Cross-Modal Meaning**

The examples shown earlier not only illustrate how cohesive and perceptual ties build four types of visual/verbal collaboration but also indicate that particular examples produce cross-modal meaning capable of clarifying, contradicting, or challenging mind’s-eye imagery and inner speech. These meaning shifts are not simply a complicated parlor trick. The stakes
for developing cross-modal meaning as a rhetorical move are quite high, because language alone is not always sufficient to communicate shared understanding.

Meaning, as shared understanding, is often considered from a basis in language (Halliday & Matthiessen, 1999), a foundation that images are said to share (Fei, 2004; Kress & van Leeuwen, 1996; Muntigl, 2004). However, as Paivio’s (1986) work suggests, the structure and function of images call those language assumptions into question. More intriguingly, the limitations of language invite questions concerning why language enjoys its status as the foundation of thought. Although text does communicate meaning more effectively when one wants to make a clear statement, question, or demand (Olson, 1994), clarity decreases as soon as language is used to represent people in environments. The words simply have not been invented to communicate specific spatial relationships (Jackendoff & Landau, 1995).

In their essay *Spatial Language and Spatial Cognition*, Jackendoff and Landau (1995) pointed out the fundamental problem concerning language’s ability to represent spatial relationships. The authors found that the few words used to describe these relationships number around 80 to 100 prepositions. Although the authors pointed out that some verbs do a similar job, those verbs can be folded into similar prepositions. When you compare the number of “relationship words and phrases to the tens of thousands of object names” (p. 107), the constraints on language become obvious. For that reason, spatial description can only provide a few clues about the environment, leaving an audience to fill in the blanks using personal associations. These personal associations can bolster engagement with the text because they help the reader establish a known context for that engagement. But when an audience does not have the experience to correctly imagine, or when an idea has critical features that imagining can miss, a basic problem with language, as the carrier of thought, surfaces. In those cases, cross-modal meaning might be more rhetorically useful.

The concept of cross-modal meaning is built on the verbal and visual information that form mental representations of thought (Paivio, 1986). From that perspective, the complementary and collaborative strengths in each modality, rather than the strengths inherent in language (Halliday & Matthiessen, 1999) or visual information (Arnheim, 1969), form the basis of thought and meaning. A simple example suggests that shared nature. When the words “Mr. Chips” stand alone without an image, “Mr. Chips” could refer to a man—or a cookie. Similarly, an image of a man or a cookie by itself remains anonymous. The contextual interaction of text and image
clarifies descriptive aspects of the image and scenic aspects of the text. In this case, Mr. Chips is a very specific man.\(^6\)

Although Mr. Chips clarifies meaning by putting the gentleman, or the cookie, in the right category, other situations clarify meaning by making spatial relationships more accessible. For example, although Levin et al. (1987) noted that mind’s-eye imagery can improve learning, their work and other studies (Carney & Levin, 2002; Mayer, 2002; Plass, Chun, Mayer, & Leutner, 1998) also suggest that concrete images more effectively improve understanding, especially for novice learners (Mayer, 2002). Many examples relate to way-finding with maps or interpreting text that explains mysteries such as blood pressure (Levin et al., 1987) or pulleys (Mayer, 2002). Visual information in collaboration with text even helps students more effectively acquire a second language (Plass et al., 1998). However, the ties that hold visual and verbal information together in collaboration have not been fully considered. To date, only proximity to the relevant text (Mayer, 2002) has been explored.

Although visual/verbal collaboration provides critical information that aids description and explanation, that collaboration might also provide critical information that aids argument in situations dealing with foreign communities or concepts. When the stakes concern the basic humanity of a community that has been diminished by a dominant group, language alone might not be convincing. Let Us Now Praise Famous Men (Agee & Evans, 1941) provides an interesting case that illustrates the problem authors have when trying to help an audience learn about people and places that they either do not know or do not like. In that book, the writer James Agee and the photographer Walker Evans seek to help outsiders understand the lives of sharecroppers in the 1940s. However, because their author/photographer interaction was practically noncollaborative, Agee did not refer out to Evans’s photographs, nor did Evans use Agee’s work to clarify the information in his images. Agee did not do this, even though he knew his readers were not knowledgeable collaborators with the necessary experience to build mind’s-eye imagery that could approximate what Agee referred to as “human actuality” (Bartholomae & Petrosky, 2002).

