Rhetorical relations in multimodal documents

Maite Taboada and Christopher Habel

Discourse Studies 2013 15: 65
DOI: 10.1177/1461445612466468

Article

Discourse Studies
15(1) 65–89
© The Author(s) 2013
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1461445612466468
dis.sagepub.com
SAGE

Rhetorical relations in multimodal documents

Maite Taboada
Simon Fraser University, Canada

Christopher Habel
University of Hamburg, Germany

Abstract

We present a corpus-based study of coherence in multimodal documents. We concern ourselves with the types of relationships between graphs and tables and the text of the document in which they appear. In order to understand and categorize the types of relations across modalities, we are making use of Rhetorical Structure Theory (Mann and Thompson, 1988), and propose that this can adequately describe these types of relations. We analyzed a corpus comprising three different genres, and consisting of about 1500 pages of material and almost 600 figures, tables and graphs. We show that figures stand in both presentational and subject matter relations to the text, and that the relationship between figures and text is one of a small set out of the larger possible rhetorical relations. We also discuss several issues that arise in the treatment of multimodal material, such as the potential for multiple connections between figure and text.

Keywords

Coherence, genre, multimodality, Rhetorical Structure Theory

Introduction

A great deal of work in the last few years has focused on the relationships between pure text and material presented through other modalities, be it visual, audio, or a combination of the two. Research on document design and learning has been trying to elucidate what kind of impact multimodal material has on the reader. Mayer (2009), for instance, reports

Corresponding author:
Maite Taboada, Department of Linguistics, Simon Fraser University, 8888 University Dr., Burnaby, BC, Canada V5A 1S6.
Email: mtaboada@sfu.ca

on years of studies showing that students learn better when they learn from words and pictures than when they learn from words alone. The learning, however, is improved only when certain principles in the presentation of the multimedia material are followed. The principles refer to the order of presentation, the coherence of the text and pictures, and the type of cross-reference used.

Much research has studied whether to use multimodal material or not, where to place it, and what effect captions or other verbal information surrounding such material have on the reader (for an extensive literature review, see Acartürk, 2010). Less frequently discussed is the nature of the relationship between depictive material and the text itself, that is, whether pictures, charts, tables, diagrams, etc., serve as illustration, merely decoration, evidence, example, or something else.

In this article, we concern ourselves with the types of relationships between pictures, figures and tables, and the text in which they appear. The rhetorical relations between figures and text can be understood as coherence links, contributing to the perceived coherence of a document (Taboada, 2004). One fundamental assumption in the study of discourse and communication is that most discourse is coherent. In Rhetorical Structure Theory (RST), one theory that tries to account for text coherence, coherence is understood as the absence of non-sequiturs, that is, as a property of texts whereby all parts of a text have a reason to be in the text and, furthermore, there is no sense that there are parts that are somehow missing (Mann and Taboada, 2010; Mann and Thompson, 1988). Our hypothesis is that multimodal documents exhibit a form of coherence that links not only the verbal material, but also the depictive material – the links being rhetorical relations.

This view of coherence is also framed within the notion of genre, the view that texts and discourses are a result of the context in which they are produced and processed, and that the specific goals of a genre have an effect on that genre's structure and lexicogrammatical realizations. For our approach to genre, we follow Systemic Functional Linguistics. Within that school, the widely quoted definition by Martin is that genre is 'a staged, goal-oriented, purposeful activity in which speakers engage as members of our culture' (Martin, 1984: 25). The study of genre within Systemic Functional Linguistics has concentrated on structural characterizations through genre staging. Stages are the constitutive elements of a genre, which follow each other in a predetermined fashion, specific to each genre. Eggins (1994) characterizes the staging, or schematic structure of a genre, as a description of the parts that form the whole, and how the parts relate to each other. This is achieved following both formal and functional criteria.

Because we believe genre constrains the types of figures present, their placement and their relationship to the text, we study three different genres: newspaper articles, magazine articles in a scientific magazine, and scientific articles. Newspaper articles have the main goal of informing and entertaining, and develop in stages determined by the 'inverted pyramid' structure of newspaper writing (see Scanlan, 2000, for a description and a history). Magazine articles tend to be more entertaining in both content and presentation, but in our case the articles originate in a scientific magazine, whose purpose is to disseminate research to a wider audience – professional, but also potentially lay. Finally, scientific articles have a much more restricted audience of researchers in a specific area. Their generic structure varies, but it tends to follow the introduction–method–result–discussion structure described by Swales (1990).

We are concerned mostly with figures (flow diagrams, bar or other charts, system overviews, maps, etc.) and tables. We will discuss briefly the role of photographs, but the study of photographs themselves and their relationship to the text will be left for future work (as we believe it is more complex than that between figures and text). We will often use the generic word ‘figure’ or ‘graph’ to refer to the types of illustrations discussed, as a whole, but prefer the term ‘depiction’ for modalities other than text (Kosslyn et al., 2006), and will use it as a cover term for pictures, graphs, figures, and tables.

There is extensive research on different aspects of multimodality, from its reception and its effect on learning and recall (Holsanova et al., 2009; Mayer, 2009; Moreno and Mayer, 1999), to the particular aspects of layout, design, and links between different modes (Jeung et al., 1997). We focus on two particular modes: text and the depictive material that accompanies the text, and in particular on the links between the two. Issues of presentation and layout are beyond the scope of this work, but they are undoubtedly important, and have been studied elsewhere, for instance, in the work of Bateman and colleagues (Bateman, 2008b; Bateman et al., 2000, 2001). One interesting avenue to pursue in this direction is the effect of layout in understanding, and research using eye tracking technology seems most promising in that regard (Acartürk et al., 2008; Chu et al., 2009).

Captions for figures are also an element to bear in mind. Extensive research has shown that they are used to bridge text and images, and to process the individual parts of images (Elzer et al., 2005). However, and in order to focus the current research, captions have been ignored, and considered a whole together with the figure. We consider that the rhetorical relation is between the text and the figure as a whole, including the caption.

In order to understand and categorize the types of relations between figures and text, we are making use of Rhetorical Structure Theory (Mann and Thompson, 1988), which we discuss in the following section.

Coherence: Rhetorical relations between text and depictive material

In this section, we refer to the relations between depictive material and the portion of the document that introduces such figure, and propose that they stand in a coherence/rhetorical/argumentative relation. Previous research on the relations between parts of a document (Bateman et al., 2001; Kong, 2006) has established that the relations can be captured making use of the taxonomy of logico-semantic relations provided by Systemic Functional Linguistics (Halliday and Matthiessen, 2004), a school of linguistics that has long been associated with the study of multimodality (Baldry and Thibault, 2006; Bateman, 2008a; Kress and Van Leeuwen, 2006; Kress et al., 2001; Lemke, 1998; O’Halloran, 2008; Royce, 2007; Stenglin, 2009; Ventola et al., 2004). Other proposals for this relationship exist, most of them nicely summarized by Bateman (2008a) and Stöckl (2009), who also proposes rhetorical-logical relations for the connection between text and depiction.

In our work, we make use of the related relations in Rhetorical Structure Theory (RST) (Mann and Thompson, 1988). We postulate that figures, graphs, pictures, and tables are in a rhetorical relation with the text that they accompany. By rhetorical

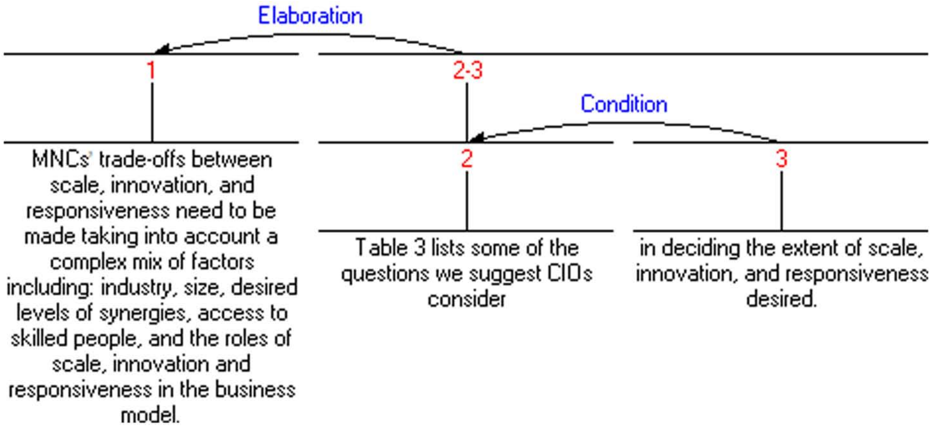

Elaboration
1
2-3
Condition
2
3
MNCs' trade-offs between scale, innovation, and responsiveness need to be made taking into account a complex mix of factors including: industry, size, desired levels of synergies, access to skilled people, and the roles of scale, innovation and responsiveness in the business model.
Table 3 lists some of the questions we suggest CIOs consider
in deciding the extent of scale, innovation, and responsiveness desired.

Figure 1. Sample RST analysis.

relation we mean a relation that establishes a coherent relation between the text and the depiction. The advantage of using RST relations is that they are well-defined and have been extensively tested.

In RST, texts are understood as coherent wholes, made up of parts that stand in rhetorical relations to each other. The parts are typically clauses or sentences, and the relations are those that capture the perceived coherence of most texts. Relations are, at the lower level of the text, closely related to the coordinate and subordinate relations of traditional grammar (Concession, Condition, Cause, Result), but they can become more abstract (Elaboration, Antithesis, Summary, Background), typically when the relation is between larger chunks of the text. Relations are recursively applied, that is, two clauses may stand in a Condition relation and, as a unit, they may become part of an Elaboration relation with another unit in the text, a unit that can in turn be as small as a clause or as large as a paragraph. Units are called spans, and they may be atomic (one clause or one sentence), or composed of other spans.

Another fundamental aspect of RST is the relative status of spans. In most relations, one part of the relation (one span) is considered to be the main part, and the other one is secondary. They are called nucleus and satellite, respectively, and are analogous to the main and subordinate clauses of traditional grammar. Relations between a nucleus and one or more satellites are hypotactic. Some relations are paratactic, consisting of two or more nuclei, in a relation similar to that between coordinated clauses.

