4 The Computer-Based Analysis of Narrative and Multimodality

Andrew Salway

INTRODUCTION

This chapter is concerned with how the processing power and multimedia capabilities of modern computers can be exploited to understand the workings of stories told in combinations of different types of media. Since the early days of digital technology, computers have been used to count and analyze words and patterns of words in texts, chiefly under the rubric of corpus stylistics, and more generally in the field of digital humanities. At the risk of oversimplification, much of this work can be characterized by its focus on single literary texts or on the works of single authors, and also by the way in which scholars look for signs of previously hypothesized linguistic or literary phenomena in the texts, which sometimes involves manual annotation of the texts prior to automated analysis. In contrast, this study advocates a novel computer-based approach to the analysis of narrative and multimodality that is characterized by the use of a computer to extract unusually frequent patterns from the surface forms of large collections of multimodal stories. Crucially, these patterns are to be extracted without the cost and bias due to prior manual annotation and the encoding of grammars, pragmatics, and world knowledge. Instead, patterns are extracted from corpora of multimodal texts on purely statistical grounds.

The extraction of such patterns is envisioned as a starting point for developing theory in two ways. Firstly, by asking why particular forms are prevalent in a corpus, we may be led to develop explanations of the observed phenomena. Secondly, because the phenomena of interest have known surface forms, it is straightforward to automatically generate more data about their occurrence in the initial corpus and in new instances of multimodal texts in order to test hypotheses. By means of two case studies I assess the potential for this kind of approach, discuss its inherent limitations, and point to opportunities for further work. A major obstacle to the computer-based analysis of multimodal texts is the current limit on what can be achieved with automatic image and video analysis techniques, compared with text analysis. This issue was sidestepped in the work presented in the first case study as feature films were analyzed by extracting patterns from corpora of texts that are surrogates for different parts of the multimodal text, i.e., film scripts, subtitles, audio description, and plot summaries. The second case study considers the extent to which image analysis techniques may be used together with language processing in the analysis of multimodal texts, like Web pages.

By way of a backdrop to the case studies, I briefly review other work that used computers in the analysis of multimodal texts. I also give a view on the current state-of-the-art technologies for text and image understanding, which should inform our expectations for, and approaches to, automating the analysis of narrative and multimodality. Baldry (2004) describes a multimodal concordancer with which a researcher annotates a multimodal text, such as a video recording of an advert or a film, with textual descriptions. The system then assists the researcher by generating statistics to help find previously specified patterns in their descriptions. To access and analyze levels of multimodal documents such as rhetorical structure, layout structure, and navigation structure—specifically with a view to finding correlations between these layers and document genres—an XML-based annotation scheme was developed and deployed by Bateman and colleagues (Bateman, Henschel, and Delin 2002; Bateman 2008). An approach to computer-assisted multimodal analysis that does not require transcription or annotation is reported in O'Halloran (2004b). This concentrates on analyzing the visual properties of film by using a standard video editing system to manipulate the image's brightness, color, contrast, etc., in order to highlight different semiotic choices and to view their effects. Our proposed use of technology for data-driven explorations of narrative and multimodality should be seen as complementary to the development of tools that enable researchers to carry out very deep and detailed analyses of multimodal documents in an interactive fashion. O'Halloran (2009) explores possible ways of using digital technology for a kind of multimodal analysis that is grounded in social semiotics. She notes how technology has led to new research paradigms in mathematics and science, and suggests analogous impacts of digital technology on the field of multimodality. In particular, attention is drawn to the potential for visualization techniques, coupled with techniques for automated low-level visual content analysis, to help multimodal discourse analysts to unpick the multiple interwoven strands of meaning in media such as video and interactive Web sites. It is argued that these techniques are better suited than static, paged-based transcriptions for helping researchers to articulate and elucidate semiotic choices in multimodal documents that are "conceptualised as continuous spatial-temporal-type relations." There are challenges here, however, to do with designing interfaces for such tools in a way that maintains theoretical consistency and ease of use (O'Halloran et al., forthcoming).

Compared with successes in text analysis and language understanding, the use of computers for the automated analysis of multimodal documents will be limited by the fact that image, video, and audio data are much less computationally tractable than text data. The nature of language and the machine encoding of written text data makes explicit the basic meaning-bearing units (words) and their meaningful sequencing, and so word frequencies and collocation data give insights into document meanings, and some degree of automatic parsing and mapping into semantic representations is possible. By contrast, the machine-level encoding of still and moving images comprises matrices of pixels: each pixel is a color value for one point in the image and so carries no meaning by itself.