Although Agee could have used Evans’s photographs, he stayed with the modality he trusted and preferred, even though his task centered on evoking an accurate vision of sharecroppers’ homes using a cumbersome, time-consuming, and ultimately ineffective tool. His awareness of the problem seems evident in this statement: “If I bore you, that is that” (Agee & Evans, 1941, p. 10). Here is one excerpt that illustrates Agee’s struggle:
Two blocks, of two rooms each, one room behind another. Between these blocks a hallway, floored and roofed, wide open both at front and rear: so that these blocks are two rectangular yoked boats, or floated tanks, or coffins, each, by an inner wall, divided into two squared chambers. The roof, pitched rather steeply from front and rear, its cards met and nailed at a sharp angle. The floor faces the earth closely. On the left of the hall, two rooms, each an exact square. On the right a square front room and, built later, behind it, using the outward weatherboards for its own front wall, a lean-to kitchen half that size.

At the exact center of each of the outward walls of each room, a window. Those of the kitchen are small, taller than wide and are glassed. Those of the other rooms are exactly square and are stopped with wooden shutters. (p. 138)

It is difficult to decide what small means but impossible to know how much taller “taller than wide” might be. Even if the audience grasps some of the geometry in the text, they are not likely to see the view. However, because both modalities were available in that book, a visual/verbal experience was possible. The sharecropper’s living situation would have been that much more difficult to consider had the photographs only been on display in a museum. Even so, trying to match Agee’s text with Evans’s images is an exercise in frustration. Although the loosest exophoric tie does hold image and text together, the lack of referencing ties between Agee’s specific descriptions and Evans’s related images leads to difficulties in understanding. Although Evans might argue that the images were meant to stand alone, and although Agee might argue that the text is independent of the images, their proximity invites the audience to make connections (Arnheim, 1974). However, this work is a landmark in that it is a starting point for new excursions into an untapped rhetorical opportunity.

Considering the Imagined and the Concrete: An Untapped Opportunity

Despite its problematic aspects, Agee and Evans’s (1941) book made an important contribution concerning the use of concrete modalities employed in a serious attempt to challenge an audience’s default understanding. However, new assumptions might produce even better outcomes. Two examples, which illustrate collaborative possibilities, follow. The first concerns the basic assumptions of a field of study. The second concerns an argument related to the basic assumptions of a community. Both challenge assumptions of ordinary meaning, but each approaches the use of the imagined and the concrete in ways that meet the needs of the situation.
In the first case, the example challenges certain problematic assumptions related to industrial design, in which an interest in visually satisfying products sometimes overrides a competing need for safe products. That collaborative message, one that challenges an emphasis on slick production, states, “good design, bad baby.” Above the text, a drawing of a 10-month-old can be seen with his head stuck in a stroller between the padding and the side frame that supports the handlebar. The ordinary meaning of the words “good design” alone could invite an aesthetic definition. The words “bad baby,” separated from the image, might indicate a proclivity, on the part of the baby, for jumping from chairs into strollers, in which case, the design would not be at fault. The image by itself shows the poor baby, but that image emphasizes the immediate problem—the child is in trouble—rather than the flawed assumptions of the designer.

In collaboration, the concrete image of the “trapped baby” lends irony to the words “bad baby,” altering the connotations, and producing dark humor. Furthermore, the concrete image of baby and stroller challenges a metaview of “good design,” one that stresses good looks over safe structure, and might also improve learning (Carney & Levin, 2002). Without the text, the image of the baby could be interpreted as simply a cruel joke. Without the image, the text generates imagery that does not create the ironic effect intended. In this case, an effective means of persuasion was developed using both concrete image and text.

However, another perspective on the use of the imagined and the concrete in visual/verbal collaboration is possible. This perspective begins when researchers and practitioners begin to reflect on why and when concrete or imagined elements should be introduced. Text has no choice but to invite imagined scenes, and images have no choice but to invite imagined texts. Visual/verbal collaboration at first seems to make its greatest contribution in situations that benefit from the inclusion of two concrete modalities, which lessens the possibility of interpretive confusion. That assumption has certainly proved useful in information design. However, from a rhetorical perspective, visual/verbal collaboration might eventually make its greatest contribution not because it makes more elements concrete but because it can weave both the imagined and the concrete in powerful ways. Consider the implications in the following example. First, it must be pointed out that this work originally included motion and sound, but it is included here for two reasons. The example can be considered for the contributions made by image and text alone. Furthermore, when a short analysis of sound is added, not only does that complementary contribution begin to emerge, but that contribution suggests that multimodal study might also begin from an entry point in difference rather than similarity.
Consider the implications in this example, one that begins by asking the audience to build a mental model of powerful women using only text as a guide (Figure 7). Continuing textual cues further encourage the audience to construct mind’s-eye imagery of strength and compassion and compare those qualities with culturally compelling but less convincing aspects of power based on physical beauty (Figure 8). These mind’s-eye images become a meeting ground, a context, in which the concrete image of a stereotypically pitied woman, an American Muslim woman in hijab, is concretely placed. She now lives in that mental space, a space that she could not have entered before (Figure 9); a space in which calf muscles, silken hair, and unblemished skin take a back seat to compassion and strength.