To illustrate these concepts, we will use the text in Example (1), represented graphically in Figure 1, and taken from our corpus.[1]

(1) MNCs' trade-offs between scale, innovation, and responsiveness need to be made taking into account a complex mix of factors including: industry, size, desired levels of synergies, access to skilled people, and the roles of scale, innovation and responsiveness in the business model. Table 3 lists some of the questions we suggest CIOs consider in deciding the extent of scale, innovation, and responsiveness desired.

In the figure, we see the representation of main and secondary parts (nuclei and satellites) as vertical lines and arrows. Leaving aside the question of how the text was segmented (there could have been further segmentation in unit 1), we see that the three units are related to each other. First of all, unit 3 is secondary to unit 2, with the two in a conditional relation (*if you need to decide the extent of scale, innovation, and responsiveness, then look at the questions in Table 3*). Those two spans become one unit, which is an elaboration of the first unit. Further multimodal analysis would include the relation of this piece of text to Table 3 itself.

In RST, relations are defined in terms of intentions that lead authors to use a particular relation. Thus, an RST diagram such as the one provided in Figure 1 provides a view of some of the author's purposes or intentions for including each part.

Spans of texts can be related recursively by using relations, which are defined by constraints on the nucleus, on the satellite, and mainly by the effect that the writer wants to achieve on the reader. When labeling a particular relation, the analyst must make a plausibility judgment, based on the contextual situation and the (presumed or declared) intentions of the writer. That is, the analyst judges whether it is plausible that the writer had such-and-such intentions or desired to obtain such-and-such effects when creating the text.

Space precludes a more extensive discussion of the theory itself. More detail can be found in the original paper on RST (Mann and Thompson, 1988), a more recent overview (Taboada and Mann, 2006a, 2006b), or the RST web site (Mann and Taboada, 2010). The types of general relations that we shall be using will be hopefully self-explanatory from their labels (Evidence, Condition, Elaboration, etc.), but full definitions are available on the RST web site.

Most research in RST has examined texts and the relations contained therein. There have been applications in spoken discourse as well (for a review of applications, see Taboada and Mann, 2006a), but not much research has addressed the connection between different types of components, for instance, different modes in a multimodal document. One exception is the work of Bateman and colleagues (Bateman, 2008b; Bateman et al., 2000, 2001; Delin and Bateman, 2002), where rhetorical relations are annotated for entire documents, and figures and other graphic material are found to be in rhetorical relations with other text elements. Bateman and colleagues have also brought in the idea of genre, the type of text under consideration, and how that affects both the layout and the relationship of text and illustrations. Other work examining multimodal documents as coherent wholes that may be built out of rhetorical structures includes the work of André (André, 1995; André and Rist, 1996), Feiner and McKeown (1993), or Stock and others (Carenini et al., 1993; Stock, 1993). In most of this work there was a computational component, with the intention of generating multimodal documents.

Corpus

This is a preliminary and mostly qualitative study, and thus our corpus size is moderate. We have studied three different types of texts: newspapers (print and/or online); scientific magazine articles; and scientific articles. The three types of texts were chosen because they were easily accessible and provide a range of comparison from the expertise point of

Table 1. Composition of the corpus.

	Page count	Pictures	Figures, graphs	Tables	Maps
Communications of the ACM	736	35	189	21	–
Computational Linguistics	645	–	137	139	–
New York Times	62	51	–	–	7

view, since we can assume that newspaper layout is generally produced by experts, whereas scientific articles are created out of intuition and exposure to the genre, but usually not by people who have expertise in layout. Magazine articles combine experts of both types: writers who have some scientific knowledge, but who devote themselves to dissemination of scientific knowledge.[2]

Composition of the corpus:

- *Communications of the ACM* (http://cacm.acm.org/), henceforth CACM:
  - ○ all issues for January–June 2010 (six issues).
- *Journal of Computational Linguistics* (http://www.mitpressjournals.org/loi/coli), henceforth CL:
  - ○ all issues for 2009 (four issues);
  - ○ of those, only main articles (about four per issue).
- *New York Times*, henceforth NYT:
  - ○ articles from the ProQuest edition of the *New York Times* (available from the Simon Fraser University Library system via the ProQuest service):
    - articles between 1 January and 31 December 2006 (most recent year available within ProQuest);
    - articles extracted with the search terms ‘oil spill’, a total of 54 documents (more were extracted, but discarded if they did not contain the words in the general sense of spilling of oil in the processes of drilling).[3]

The corpus as a whole contains about 1500 pages of material, with 579 instances of depictive material (pictures, figures, tables, and maps). The annotation process involved identifying, for each issue and article, what types of depictive material the articles contain. We then determined the type of relation between depictive material and text. In this section, we provide a summary of the numbers of articles annotated, and the types of depictive material by genre. Table 1 summarizes the counts for each component in the corpus.

Analysis of rhetorical relations

In this section, we provide a summary of the type of relations and the signaling between depiction and text, again divided into the three corpora studied.

A few notes on methodology are in order here. First of all, we see the relation between depiction and text as happening at different levels, and affecting different types of spans. In some cases, the depictive material stands in a relationship to the entire text; such is the

Table 2. Rhetorical relations in the CACM corpus.

	Pictures	Figures	Tables	Total
Elaboration	7	134	10	151
Enablement	–	1	–	1
Evidence	2	48	9	59
Motivation	25	–	–	25
Preparation	1	–	–	1
Restatement	–	3	–	3
Summary	–	3	2	5
Total	35	189	21	245

case with some of the Background and Preparation pictures in the *New York Times* articles. The relationship is akin to that of a title to the body of the article. In other cases, however, the relationship is between one paragraph that introduces and describes the depiction and the depiction itself. In some other cases, the relation is much more local, linking a single span of text (clause or sentence) and the depictive material, which together form a multimodal cluster (Baldry and Thibault, 2006).

Second, most of the relations found were hypotactic, that is, they have two unequal members, one the nucleus and one the satellite. The vast majority of the examples have the depictive material as satellite. In the analysis, we found that the depictive material was almost always secondary to the information presented in the text. To ensure the accuracy of the annotation, we used the deletion test (Mann and Thompson, 1988; Marcu et al., 1999): whichever one of the text or the depictive material can be deleted, that one is the satellite. As we mention in the Discussion section, we found that most depictive material acts as illustration, in Barthes' (1977) terms, certainly a characteristic of the genres studied.

In general, the annotation process involved reading the entire text, paying attention to depictive material when the author(s) included a deictic reference to it, or when it appeared in the layout, in the cases where there was no reference. A note was made of the most likely span of text that had a relation to the depictive material. Upon a second reading, nuclearity and type of relation were decided.

Communications of the ACM

Table 2 provides a summary of the relations found within the CACM corpus, broken down by type of depictive material. As is clear from the table, Elaboration relations are the most frequent. This is because the depictive material, especially the figures, takes the material presented on the text further and provides additional information. This can take many different forms, as suggested in the RST definition of Elaboration, where the nucleus and the satellite (in this case, the graphical material) can be in one of the following relations: set-member, abstraction-instance, whole-part, process-step, object-attribute, generalization-specific.

By type of material, pictures are most often used for motivation, figures for elaboration, and tables for evidence. We discuss next some examples of these relations.

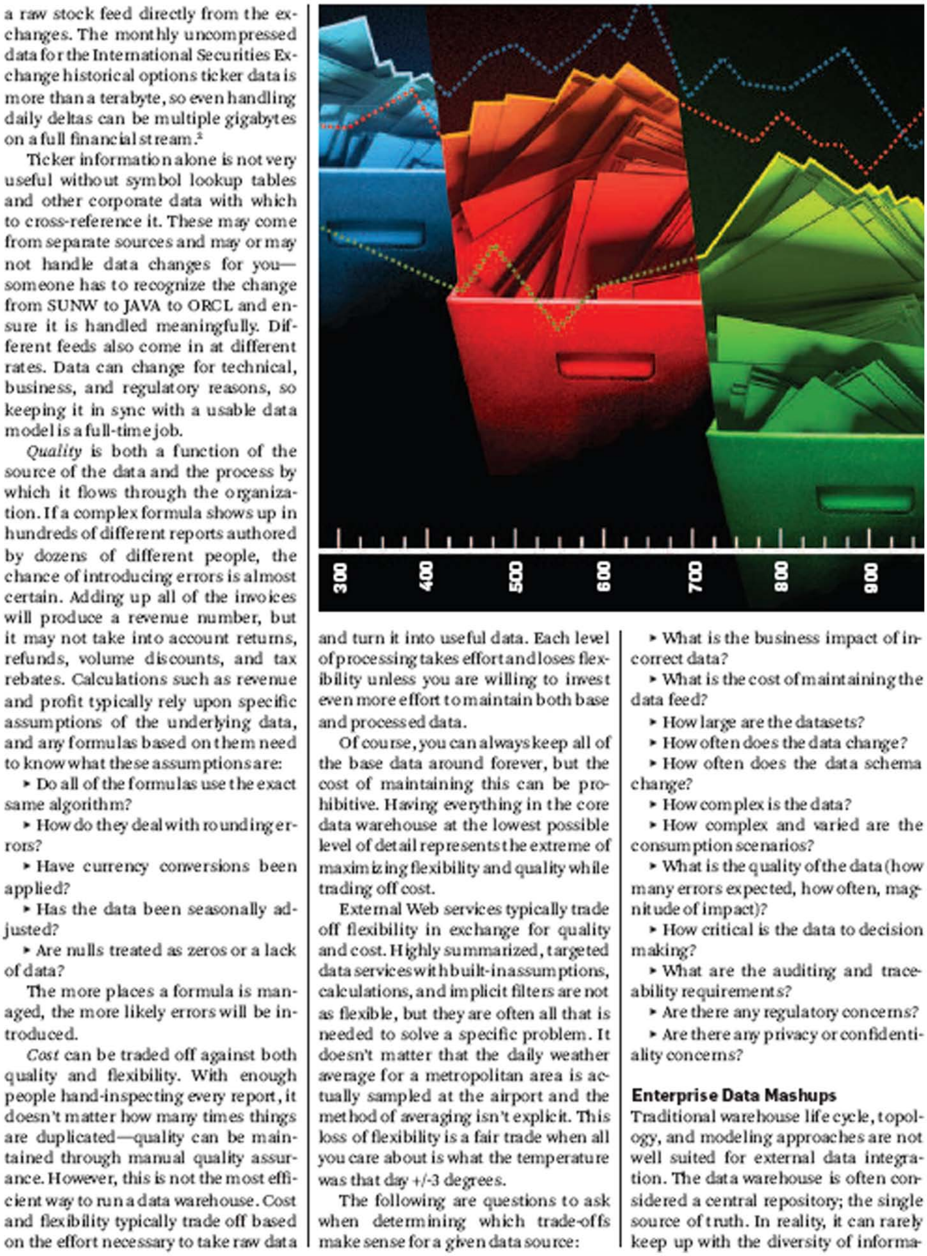
We saw, in the statistics presented earlier, that the CACM corpus had a relatively high number of pictures, mostly presented without captions. These tend to serve as illustrations to the text, and are often abstract representations relating to the content of the article. It was difficult to decide what rhetorical relation to assign to this picture–text relationship, as there was no signaling to make that connection more transparent. In the cases where the picture appeared towards the beginning of the article, it was straightforward to label them as Preparation. The Preparation relation states that the satellite precedes the nucleus and tends to make the reader more ready, interested, or oriented for reading the nucleus (in our case, the text itself). We will see later on that in the NYT corpus we have relaxed the precedence constraint, and have included cases where the picture appears towards the beginning of the article, not necessarily preceding the text, but in a salient position (see Figure 4). In the CACM corpus, we find some instances of pictures as Preparation. We also find, however, instances of similar pictures that appear later on in the article, sometimes on pages 2 or 3 in a multi-page article. We can hardly label those as Preparation to the entire article, as we have done with pictures at the beginning of the article. We decided, in those cases, to label them as Motivation. The Effect of Motivation, in the RST definition, is that the reader's desire to perform the action in the nucleus is increased. We considered that the picture serves as motivation for the reader to continue reading the text.