The state of the art in language understanding technology is represented by work in the field of information extraction; for an overview, see Gaizauskas and Wilks (1998). An information extraction system extracts certain kinds of facts from a given type of text: for example, financial news stories about company takeovers are analyzed to fill a database of "Takeover Events" with details for each event about which company bought the other, for how much, and the people involved. Whilst useful for commercial applications, information extraction technology still falls a long way short of what would be considered to be story understanding in the field of narratology. The recent development of a system to model space and time in restaurant narratives (Mueller 2007) points to the challenges here. It required thirteen thousand lines of computer code to extract information from restaurant narratives and to reason about this information in order to answer generic questions about the dining experience. Whilst efforts are ongoing to produce encodings of general world knowledge,[1] it seems to me that Dreyfus's arguments as to what computers still can't do hold sway here (Dreyfus 1992). So, I argue, if we are interested in generally applicable approaches to the computer-based analysis of narrative and multimodality, we should concentrate on exploiting what computers can do, and favor approaches that analyze patterns in the surface forms of multimodal texts.

As noted earlier, the machine encodings of written texts make explicit something of the meaning-bearing elements, but this is not so for image and video data. Thus we must consider what can be done to get beyond pixel-level representations using technologies from the interrelated and overlapping fields of image processing, image analysis, and image understanding (or computer vision); for an introduction to these fields see Gonzalez, Woods, and Eddins (2004). The task of automatic image understanding, i.e., producing a meaningful description of what can be seen in an image, is only achievable in highly constrained images—such as single objects set against a plain background and shot with good lighting (for more on the challenges here, see Smeulders et al. 2000). That said, there are some techniques for image and video analysis that have matured to the extent that they could be applied to the analysis of multimodal documents. A review of multimedia technologies by Smeaton (2004) included: face detection and recognition in image and video data; video segmentation into shots and the selection of representative key frames for each shot; and the detection of commonly occurring image content like "children," "sand," "water," "outdoors," and "sky." For the analysis of feature films specifically, a technique has been demonstrated that classifies film sequences into either "action," "montage," or "dialogue" (Lehane and O'Connor 2006). A combination of low-level video features to do with the rate of shot change and the amount of motion within the frame were combined to give a measure of "Tempo" in feature films by Adams, Dorai, and Venkatesh (2002): changes in Tempo were shown to coincide with dramatically important moments in films. Such image and video analysis techniques suggest that it may be possible to extract useful features from visual information as part of the analysis of multimodal stories. This potential is discussed further in the second case study, but first we look at work that analyzed only textual surrogates for multimodal stories.

CASE STUDY I: ANALYZING NARRATIVE IN FILM

Implicit in the approach that we are advocating is the assumption that interesting narrative characteristics manifest in regular ways in the digital forms of multimodal texts, like films, so that we can be led to them by extracting forms that are unusually frequent in a large collection of the texts. The work discussed in this section comprises investigations into how narrative functions in film, based on the automated analysis of corpora of film scripts, plot summaries, subtitles, and audio description. The use of these corpora (summarized in Table 4.1), allows us to deal separately with different information streams, or semiotic strands, in the complex multimodal artefact that is film: audio description[2] stands as a surrogate for the characters, props, scenes, and actions depicted on screen; subtitles as a surrogate for the dialogue, and film scripts and plot summaries do both. Of course we must recognize the importance of studying the interplay between the moving image and the sound track, and the diverse semiotic choices available within each; for a detailed inventory of the meaning-making elements in film, see O'Halloran (2004b). We must also recognize that these textual surrogates are only an imperfect and incomplete record of the film. For now though, in the context of the current discussion, we seek to provide evidence in support of our underlying assumption by showing that some narrative aspects of film manifest in regular forms.

Table 4.1 Four Corpora of Text Surrogates for Film

Text Type	*Source*	*Number of texts*	*Number of words*
Audio Description	Major UK producers of audio description—RNIB, BBC, and ITFC	73	c. 713,000
Film Scripts	www.imsdb.com www.simplyscripts.com	75	c. 1,900,000
Subtitles	www.subscene.com	80	c. 516,000
Plot Summaries	www.imdb.com	111	c. 14,000