However, as stated earlier, this example is also meant to include motion and sound. The image of this woman is not the only concrete element. Her strong voice, rather than an imagined voice, describes why she belongs in that mental space when she states, “I get my power from my compassion, my strength, my understanding, my patience, and my determination. Do not pity me. I take power from my piety.” A strong voice might not have been one that a non-Muslim population could construct as human actuality. The example suggests that the imagined and the concrete extends to sound as
well as image and text. Each modality, including sound, contributes both concrete and imagined elements, setting up a new space for reflection concerning the concept of power and how it might be defined. For that reason, I argue that these propositions for visual/verbal collaboration can be used as a springboard for multimodal study.

Weaving the imagined and the concrete is a vastly underexplored method of encouraging critical thought. But when researchers consider how that method might be used to contradict or challenge mind’s-eye imagery or inner speech, an answer to an important question Mitchell (1994) posed emerges; “the real question to ask when confronted with these kinds of image/text relations is not ‘what is the difference (or similarity) between

the words and the images?’ but ‘What difference do the differences (and similarities) make?’” (p. 91).

**Conclusion**

This article has presented complementary differences between visual and verbal elements rooted in the statements, questions, and demands text provides and the spatial relationships images provide. However, as this work has also shown, the simple presence of complementary modalities is not enough to explain visual/verbal collaboration. Visual/verbal collaboration,
as shared understanding, does not emerge unless visual and verbal elements have potential denotative and connotative meaning relationships. When these relationships are in place, cohesive ties work with perceptual ties to build four types of collaboration. The first is typographic interplay, in which language and typography blend in a shared location. Typographic interplay in turn collaborates with other image elements to produce interplay in parallel, interplay in sequence, and interweaving. Interplay in parallel is loose visual/verbal collaboration. Typographic interplay, interplay in sequence, and interweaving are tight visual/verbal collaborations. These four types of collaboration are potentially capable of clarifying, contradicting, or challenging ordinary meaning for a particular audience. When types of visual/verbal collaboration produce cross-modal meaning that deliberately weaves imagined and concrete elements, an untapped opportunity to encourage reflection, broaden motivation, and improve learning outside of the audience’s stated goals, and toward larger social concerns, might be present. As researchers examine the relevance of rhetorical study in this century and seek to build a framework for the study of this singular category, those untapped opportunities should also be explored. I hope this work will encourage researchers to look at these artifacts as of their own kind, demanding an exploration based on their own unique qualities.

Notes

1. Kress and van Leeuwen (1996) noted that in layouts of image and text, positioning on the page does affect what is stressed, as well as what is perceived as the given and new.
2. Flower and Hayes (1984) have also written about multiple representations of knowledge, but they considered the effect those representations have on the writer’s process of composing, which lies outside of the scope of this project.
3. Image interpretation seems to be influenced by textual cues that affect what an individual explicitly notices or loses in the visual smear. Yarbus (1967) asked participants a question before they looked at a painting titled “The Unexpected Visitor.” A question such as “What has the family been doing before the arrival of the ‘unexpected visitor’?” produced a very different eye scan than a question that asked the participant to estimate how long the “unexpected visitor” had been away from the family.
4. Eckkrammer (2004) proposed degrees of inseparability between text and image in intersemiotic relationships. But her work did not include the separate and necessary category for typography/text interaction, nor did it consider the range of meaning relationships present in each type.
5. “Figure X” is not a referencing tie, because there is no suggestion or command to look out to the image.
6. This example brings up an important point. Although cross-modal meaning, within this perspective on visual/verbal collaboration, is dependent on cohesive as well as perceptual ties, visual/verbal interaction is not. When contextually shared denotative or connotative elements
prove elusive, or when the pull of individual histories and interests becomes too strong, the interaction between image and text begins to rely on perceptual ties coupled with the associative relationships that any particular individual cares to construct. Even though associative meaning in visual/verbal collaboration is outside of the scope of this article, it can be considered by making one alteration to this work. The altered work would replace cohesive ties with associative relationships and reconsider the subcategory of cross-modal meaning.

References


Downloaded from http://wcx.sagepub.com at CNTRL INST INDIAN LANG on December 29, 2008


**Susan M. Hagan** received her master's degree in design and her PhD in rhetoric from Carnegie Mellon University. She focuses her research interests on the collaboration of text and visual information as interdependent contributors to meaning. A teacher, writer, and member of the advisory board of the peer-reviewed design journal *Artifact*, she has presented on this topic at conferences sponsored by rhetoric and design societies and is now editing her dissertation for journal and book publication.