An example is Figure 2, a small-scale rendition of a page from the CACM corpus ('Other people's data', January 2010, p. 55). The article deals with the issues surrounding storage of large amounts of data. The picture represents the second page of the article, and depicts three file cabinet drawers overflowing with documents, and a ruler with a seemingly arbitrary scale at the bottom. The picture does not prepare us to read the text, or elaborate on it. It simply motivates the reader to continue reading and maybe helps to break the flow of the page, so as to avoid a full page of text. The placement of the picture is consistent with research (Garcia and Stark, 1991; Holsanova et al., 2006) that shows that readers scan text and start reading at certain entry points. Entry points can be headlines, boxes, or any other elements that break the flow, and pictures and graphics have been found to be entry points.

The original RST definition of Restatement (Mann and Thompson, 1988) proposed it as a hypotactic relation, with a satellite as a restatement of a nucleus of comparable bulk, but of more central importance. The issue of central importance is much more difficult to decide when one of the spans is in a different mode. For that reason, the GeM project (Bateman et al., 2007; Henschel, 2003) proposed a new type of multinuclear Restatement relation, already present in the RST Discourse Treebank corpus annotations (Carlson et al., 2002; Marcu, 1999). In our corpus, figures restate, in a different mode, the information already presented in the text (or, vice versa, the text restates the information provided by the figure). Restatements are one of the few cases where the depictive material is not a satellite to the text.

Computational Linguistics

The articles in *Computational Linguistics* (a total of 18 studied, of which two had no tables or figures) contain a less varied distribution of rhetorical relations. As Table 3 shows, the majority of the figures and tables stand in an Elaboration and Evidence relation to the text.



a raw stock feed directly from the exchanges. The monthly uncompressed data for the International Securities Exchange historical options ticker data is more than a terabyte, so even handling daily deltas can be multiple gigabytes on a full financial stream.[2]

Ticker information alone is not very useful without symbol lookup tables and other corporate data with which to cross-reference it. These may come from separate sources and may or may not handle data changes for you—someone has to recognize the change from SUNW to JAVA to ORCL and ensure it is handled meaningfully. Different feeds also come in at different rates. Data can change for technical, business, and regulatory reasons, so keeping it in sync with a usable data model is a full-time job.

Quality is both a function of the source of the data and the process by which it flows through the organization. If a complex formula shows up in hundreds of different reports authored by dozens of different people, the chance of introducing errors is almost certain. Adding up all of the invoices will produce a revenue number, but it may not take into account returns, refunds, volume discounts, and tax rebates. Calculations such as revenue and profit typically rely upon specific assumptions of the underlying data, and any formulas based on them need to know what these assumptions are:

- Do all of the formulas use the exact same algorithm?
- How do they deal with rounding errors?
- Have currency conversions been applied?
- Has the data been seasonally adjusted?
- Are nulls treated as zeros or a lack of data?

The more places a formula is managed, the more likely errors will be introduced.

Cost can be traded off against both quality and flexibility. With enough people hand-inspecting every report, it doesn't matter how many times things are duplicated—quality can be maintained through manual quality assurance. However, this is not the most efficient way to run a data warehouse. Cost and flexibility typically trade off based on the effort necessary to take raw data and turn it into useful data. Each level of processing takes effort and loses flexibility unless you are willing to invest even more effort to maintain both base and processed data.

Of course, you can always keep all of the base data around forever, but the cost of maintaining this can be prohibitive. Having everything in the core data warehouse at the lowest possible level of detail represents the extreme of maximizing flexibility and quality while trading off cost.

External Web services typically trade off flexibility in exchange for quality and cost. Highly summarized, targeted data services with built-in assumptions, calculations, and implicit filters are not as flexible, but they are often all that is needed to solve a specific problem. It doesn't matter that the daily weather average for a metropolitan area is actually sampled at the airport and the method of averaging isn't explicit. This loss of flexibility is a fair trade when all you care about is what the temperature was that day +/-3 degrees.

The following are questions to ask when determining which trade-offs make sense for a given data source:

- What is the business impact of incorrect data?
- What is the cost of maintaining the data feed?
- How large are the datasets?
- How often does the data change?
- How often does the data schema change?
- How complex is the data?
- How complex and varied are the consumption scenarios?
- What is the quality of the data (how many errors expected, how often, magnitude of impact)?
- How critical is the data to decision making?
- What are the auditing and traceability requirements?
- Are there any regulatory concerns?
- Are there any privacy or confidentiality concerns?

Enterprise Data Mashups

Traditional warehouse life cycle, topology, and modeling approaches are not well suited for external data integration. The data warehouse is often considered a central repository; the single source of truth. In reality, it can rarely keep up with the diversity of informa-

Figure 2. Picture as Motivation.

As in the other corpora, tables most often have an Evidence function, showing data that the authors believe will bolster the claims made in the text. Some of the tables also have a simple Elaboration relation, providing additional quantitative material that would be much more difficult to read in prose form.

Table 3. Rhetorical relations in the CL corpus.

	Figures	Tables	Total
Elaboration	120	67	187
Evidence	16	72	88
Preparation	1	–	1
Total	137	139	276

Table 4. Rhetorical relations in the NYT corpus.

	Pictures	Maps	Total
Circumstance	1	7	8
Contrast	1	–	1
Elaboration	19	–	19
Evidence	9	–	9
Motivation	1	–	1
Preparation	20	–	20
Total	51	7	58

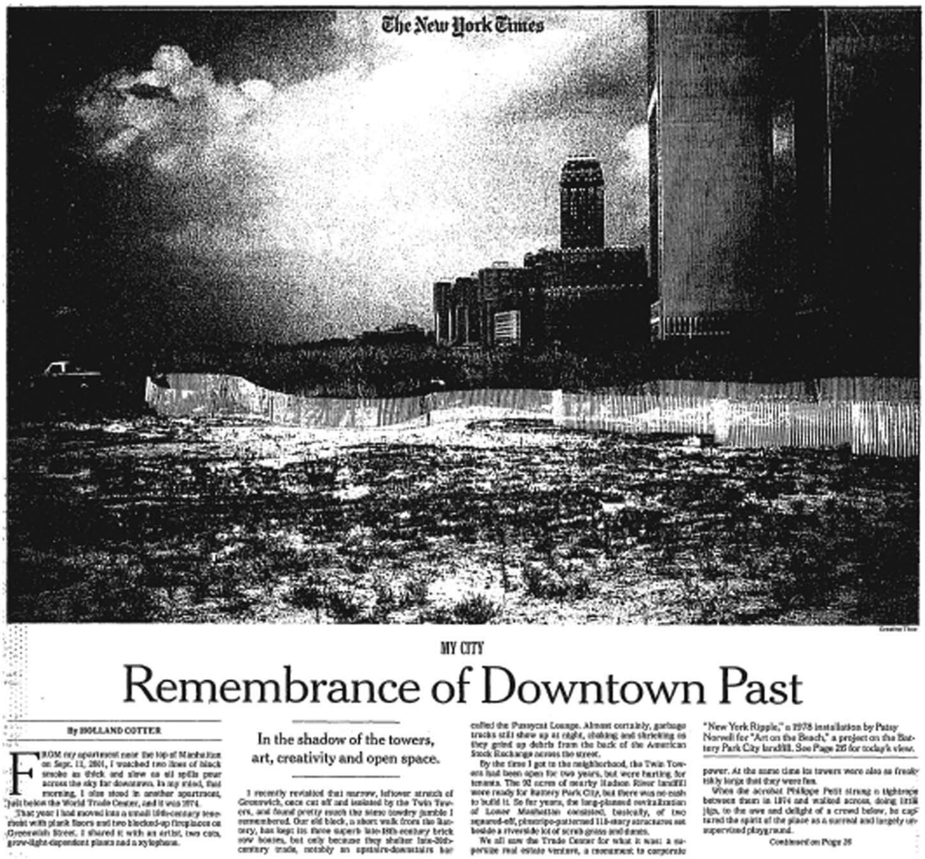
In terms of distribution across the stages of the research article genre, Elaboration relations tend to occur at the beginning of the article, whereas Evidence appears towards the middle and end, in the results section or sections.