The research summarized and discussed here used a variety of automated techniques to investigate what kinds of information are commonly provided by audio description, film scripts, subtitles, and plot summaries. Though the details varied, the general method was: (a) to identify unusually frequent words in each corpus by comparing the frequencies of words in the corpus with their frequencies in a general language sample, e.g., the British National Corpus; (b) to identify collocations of the unusually frequent words, i.e., statistically significant word sequences that contained them; and, in the analysis of film scripts and audio description, (c) to merge and generalize the collocations to produce fragments of a local grammar that described syntagmas and paradigms induced from the corpora. This third step can be viewed as an implementation of Harris's approach to the study of language and information, whereby patterns induced from a text corpus reveal the information structures of a domain (Harris 1988). The general method followed, and the link to Harris, were proposed in work done on information extraction from financial news stories (Traboulsi, Cheng, and Ahmad 2004; Almas and Ahmad 2006). For details of the techniques used to analyze film texts, and for more comprehensive discussion of the results, see the following references: for the analysis of audio description and film scripts see Salway, Vassiliou, and Ahmad (2005) and Salway (2007), and in particular for the induction of local grammars from these texts see Vassiliou (2006); for subtitles, see Lingabavan and Salway (2006); and for plot summaries see Tomadaki and Salway (2006) and Tomadaki (2006). Results from these studies are selected and presented here with a view to showing that patterns identified in these corpora lead us to some of the narrative functioning of film.

Starting with the corpora of film scripts and audio description—which we discuss together because they exhibit many frequent words and collocations in common—we see that the most unusually frequent words include the following: *looks, door, turns, away, head, towards, eyes, room, takes, around, walks, behind*. These appear in collocations such as *X looks at Y*, *X walks to the* Y, *X opens the door and leaves*, and *X nods her head*. It is straightforward to infer that these word sequences are common in film scripts and audio description because the actions that they refer to are frequently depicted in mainstream film. On the basis of the four sequences, we might go on to say that storytelling in film requires the filmmaker to ensure that the audience knows who is looking at what or whom, who is coming and going where, and about nonverbal communication between characters, but of course this selection of four phrases is arbitrary. A more systematic approach was taken by Vassiliou (2006) to merge and generalize collocation patterns to automatically generate a description of local grammar fragments around unusually frequent words. Figure 4.1 shows a local grammar fragment for the word *looks*. Note that the shaded parts of the diagram required some human intervention, but the rest was generated entirely automatically. The diagram, read left to right, captures many observed, and predicted, word sequences around *looks*, and around other words that appeared in similar contexts, i.e., *stares*, *gazes*, etc. This local grammar fragment can be interpreted as reflecting one kind of information that these texts, and hence films, frequently convey, i.e., information about what and whom characters are looking at.

By automatically generating such local grammar diagrams for the ten most unusually frequent words in audio description and film scripts, Vassiliou (2006) produced a clear picture of some of the kinds of information commonly provided by these texts. By manually abstracting from the diagrams he postulated three main kinds of film events, which were all grounded in empirically observed text forms. He labeled these "Focus_of_Attention" (comprising *X looks at Y*, etc.), "Change_of_Location" (*X walks to Y*, etc.) and "Non-verbal_Communication" (*X nods her head*, etc.). Whilst an unsurprising finding for anyone with a passing knowledge of how stories are told in film, what is significant about the postulation of these events is that each is tied to a realization in the form of word sequences. An important corollary of this is that it is easy to develop an information extraction system to automatically generate a database comprising data about these events in any number of films for which film scripts or audio description is available, as Vassiliou went on to do. The information extraction system, albeit imperfectly, instantiated a database containing data about Focus_of_Attention events, Change_of_Location events and Non-verbal_Communication events in 193 films, including details of the characters, objects, and locations involved, and estimates of when they happened (from audio description time codes and relative positions in film scripts).

A similar approach, though implemented less extensively, was taken to the analysis of a subtitle corpus (Lingabavan and Salway 2006). Preliminary results showed unusually frequent words being: *don't, gonna,*

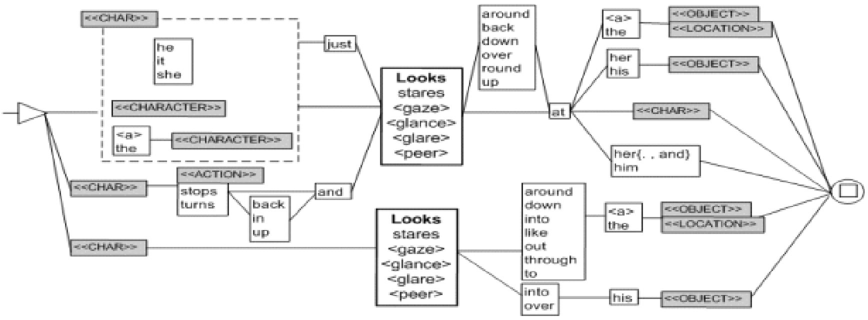

Figure 4.1 A local grammar fragment induced from corpora of film scripts and audio description, from Vassiliou (2006).