The New York Times

Table 4 summarizes our analysis of the pictures and maps in the NYT corpus. We find that maps have the unique function of setting up a framework, and then relate to the text in a Circumstance relation (see below for an example). With respect to pictures, they have two main functions: Elaboration and Preparation. We discuss some examples below.

One of the characteristics of the pictures in NYT is that they function as Preparation to the rest of the story. For example, a picture in ‘Remembrance of downtown past’ (1 September 2006, p. E21) portrays the landscape around the former location of the World Trade Center in New York in 1978. The article is a personal reminiscence of that space in the 1970s and 1980s, interspersed with factual information about the art scene in that location. The picture serves as preparation for the rest of the story, a prompt that the location of the World Trade Center is still barren after the terrorist attacks of 11 September 2001, and a pathway into the time before the World Trade Center was erected, when the land was also barren. Figure 3 is a low-quality[4] rendition of the beginning of the article, with a picture that dominates the page (this part of the article occupies the first half of the newspaper page). The picture, according to the caption, is a 1978 art installation on Battery Park. Part of the World Trade Center towers can be seen in the background, to the right.

A large number of the figures and pictures in NYT, as in CACM, serve as Preparation for the rest of the text. Eye-tracking studies of newspaper reading show that pictures, especially large ones, are processed first, before the text is read.[5]



The New York Times

MY CITY

Remembrance of Downtown Past

By HOLLAND COTTER

In the shadow of the towers, art, creativity and open space.

“New York Ripple,” a 1978 installation by Patsy Norvell for “Art on the Beach,” a project on the Battery Park City landfill. See Page 26 for today's view.

FROM my apartment near the top of Manhattan on Sept. 11, 2001, I watched two lines of black smoke as thick and slow as oil spills pour across the sky far downtown. In my mind, that morning, I also stood in another apartment, just below the World Trade Center, and it was 1974.

That year I had moved into a small 19th-century tenement with plank floors and two blocked-up fireplaces on Greenwich Street. I shared it with an artist, two cats, grow-light-dependent plants and a xylophone.

I recently revisited that narrow, leftover stretch of Greenwich, once cut off and isolated by the Twin Towers, and found pretty much the same tawdry jumble I remembered. Our old block, a short walk from the Battery, has kept its three superb late-18th-century brick row houses, but only because they shelter late-20th-century trade, notably an upstairs-downstairs bar called the Pussycat Lounge. Almost certainly, garbage trucks still show up at night, choking and shrieking as they grind up debris from the back of the American Stock Exchange across the street.

By the time I got to the neighborhood, the Twin Towers had been open for two years, but were hurting for tenants. The 92 acres of nearby Hudson River landfill were ready for Battery Park City, but there was no cash to build it. So for years, the long-planned revitalization of Lower Manhattan consisted, basically, of two squared-off, pinstripe-patterned 110-story structures set beside a riverside lot of scrub grass and dunes.

We all saw the Trade Center for what it was: a supersize real estate venture, a monument to corporate power. At the same time its towers were also so freakishly large that they were fun.

When the acrobat Philippe Petit strung a tightrope between them in 1974 and walked across, doing little jigs, to the awe and delight of a crowd below, he captured the spirit of the place as a surreal and largely unsupervised playground.

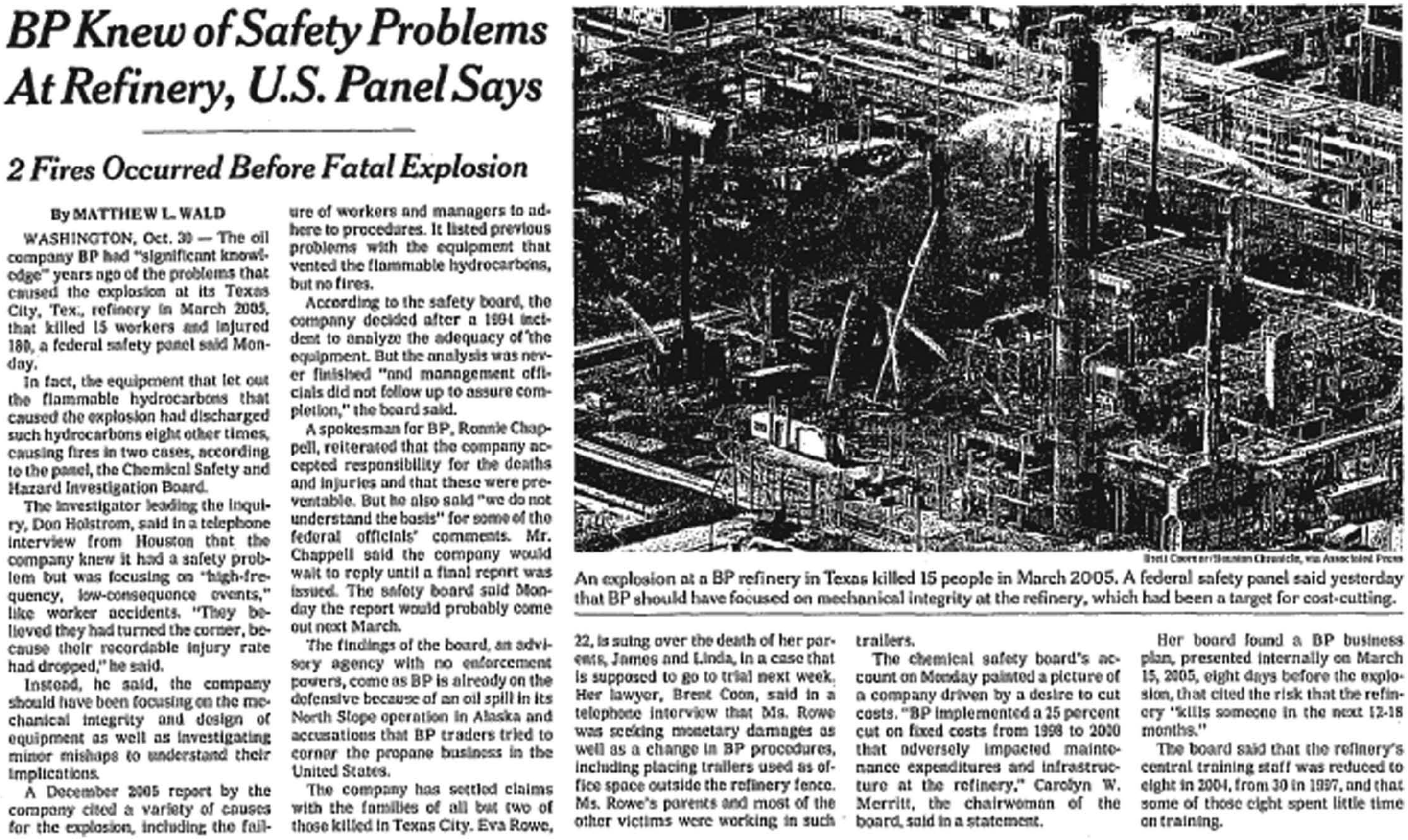
Continued on Page 26

Figure 3. Beginning of *New York Times* article, 1 September 2006.

One issue, however, is that the preparatory pictures are not always *before* the text, as the standard definition for Preparation states. The constraints on the relation read: ‘S precedes N in the text; S tends to make R more ready, interested or oriented for reading N’ (Mann and Taboada, 2010). We have decided to waive the precedence requirement, as the salience of pictures has, in a sense, certain precedence: the picture is seen before the text (or at least the body of the text) is read. In all cases where we labeled a picture as Preparation, the picture was towards the beginning of the article, that is, never on a continuation page. Such is the case in Figure 4 (‘BP knew of safety problems at refinery, U.S. panel says’, 31 October 2006, p. C3). The picture, although to the right of the beginning of the text, is quite prominent, and larger than either the body text to its left or the heading.

We also find many pictures of the people involved in the story, which were also labeled as Preparation, since they seem to prepare and motivate the reader, by reading the article, to find out more about the people portrayed in the picture.

There is probably much more to the placement of the pictures than we have space to discuss here. In general, we have avoided discussing layout, as it would make the analysis



BP Knew of Safety Problems At Refinery, U.S. Panel Says

2 Fires Occurred Before Fatal Explosion

By MATTHEW L. WALD

WASHINGTON, Oct. 30 — The oil company BP had "significant knowledge" years ago of the problems that caused the explosion at its Texas City, Tex., refinery in March 2005, that killed 15 workers and injured 180, a federal safety panel said Monday.

In fact, the equipment that let out the flammable hydrocarbons that caused the explosion had discharged such hydrocarbons eight other times, causing fires in two cases, according to the panel, the Chemical Safety and Hazard Investigation Board.

The investigator leading the inquiry, Don Holstrom, said in a telephone interview from Houston that the company knew it had a safety problem but was focusing on "high-frequency, low-consequence events," like worker accidents. "They believed they had turned the corner, because their recordable injury rate had dropped," he said.

Instead, he said, the company should have been focusing on the mechanical integrity and design of equipment as well as investigating minor mishaps to understand their implications.

A December 2005 report by the company cited a variety of causes for the explosion, including the failure of workers and managers to adhere to procedures. It listed previous problems with the equipment that vented the flammable hydrocarbons, but no fires.

According to the safety board, the company decided after a 1994 incident to analyze the adequacy of the equipment. But the analysis was never finished "and management officials did not follow up to assure completion," the board said.

A spokesman for BP, Ronnie Chappell, reiterated that the company accepted responsibility for the deaths and injuries and that these were preventable. But he also said "we do not understand the basis" for some of the federal officials' comments. Mr. Chappell said the company would wait to reply until a final report was issued. The safety board said Monday the report would probably come out next March.

The findings of the board, an advisory agency with no enforcement powers, come as BP is already on the defensive because of an oil spill in its North Slope operation in Alaska and accusations that BP traders tried to corner the propane business in the United States.