didn't, hey, fuck, guy, shit, uh, gotta, doesn't. Taking these words as candidate nucleates, the standout collocations were *I don't know*, *I'm gonna*, *I wanna*, and *I gotta*. On the basis of these results two common dialogue events were postulated: "Statement_of_Lack_of_Knowledge" (realized by *I don't know*) and "Statement_of_Future_Intended_Action" (realized by the other three collocations), which had three varieties according to whether the future action was something that the speaker was planning to do to meet one of their own goals, or because of some external force acting on them. Such an analysis is impoverished compared to the rich discussion of the functions of film dialogue provided by Kozloff (2000), but again the key point is that the postulated dialogue events are tied to formal realizations and so data about their occurrence could be extracted automatically from subtitles for further analysis. The corpus of plot summaries analyzed in Tomadaki and Salway (2006) and Tomadaki (2006) was not large enough to allow for the automatic identification of collocation patterns on statistical grounds, but manual inspection of concordances of frequent words did suggest some common kinds of events being described. Compared with the relatively basic actions described in audio description and film scripts (*look, walk, take*), plot summaries describe larger-scale events such as: *help*, *meet, kill, bring, tell, force, find, discover, love,* and *murder*, with common patterns including *fall in love with*, *X helps Y*, *X discovers that* . . . and *X finds Y.*

The findings reported in this section have made a reasonable case that an automated analysis of narrative texts can be revealing about some of the kinds of information that must be conveyed for successful storytelling, and hence inform theories of how narrative functions. In particular, by analyzing corpora of different surrogate texts it seems possible to gain insights into how the different semiotic strands of film contribute to the whole. All the work presented earlier is incomplete in the sense that the corpus data could have been drilled more deeply, and the corpora could have been increased in size. With further work we would expect to see more kinds of commonly occurring events identified on the basis of statistically significant patterns. There is also much to be done analyzing the data already extracted about the commonly occurring events. One piece of ongoing research is looking at word patterns in audio description that express information about characters' mental states—common patterns of this type include *CHARACTER looking *ed|*ly* (where *ed stands for any word ending -ed, and *ly any word ending -ly), and *CHARACTER smiles|stares|looks|walks *ly*. For example, the description of an on-screen action as *John smiles nervously* conveys something about John's mental state, cf. Palmer's thought-action continuum and his arguments for the centrality of characters' mental states to story (Palmer 2004). The description of thoughts and actions in audio description is analyzed by Salway and Palmer (2007).

CASE STUDY II: ANALYZING MULTIMODALITY IN WEB PAGES

I now turn to address the question of whether interesting patterns can be extracted automatically from the visual components of multimodal documents. As noted previously, we would not expect such patterns to emerge directly from image and video data represented at the level of pixels, but rather from the analysis of prespecified image and video features that can be derived automatically from image and video data, e.g., color distributions in images, rate of shot change in videos, etc. To date, there have been no attempts to derive insights into the narrative functioning of multimodal documents by analyzing image and video features with a method analogous to that described in the previous section for text corpora. Thus, here, the best we can do is to present a framework that was developed manually to account for image–text relations in multimodal documents, and to assess the extent to which current image analysis techniques could automate the framework in the future. Specifically, we argue that patterning in the digital forms of multimodal documents cues various kinds of relations between the elements making up these documents, with examples of how the functioning of image–text relations in Web pages is cued by combinations of text features, image features, and page-layout features. The core of the work discussed here is the system of image–text relations proposed by Martinec and Salway (2005), which combines two kinds of relations—drawing on the ideas of Barthes (1977) and Halliday (1985)—the relative status of images and text, and how they relate to one another in terms of logico-semantics. Compared with previous functional-systemic approaches to multimodality, such as Kress and van Leeuwen (1996), and most of the papers in O'Halloran (2004a), and in Royce and Bowcher (2007), this system of image–text relations may be considered distinct on two counts: (i) it is intended to be general to all genres of multimodal discourse where images and text co-occur; (ii) in the identification of each image–text relation, great emphasis was placed on the specification of tangible, machine-processable realizations of each relation, e.g., some mix of text, image, and page-layout features detectable in the surface form of multimodal documents. That said, in the context of the current discussion, it must be stressed that the system of image–text relations was developed through an entirely manual analysis of multimodal documents. This section first summarizes Martinec and Salway's system of image–text relations, and then discusses the extent to which these relations may be realized in forms that a computer can detect automatically, mentioning recent work that has begun to test empirically one kind of image–text relation in online news stories.