The company has settled claims with the families of all but two of those killed in Texas City. Eva Rowe, 22, is suing over the death of her parents, James and Linda, in a case that is supposed to go to trial next week. Her lawyer, Brent Coon, said in a telephone interview that Ms. Rowe was seeking monetary damages as well as a change in BP procedures, including placing trailers used as office space outside the refinery fence. Ms. Rowe's parents and most of the other victims were working in such trailers.

The chemical safety board's account on Monday painted a picture of a company driven by a desire to cut costs. "BP implemented a 25 percent cut on fixed costs from 1998 to 2000 that adversely impacted maintenance expenditures and infrastructure at the refinery," Carolyn W. Merritt, the chairwoman of the board, said in a statement.

Her board found a BP business plan, presented internally on March 15, 2005, eight days before the explosion, that cited the risk that the refinery "kills someone in the next 12-18 months."

The board said that the refinery's central training staff was reduced to eight in 2004, from 30 in 1997, and that some of those eight spent little time on training.

Brett Coomer/Houston Chronicle, via Associated Press

An explosion at a BP refinery in Texas killed 15 people in March 2005. A federal safety panel said yesterday that BP should have focused on mechanical integrity at the refinery, which had been a target for cost-cutting.

Figure 4. Picture as Preparation.

and discussion more complex. We would simply like to point out, in the discussion of the Preparation relation and the placement of pictures in a Preparation relation, that one could analyze the two spans of the relation (text and picture) with respect to placement on the page, and using Kress and Van Leeuwen's (2006) Given–New/Ideal–Real distinctions, where Given is presented to the left of New, and Ideal on top of Real. The pictures at the beginning of the text act as Given material, with the text to be read as New.

Another interesting characteristic of NYT articles is the presence of maps. Maps allow the reader to locate a city, area or country that the reader may not be familiar with. Quite often, they provide additional information, a framework or grid to locate information presented in the article. In the same article that we discussed earlier ('Remembrance of downtown past', 1 September 2006, p. E21), a map identifies the locations discussed in the article, such as the location of the World Trade Center, Battery Park, and other, perhaps less familiar locations such as Herman Melville's birthplace and other sites of artistic interest. This type of relation we labeled as Circumstance, with the map as satellite. In RST terms, the satellite of a Circumstance relation sets the framework within which the reader is to understand the nucleus. The framework may be temporal, spatial, or of a similar type. Most maps clearly provide a spatial framework.

Reliabilty study

We have, so far, presented results of our analysis, and would like in this section to show that the analyses are reliable. RST analyses have been argued to be subjective, and the judgment of the analyst has always been part of how the theory is applied (Mann and Thompson, 1988). Critics may argue that transferring RST to multimodal documents could make it even more subjective. In order to demonstrate that our analyses are reproducible, we conducted a reliability study.

For the study, we selected several documents: one from the CACM, one from CL, and three NYT articles. In total, they contain 56 different relations, about 10% of the relations in the corpus. We asked an independent analyst, trained in RST analyses, to label the relationship between text and depiction following some basic guidelines, mostly the RST definitions, with the additional information that we have presented here, namely that depictive material tends to enter into only a handful of relations with text, and that nuclearity can be tested with the deletion test.

Out of the 56 relations, the analyst agreed with our analyses in 41 of the cases, that is, 73.21% of the time. Percentage agreement can be misleading, since it is not sensitive to the range of options (in our case, an analyst can choose one of the 30 RST relations). To account for the fact that multiple categories are possible, we calculated agreement using Cohen's kappa, for nominal data and unweighed (Cohen, 1960). The kappa value for the 56 examples, labeled with one of nine relations (the actual relations used by either us or the analyst) is 0.616. This is considered 'substantial' agreement according to Landis and Koch (1977), and much higher than expected by chance. We can conclude that the methodology that we followed is reproducible, given a trained RST analyst.

Disagreements between our original analyses and those of the third analyst had mostly to do with the function of tables, and whether they were considered to be in Elaboration or Evidence relations with the text. Similarly, Preparation and Background were also annotated differently by the analyst. These are natural grey areas in the analysis, and showed intra-annotator consistency (i.e. the analyst tended to use more Evidence for tables, whereas the original analyses tended to use Elaboration).

Discussion

We can summarize the findings of our analysis as follows:

- figures tend to elaborate on the text;
- tables tend to provide evidence for claims and proposals in the text;
- pictures provide background and motivation for the information in the text.

Pictures, graphs, and diagrams are most often subsidiary to the text, whereas tables provide either additional information or data demonstrating the validity of the methods and experiments in the text.

This secondary nature of illustrations is a result of the types of texts studied. In all three cases, the genre is one where words are most important and communication of verbal information is primary. The genres studied are examples of a text-flow semiotic mode (Bateman, 2009), where the most important information is conveyed in the text, and the depictive material is used to support the text.

We found that, from the range of RST relations typically used (25–30 in most applications of RST), we used only a handful, and that those tended to be presentational, that is, relations that facilitate the presentation process and are internal to the text, as opposed to subject matter relations, which express parts of the subject matter of the text and reflect the state of affairs outside the text. We believe this is because of the wordy, text-flow characteristic of the documents. Liu and O'Halloran (2009), in analyzing news articles, also used only four of the possible conjunctive relations: Comparison, Addition, Consequence, and Time. Other research in document design (e.g. Schriver, 1997) seems to point to a limited number of functions that depictive material has with respect to the text.

In the rest of this section, we would like to discuss more extensively three particular aspects of the analysis, and their implications for further studies of multimodal documents: the consequences of the creation process, the nature of the Elaboration relation, and the possibility that there are multiple relations connecting the same depiction to different parts of the text.

The creation process

The creation process in some of the documents studied is such that there may not be a single author, or even the same author involved in all parts of the document. In particular, in newspaper articles (and possibly some articles in CACM), one author may write the text, and a graphic artist may insert the picture, map, or graph, while an editor oversees the process and final product. It is also the case that many of the CL articles have multiple authors. RST assumes a single writer for the whole text, with particular intentions and effects that he or she wants to achieve. The presence of possible multiple authors

Table 5. Summary, RST relations in entire corpus.

	Pictures	Figures	Tables	Maps	Total	%
Elaboration	26	254	77	0	357	61.66
Evidence	11	64	81	0	156	26.94
Motivation	26	0	0	0	26	4.49
Preparation	21	1	0	0	22	3.80
Circumstance	1	0	0	7	8	1.38
Summary	0	3	2	0	5	0.86
Restatement	0	3	0	0	3	0.52
Enablement	0	1	0	0	1	0.17
Contrast	1	0	0	0	1	0.17
Total	86	326	160	7	579	

complicates a straightforward set of constraints and desired effects that every RST relation is supposed to have.

In our analyses, we have assumed the usual situation of a single creator for the document, or rather, a single 'mind'. One could consider that the authors/contributors have certain purposes and effects they, as a group, want to achieve with the multimodal document. This is the view that we have taken, but we also understand that more fine-grained analyses would have to study the creation process in order to understand the contribution of different authors and contributors to the final product.

The Elaboration relation

The second aspect of the analysis that we would like to discuss is the predominance of the Elaboration relation. Taken together, all the corpora contain 579 relations. Of those, 61.66% are Elaboration relations (see Table 5).

The main reason that such a high proportion of the relations are Elaboration is that the documents analyzed convey most of their information through the text. Figures, pictures, tables, and maps are brought in to support information provided in the text. That type of relation is, in essence, an Elaboration relation.

Example (2) and Figure 5 show an instance of an Elaboration relation from the CL corpus.[6] The text introduces the figure as presenting the overall architecture, and refers only to one part of the figure (MCUBE). The figure is the whole of the system, whereas the text contains a reference to only a part of it.

(2) The underlying architecture that supports MATCH consists of a series of re-usable components which communicate over IP through a facilitator (MCUBE) (Figure 5).

Similarly, in Example (3) and Figure 6, from the CACM corpus,[7] we see an Elaboration where the table provides additional information to the material presented in the text. Although the text uses the keyword 'summarizes', the summary is of ideas external to the text (i.e. held by the authors), whereas the table simply provides specific details to the generalization that there are critical obstacles.

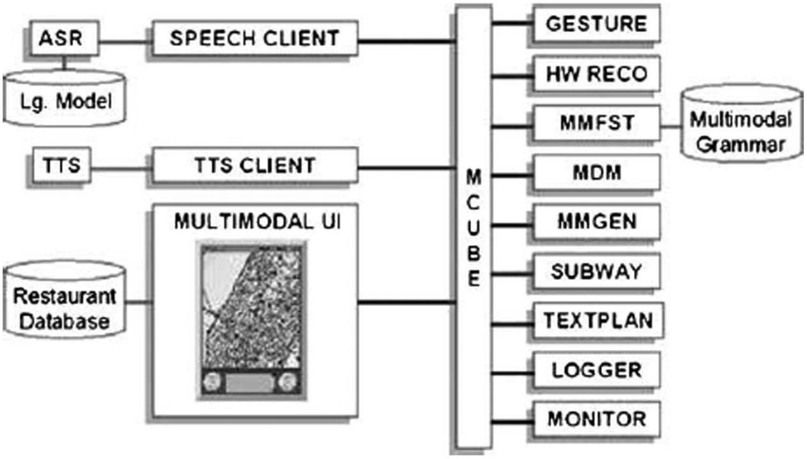


Figure 5. A diagram in an Elaboration relation.

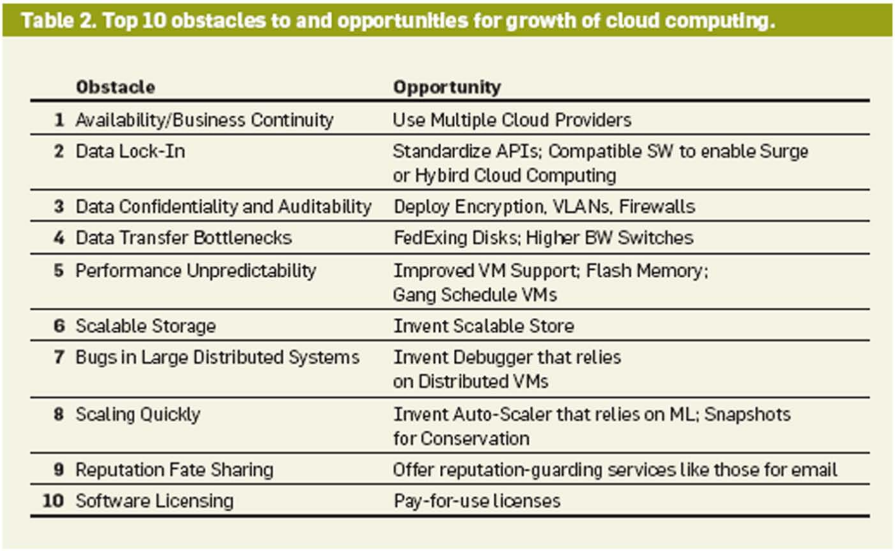


Table 2. Top 10 obstacles to and opportunities for growth of cloud computing.

	Obstacle	Opportunity
1	Availability/Business Continuity	Use Multiple Cloud Providers
2	Data Lock-In	Standardize APIs; Compatible SW to enable Surge or Hybird Cloud Computing
3	Data Confidentiality and Auditability	Deploy Encryption, VLANs, Firewalls
4	Data Transfer Bottlenecks	FedExing Disks; Higher BW Switches
5	Performance Unpredictability	Improved VM Support; Flash Memory; Gang Schedule VMs
6	Scalable Storage	Invent Scalable Store
7	Bugs in Large Distributed Systems	Invent Debugger that relies on Distributed VMs
8	Scaling Quickly	Invent Auto-Scaler that relies on ML; Snapshots for Conservation
9	Reputation Fate Sharing	Offer reputation-guarding services like those for email
10	Software Licensing	Pay-for-use licenses

Figure 6. Table in an Elaboration relation.

(3) Table 2 summarizes our ranked list of critical obstacles to growth of cloud computing. The first three affect adoption, the next five affect growth, and the last two are policy and business obstacles. Each obstacle is paired with an opportunity to overcome that obstacle, ranging from product development to research projects.

In many cases, the depiction is a specific member, instance or part of a general case discussed in the text. In other cases, the relationship works in the other direction (thus reversing the status of the depiction from satellite to nucleus): the depiction presents an entire process, of which only one or two steps are discussed in the text. In a sense, most Elaboration relations are specific cases of the general *illustration* function that Barthes (1977) defined.

The Elaboration relation has been criticized as not being a true rhetorical relation (a relation of coherence), rather a relation of cohesion, that is, a relation among entities in the discourse. Knott et al. (2001) discuss one of the cases of Elaboration, object–attribute Elaboration, and conclude that it is different in nature from other types of Elaboration and from other relations, since it is a relation between entities: one of the spans contains an attribute on an entity present in the other span. All other RST relations are relations between propositions, thus making (object–attribute) Elaboration a global relation, linking entities that are within focus spaces in the discourse, of the type proposed by Grosz and Sidner (1986). Knott et al.'s proposal involves removing object–attribute Elaboration from the set of RST relations. It is not clear, however, whether this applies to all types of Elaboration.

In our corpus, there are few entity-based Elaboration relations between text and figures. Most of the Elaboration cases are abstraction-instance, whole-part, process-step, or generalization-specific. It does seem, however, that labeling a text–graphic relation as Elaboration in over 60% of the cases provides little information about the type of relation holding. A proposal for multimodal documents would be to incorporate the additional labels to the Elaboration label, thus specifying what type of Elaboration relation holds.

Discussions by Baldry and Thibault (Baldry and Thibault, 2006; Thibault, 1997) and Bateman (2008b) also deal with other problems with elaboration relations. Bateman points out (2008b) that, for instance, the relationship between a depiction and labels identifying different parts of it could be classified as Elaboration, but one that holds between elements that would not be naturally considered units in RST, since they are fragments. Another aspect worth mentioning in the context of the Elaboration relation is its relationship to cohesive links, that is, connections between elements that are related through relations such as reference, synonymy, or hypernymy. The connections may take place within modality, such as the words *bird* and *gannet*, but also across, such as the phrase *off the NW coast of Europe* and a map depicting that area (Bateman, 2008a).

Multiple relations

We have, in our analyses, always assumed that there is some relation between text and graphical material. We have, furthermore, assumed that such relation is unique, between a particular portion of a text (the scope of which may be under-defined) and the corresponding illustration. We would like now to explore the possibility that graphical material stands in multiple relationships to the text in the document.

RST assumes a linear order of processing. That is often the case in reading, and certainly so in speech, although one can look back in text, and also exploit some of the benefits of echoic memory for discontinuous processing of speech. In general, however, we follow the flow of reading or speaking. In multimodal documents, on the other hand, linear processing cannot be assumed. In a multimodal document, the text is processed in linear order, but with 'excursions' to the graphical material (Holmqvist et al., 2003;

Lewenstein et al., 2000; Stark Adam et al., 2007), and we know that, upon first contact with a multimodal document, we may scan back and forth, rather than read (Kress and Van Leeuwen, 1998). It may well be the case that a depiction is examined multiple times as the document is read. If so, then the relation between depiction and text may be a different one in each of those situations.

Let us consider one example, from a paper in CL.[8] A table, reproduced in Figure 7, is presented towards the beginning of the article, as Preparation to make the reader more able to understand the rest of the text. In (4), we reproduce the text used to justify the presence of the table. The text is in the same section as the table.

(4) Because this section combines notation from different theoretical frameworks (in particular, from formal semantics and statistical time-series modeling), a notation summary is provided in Table 1.

At this point in the article, some of the terminology and notations in the table have already been introduced. The table could then serve as a Summary of definitions presented earlier (in Section 3 in the article), Preparation for the terminology presented in Section 4, and a Summary throughout the rest of the paper. It is quite likely that the reader will flip back to this table as he or she reads the paper, then establishing new links between text and table.

Another example of potential multiple relations is the already discussed article 'Remembrance of downtown past', from the *New York Times* corpus. As we mentioned earlier, the picture at the beginning of the article stands in a Preparation relation to the rest of the article: the reader, by looking at the photo, is more prepared to understand that the article is about a time in the past in New York City. However, as the article progresses, we also find that it is not only about a past time, but also, more specifically, about the art scene at that time. The fourth paragraph in the article is as follows:

(5) By the time I got to the neighborhood, the Twin Towers had been open for two years, but were hunting for tenants. The 92 acres of nearby Hudson River landfill were ready for Battery Park City, but there was no cash to build it. So for years, the long-planned revitalization of Lower Manhattan consisted, basically, of two squared-off, pinstripe-patterned 110-story structures set beside a riverside lot of scrub grass and dunes.

It is plausible that, at the point that 'a riverside lot of scrub grass and dunes' is mentioned, the reader looks back at the photo (right on top of this paragraph) and establishes a new relationship between text and photo, this time one of Elaboration. The photo elaborates, pictorially, what the text is describing.

On the next page, we read the following paragraph:

(6) Creative Time, led by Anita Contini, struck gold when she persuaded the Battery Park City Authority to let her use its empty landfill for art events. The location, which we were already using for sunbathing and kite flying, was stark, stunning and slightly eerie. [. . .]

The justification for the photo at the beginning seems now complete, with a new Elaboration relation. At first, we just see the photo as Preparation that, together with the

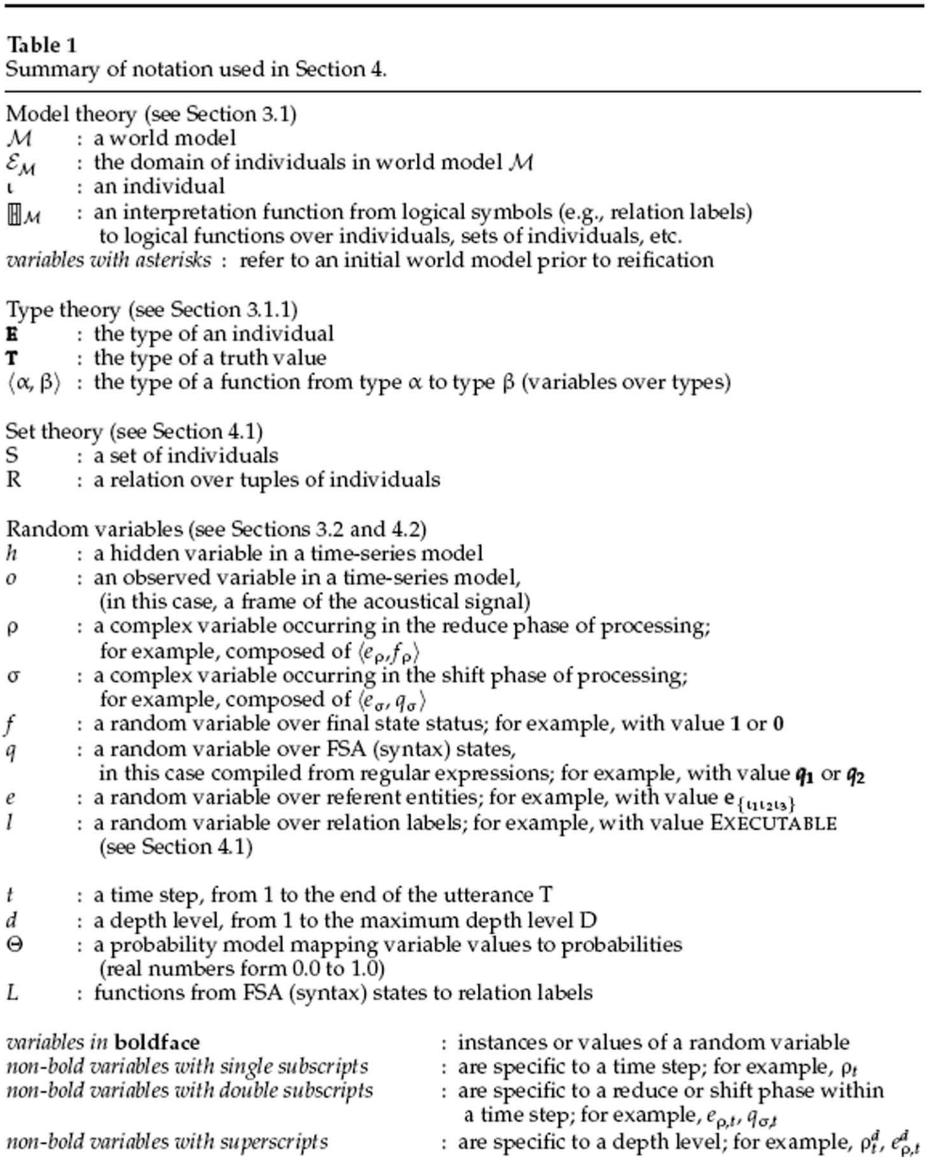


Table 1
Summary of notation used in Section 4.

Model theory (see Section 3.1)
$\mathcal{M}$: a world model
$\mathcal{E}_{\mathcal{M}}$: the domain of individuals in world model $\mathcal{M}$
ι : an individual
$[\![\,]\!]_{\mathcal{M}}$: an interpretation function from logical symbols (e.g., relation labels) to logical functions over individuals, sets of individuals, etc.
variables with asterisks : refer to an initial world model prior to reification

Type theory (see Section 3.1.1)
$\mathbf{E}$: the type of an individual
$\mathbf{T}$: the type of a truth value
$\langle\alpha, \beta\rangle$: the type of a function from type α to type β (variables over types)

Set theory (see Section 4.1)
S : a set of individuals
R : a relation over tuples of individuals

Random variables (see Sections 3.2 and 4.2)
h : a hidden variable in a time-series model
o : an observed variable in a time-series model, (in this case, a frame of the acoustical signal)
ρ : a complex variable occurring in the reduce phase of processing; for example, composed of $\langle e_\rho, f_\rho\rangle$
σ : a complex variable occurring in the shift phase of processing; for example, composed of $\langle e_\sigma, q_\sigma\rangle$
f : a random variable over final state status; for example, with value **1** or **0**
q : a random variable over FSA (syntax) states, in this case compiled from regular expressions; for example, with value $\boldsymbol{q_1}$ or $\boldsymbol{q_2}$
e : a random variable over referent entities; for example, with value $\mathbf{e}_{\{\iota_1\iota_2\iota_3\}}$
l : a random variable over relation labels; for example, with value EXECUTABLE (see Section 4.1)

t : a time step, from 1 to the end of the utterance T
d : a depth level, from 1 to the maximum depth level D
Θ : a probability model mapping variable values to probabilities (real numbers form 0.0 to 1.0)
L : functions from FSA (syntax) states to relation labels

variables in **boldface** : instances or values of a random variable
non-bold variables with single subscripts : are specific to a time step; for example, ρ_t
non-bold variables with double subscripts : are specific to a reduce or shift phase within a time step; for example, $e_{\rho,t}$, $q_{\sigma,t}$
non-bold variables with superscripts : are specific to a depth level; for example, ρ_t^d, $e_{\rho,t}^d$

Figure 7. Table with multiple relations to the text.

title, leads us to think that the article is about a previous time. Then we realize that it is about how Lower Manhattan had empty spaces next to the World Trade Center. Finally, we understand that this space was the location of an art installation.

Thus, the reader has presumably established three different relations between photo and text. There are potential additional relations created with the help of the photo caption, which reads:

(7) "New York Ripple," a 1978 installation by Patsy Norvell for "Art on the Beach," a project on the Battery Park City landfill. See Page 26 for today's view.

The caption helps us establish the later two Elaboration relations, keying the concepts of art installation and landfill to what the text describes.

The possibility of multiple relations between text and depiction is also entertained by Liu and O'Halloran, although in their paper they explore only single conjunctive relations between 'two contiguous visual-linguistic messages' (Liu and O'Halloran, 2009: 379).

The presence of multiple relations to the same illustration may re-open a debate on RST on the cognitive status of rhetorical relations. If the reader can establish multiple relations between one span and another (in this case, between a table or figure and portions of the text), then it is also possible that different readers will establish different relations. For instance, it is entirely plausible that some readers will flip back to the table presented in Figure 7, but that other readers will consult the table fewer times, or not at all. The relations established are then in the mind of the reader, and are not necessarily what the writer intended. It is also possible that the relation is established without the need to look back at the picture, making use of mental imagery (e.g. Kosslyn et al., 2006; Paivio, 1986).

RST has often been pegged as presenting a view of text as product, as opposed to text and discourse as a process. We believe RST does not need to be necessarily reduced to analyzing texts as finished products, and that it is therefore consistent with a treatment where different readers have different interpretations of the document, and where interpretation of the role of depictive material changes as the document is processed. We have shown, in previous work, that RST can be used in the analysis of process-based language, such as conversation (Taboada, 2004).

6. Conclusions

We have presented a corpus-based analysis of the coherence and cohesion relations established between text and depictive material in multimodal documents. We argue that text and depictions stand in coherence relations, which we have chosen to define as rhetorical relations.

We have tried to show that the types of coherence relations that we find in verbal discourse also exist in multimodal discourse. This is at the risk of what Bateman calls 'linguistic imperialism' (Bateman, 2009), whereby researchers tend to assume that other semiotic modes will behave like language, the semiotic mode that we know best. Bateman suggests that such an assumption needs to be empirically investigated, and that is precisely what we have attempted to do in this article. We have found the hypothesis wanting in some aspects (very few of the RST relations are used), but applicable in many others (relations are identifiable, and they capture the functions of depiction).

Our work contributes to ongoing research in the structure of multimodal documents. We follow Bateman's framework (e.g. Bateman, 2008b) in using Rhetorical Structure Theory, and extend it by applying RST to a large collection of assorted documents. We also see a relationship to the work of Holsanova and others (Holsanova, 2008; Holsanova et al., 2009), adding a corpus dimension.

The most significant implication of the research that we would like to conclude with is the potential presence of multiple relations between a particular instance of depictive material and different parts of the text. This could lead to a re-thinking of the structure of RST relations. The most common representation of the structure of a text in RST terms is in the form of a tree. It has been argued that trees are insufficient in some cases, such as parallelism, and that reported speech and other phenomena lead to crossed dependencies that trees cannot capture (Wolf and Gibson, 2005). There is, a priori, no theoretical commitment to trees (Taboada and Mann, 2006b), and thus other structures are possible.

Acknowledgments

We would like to thank audiences at the CINACS Summer School (Hamburg, September 2010), the workshop 'Language and Depiction' (Hamburg, November 2010), the Discourse Research Group (Utrecht, November 2010), and the Cognitive Science Speaker Series (Vancouver, February 2011) for useful feedback and suggestions, in particular Sabine Bartsch, John Bateman, and Manfred Stede. Thanks also to Debopam Das for carrying out the reliability study.

Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada (Discovery Grant 261104-2008). The research was conducted while Maite Taboada was a guest at the University of Hamburg, sponsored by a Fellowship for Experienced Researchers from the Alexander von Humboldt Foundation. We gratefully acknowledge their support.

Notes

1. From the *Communications of the Association for Computing Machinery* 53(3): 63, 'Global IT Management: Structuring for scale, responsiveness and innovation'. The segmentation of this example is done at a coarse level, showing only higher-level relations.
2. The articles in *Communications of the Association for Computing Machinery* have a range of authors, from those more frequently involved in science dissemination, to regular academics who wish to publish in a more accessible venue.
3. The choice of a term was opportunistic (collection was done at around the time of the oil spill in the Gulf of Mexico in 2010) and differed from the approach for the other two genres, where entire collections (a year, or a few months) were used. We could not study a chronological set for the entire newspaper (week, month, or more), and that would have also resulted in many different genres. This was also a practical issue. It was not easy to find an electronic version of a complete issue of *The New York Times*, with pictures as they appeared in print. Their historical archive, which does contain pdf files of the original layout, is searchable, but not browsable by issue.
4. The article, with higher-quality pictures, is available at: http://www.nytimes.com/2006/09/01/arts/design/01city.html (accessed 22 June 2012).
5. An initial study by the Poynter Institute showed that readers attend to images first, then to headlines and the rest of the text. A subsequent study by the same organization attested first attention to headlines, then to pictures. The experimenters attribute the difference to experimental

design: in the second study, participants had been asked to refrain from reading the day's newspapers. The subjects were then presumably genuinely interested in the news, and thus read the headlines first. Full details of the studies are available in the book *Eyetracking the News* (Stark Adam et al., 2007), or from the Institute's web site: http://www.poynter.org/, with the search term 'eye-tracking' (accessed 22 June 2012). There is also a difference in behavior between print and online news reading, with online readers attending first to text. See Lewenstein et al. (2000) for a description of these studies.

6. Bangalore S and Johnston M (2009) Robust understanding in multimodal interfaces. *Computational Linguistics* 35(3): 345–397.
7. Armbrust M, Fox A, Griffith R et al. (2010) A view of cloud computing. *Communications of the Association for Computing Machinery* 53(4): 50–58.
8. Schuler W, Wu S and Schwartz L (2010) A framework for fast incremental interpretation during speech decoding. *Computational Linguistics* 35(3): 314–343.

References

Acartürk C (2010) Multimodal comprehension of graph–text constellations: An information processing perspective. PhD dissertation, University of Hamburg, Hamburg.

Acartürk C, Habel C, Cagiltay K and Alacam O (2008) Multimodal comprehension of language and graphics: Graphs with and without annotations. *Journal of Eye Movement Research* 1(3): 1–15.

André E (1995) *Ein planbasierter Ansatz zur Generierung multimedialer Präsentationen* [A plan-based approach to the generation of multimedia presentations]. St Augustin: Infix.

André E and Rist T (1996) Coping with temporal constraints in multimedia presentation planning. Proceedings of Thirteenth National Conference on Artificial Intelligence, Portland, OR, pp. 142–147.

Baldry A and Thibault PJ (2006) *Multimodal Transcription and Text Analysis: A Multimedia Toolkit and Coursebook*. London: Equinox.

Barthes R (1977) *Image, Music, Text*, trans. Heath S. London: Fontana.

Bateman J (2008a) *Basic techniques and problems in multimodal analysis*. Available at: http://www.fb10.uni-bremen.de/anglistik/langpro/webspace/jb/repository/talks/session-statics.pdf (accessed 22 June 2012).

Bateman J (2008b) *Multimodality and Genre: A Foundation for the Systematic Analysis of Multimodal Documents*. London: Palgrave Macmillan.

Bateman J (2009) Discourse across semiotic modes. In: Renkema J (ed.) *Discourse, Of Course*. Amsterdam and Philadelphia, PA: John Benjamins, pp. 55–66.

Bateman J, Delin J and Allen P (2000) Constraints on layout in multimodal document generation. Proceedings of First International Natural Language Generation Conference, Workshop on Coherence in Generated Multimedia, Mitzpe Ramon, Israel.

Bateman J, Delin J and Henschel R (2007) Mapping the multimodal genres of traditional and electronic newspapers. In: Royce TD and Bowcher WL (eds) *New Directions in the Analysis of Multimodal Discourse*. London: Routledge, pp. 147–172.

Bateman J, Kamps T, Kleinz J and Reichenberger K (2001) Towards constructive text, diagram, and layout generation for information presentation. *Computational Linguistics* 27(3): 409–449.

Carenini G, Pianesi F, Ponzi M and Stock O (1993) Natural language generation and hypertext access. *Applied Artificial Intelligence* 7(2): 135–164.

Carlson L, Marcu D and Okurowski ME (2002) *RST Discourse Treebank*, LDC2002T07 [Corpus]. Philadelphia, PA: Linguistic Data Consortium.

Chu S, Paul N and Ruel L (2009) Using eye tracking technology to examine the effectiveness of design elements on news websites. *Information Design Journal* 17(1): 31–43.

Cohen J (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1): 37–46.

Delin J and Bateman J (2002) Describing and critiquing multimodal documents. *Document Design* 3(2): 140–155.

Eggins S (1994) *An Introduction to Systemic Functional Linguistics*. London: Pinter.

Elzer S, Carberry S, Chester D, Demir S, Green N, Zukerman I and Trnka K (2005) Exploring and exploiting the limited utility of captions in recognizing intention in information graphics. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, MI, pp. 223–230.

Feiner SK and McKeown K (1993) Automating the generation of coordinated multimedia explanations. In: Maybury M (ed.) *Intelligent Multimedia Interfaces*. Menlo Park, CA: AAAI Press, pp. 117–138.

Garcia M and Stark P (1991) *Eyes on the News*. St Petersburg, FL: The Poynter Institute.

Grosz BJ and Sidner CL (1986) Attention, intentions, and the structure of discourse. *Computational Linguistics* 12(3): 175–204.

Halliday MAK and Matthiessen CMIM (2004) *An Introduction to Functional Grammar*, 3rd edn. London: Arnold.

Henschel R (2003) *GeM Annotation Manual*. Bremen: University of Bremen.

Holmqvist K, Holsanova J, Barthelson M and Lundqvist D (2003) Reading or scanning? A study of newspaper and net paper reading. In: Hyönä J, Radach R and Deubel H (eds) *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*. Amsterdam: Elsevier, pp. 657–670.

Holsanova J (2008) *Discourse, Vision, and Cognition*. Amsterdam: John Benjamins.

Holsanova J, Holmberg N and Holmqvist K (2009) Reading information graphics: The role of spatial contiguity and dual attentional guidance. *Applied Cognitive Psychology* 23: 1215–1226.

Holsanova J, Rahm H and Holmqvist K (2006) Entry points and reading paths on newspaper spreads: Comparing a semiotic analysis with eye-tracking measurements. *Visual Communication* 5(1): pp. 65–93.

Jeung H-J, Chandler P and Sweller P (1997) The role of visual indicators in dual sensory mode instruction. *Educational Psychology* 17(3): 329–345.

Knott A, Oberlander J, O'Donnell M and Mellish C (2001) Beyond elaboration: The interaction of relations and focus in coherent text. In: Sanders T, Schilperoord J and Spooren W (eds) *Text Representation: Linguistic and Psycholinguistic Aspects*. Amsterdam and Philadelphia, PA: John Benjamins, pp. 181–196.

Kong KCC (2006) A taxonomy of the discourse relations between words and visual. *Information Design Journal* 14(3): 207–230.

Kosslyn SM, Thompson WL and Ganis G (2006) *The Case for Mental Imagery*. Oxford: Oxford University Press.

Kress G and van Leeuwen T (1998) Front page: (The critical) analysis of newspaper layout. In: Bell A and Garrett P (eds) *Approaches to Media Discourse*. Oxford: Blackwell, pp. 186–219.

Kress G and van Leeuwen T (2006) *Reading Images: The Grammar of Visual Design*, 2nd edn. London: Routledge.

Kress G, Jewitt C, Ogborn J and Tsatsarelis C (2001) *Multimodal Teaching and Learning: The Rhetorics of the Science Classroom*. London: Continuum.

Landis JR and Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33(1): 159–174.

Lemke J (1998) Multiplying meaning: Visual and verbal semiotics in scientific text. In: Martin JR and Veel R (eds) *Reading Science: Critical and Funcitonal Perspectives on Discourses of Science*. London: Routledge, pp. 87–113.

Lewenstein M, Edwards G, Tatar D and DeVigal A (2000) *Poynter Eyetrack Study*. St Petersburg, FL: The Poynter Institute. Available at: http://www.poynter.org/eyetrack2000.

Liu Y and O'Halloran K (2009) Intersemiotic texture: Analyzing cohesive devices between language and images. *Social Semiotics* 19(4): 367–388.

Mann WC and Taboada M (2010) *RST Web Site*. Available at: http://www.sfu.ca/rst.

Mann WC and Thompson SA (1988) Rhetorical Structure Theory: Toward a functional theory of text organization. *Text* 8(3): 243–281.

Marcu D (1999) *Instructions for Manually Annotating the Discourse Structures of Texts* (Manual). Marina del Rey, California.

Marcu D, Romera M and Amorrortu E (1999) Experiments in constructing a corpus of discourse trees: Problems, annotation choices, issues. Workshop on Levels of Representation in Discourse, Edinburgh, UK, pp. 71–78.

Martin JR (1984) Language, register and genre. In: Christie F (ed.) *Children Writing: Reader*. Geelong, Victoria: Deakin University Press, pp. 21–30.

Mayer RE (2009) *Multimedia Learning*, 2nd edn. Cambridge: Cambridge University Press.

Moreno R and Mayer RE (1999) Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of Educational Psychology* 91(2): 358–368.

O'Halloran K (2008) Systemic Functional-Multimodal Discourse Analysis (SF-MDA): Constructing ideational meaning using language and visual imagery. *Visual Communication* 7(4): 443–475.

Paivio A (1986) *Mental Representations: A Dual Coding Approach*. Oxford: Oxford University Press.

Royce TD (2007) Inter-semiotic complementarity: A framework for multimodal discourse analysis. In: Royce TD and Bowcher WL (eds) *New Directions in the Analysis of Multimodal Discourse*. London: Routledge, pp. 63–109.

Scanlan C (2000) *Reporting and Writing: Basics for the 21st Century*. Oxford: Oxford University Press.

Schriver KA (1997) *Dynamics in Document Design*. New York: John Wiley.

Stark Adam P, Quinn S and Edmonds R (2007) *Eyetracking the News*. St Petersburg, FL: The Poynter Institute.

Stenglin M (2009) Space odyssey: Towards a social semiotic model of three-dimensional space. *Visual Communication* 8(1): 35–64.

Stock O (1993) ALFRESCO: Enjoing the combination of natural language processing and hypermedia for information exploration. In: Maybury M (ed.) *Intelligent Multimedia Interfaces*. Menlo Park; CA: AAAI Press, pp. 197–224.

Stöckl H (2009) Beyond depicting: Language-image-links in the service of advertising. *AAA – Arbeiten aus Anglistik und Amerikanistik* 34(1): 3–28.

Swales JM (1990) *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Taboada M (2004) *Building Coherence and Cohesion: Task-oriented Dialogue in English and Spanish*. Amsterdam and Philadelphia, PA: John Benjamins.

Taboada M and Mann WC (2006a) Applications of Rhetorical Structure Theory. *Discourse Studies* 8(4): 567–588.

Taboada M and Mann WC (2006b) Rhetorical Structure Theory: Looking back and moving ahead. *Discourse Studies* 8(3): 423–459.

Thibault PJ (1997) *Re-reading Saussure: The Dynamics of Signs in Social Life*. London: Routledge.

Ventola E, Charles C and Kaltenbacher M (eds) (2004) *Perspectives on Multimodality*. Amsterdam: John Benjamins.

Wolf F and Gibson E (2005) Representing discourse coherence: A corpus-based analysis. *Computational Linguistics* 31(2): 249–287.

Author biographies

Maite Taboada is Associate Professor of Linguistics at Simon Fraser University, Canada. She works in the areas of discourse analysis, systemic functional linguistics and computational linguistics. She has carried out research on coherence and cohesion, information structure and turn-taking. She has also participated in research projects in natural language generation, machine translation and software agents. Ongoing research addresses the study of opinion and sentiment in text, including a system that extracts sentiment automatically. Other current areas of research are coherence in multimodal documents, and the study of cataphoric relations.

Christopher Habel is Full Professor in Computer Science (Artificial Intelligence) and Adjunct Professor of Linguistics at the University of Hamburg, Germany. He holds a doctoral degree from the University of Osnabrück, Germany. He is the director of the Knowledge and Language Processing Group and the director of the study program Human–Computer-Interaction at the Department of Informatics. His areas of research are discourse and event structure; representation of knowledge, in particular, about space, time and events; language comprehension and language production. Some of his current research projects are: multimodal representations and communication; verbal assistance for tactile map explorations.