With regards to their relative status, the relation between an image and a text is *equal* when either: both the image and the text are required for successful communication, in which case they are *equal-complementary*; or, both the image and the text can be understood individually, in which case they are *equal-independent*. The relation between an image and a text is *unequal* when either the image or the text can be understood individually; that which cannot be understood individually is *subordinate* to the other. Consider some examples of common kinds of Web pages. Images tend to be subordinate to the main text on news Web pages whereas the text tends to be subordinate to the image on art gallery Web pages. Image and text often share an equal relationship in Web pages used for science teaching and learning. The relationship is equal-independent when both convey the same information in different ways, and when there is a cross-reference between the image and the text. Some technical images may require text to identify them, in which case the image and the text, probably a caption, are equal-complementary.

We identified three main kinds of logico-semantic relations between images and texts. A text *elaborates* the meaning of an image, and vice versa, by further specifying or describing it. Photographs on news Web pages frequently elaborate the text, specifically the first paragraph. For example, an image often specifies what a person referred to in the news story looks like—this information is missing from the text. On art gallery Web pages, parts of painting captions elaborate the images (paintings) with a description of image content. It is possible for an image to elaborate a text and a text to elaborate an image at the same time, as in the case of a scientific diagram and accompanying text in an equal-independent relationship. A text *extends* the meaning of an image, and vice versa, by adding new information; for example, car adverts that comprise images of cars and text giving information about price and performance. A text *enhances* the meaning of an image, and vice versa, by qualifying it with reference to time, place, and/or cause-effect. In one online news story we looked at there was an enhancement relation between the image and the first paragraph of the text because the image depicted the effect of the explosion reported in the text.

Salway and Martinec (2005) suggested some image features, text features, and page-layout features that might be used together to classify image–text relations automatically, but these were not tested experimentally. It is expected that such classification will be easier if the genre of the image–text combination is known because there may be preferred image–text relations with genre-specific realisations. The features that were considered included: page layout and formatting—the relative size and position of the image and the text, font type and size, image border; references in the text—for example, "this picture shows . . . ," "see Figure 1," "on the left," "is shown by"; the grammatical characteristics of the text—tense, quantification, use of full sentences or short phrases; modality of images—a scale from realistic to abstract, or from photographic to graphic; and framing of images—for example, one centered subject or no particular subject. Crucially, all these features can be analyzed automatically with current multimedia technologies. The text features require straightforward string matching, or relatively simple part-of-speech tagging.[3] Page-layout features, whose importance for the reading of multimodal documents is made clear by Kress and van Leeuwen (1996), can be analyzed automatically with page structure algorithms such as those developed by Cai et al. (2003) and Song et al. (2004). The face detection techniques mentioned in the introduction section of this chapter make it feasible to extract framing features for images of people. As for image modality, defined by Kress and van Leeuwen (1996) as a function of a function of depth, color saturation, color differentiation, color modulation, contextualization, pictorial detail, illumination, and degree of brightness, this can perhaps be interpreted in terms of low-level image features extracted from the analysis of pixel data.

Two kinds of features that seem most indicative of status relations are page layout and lexical references. In languages that are read left to right, and top to bottom, we expect the subordinate media type to appear to the right of or below the other and to occupy less space. References in text such as "this painting . . ." or "Figure X shows . . ." are strong indications that the text is about (and therefore subordinate to) the image. Other references like "this is shown in Figure X" may suggest equal status—especially when near the end of the text. Determining logico-semantic relations involves a comparison between what is depicted in the image and what is referred to by the text. If exactly the same people, objects, and events are depicted and referred to, then there is elaboration. If completely new things are depicted or referred to then there is extension. If related temporal, spatial, or causal information is provided then there is enhancement. The question is how may such comparisons be computed? Information extraction techniques can recognize proper nouns and work out who is the subject of a story, and determine what kind of event or state is being referred to in a text. Image analysis techniques can detect faces, indoor versus outdoor scenes, and framing—all of which may give clues about the main subject being depicted. It might also be interesting to compare the complexity of the image and of the text: image complexity could perhaps be measured as a function of the number of edges/regions or graphic elements. Measures of text complexity relate to sentence length, average word length, and use of embedded clauses. When a text elaborates an image we noticed that often present tense is used, or short phrases rather than complete sentences. When an image elaborates a text in the news domain the image is of a realistic modality and typically depicts a person who is framed so that the head and shoulders fill the photograph, and the text tends to repeat the name of the person depicted. The enhancement relation of cause-effect is realized when the image depicts a process and the text refers to a state, or vice versa. The image seems to normally be a general scene, rather than a closely cropped photograph with one main subject.

The first empirical investigation of image–text relations, with both human subjects and automatic classification, was carried out by Hughes et al. (2007). An experiment was conducted to test the hypothesis that humans can predict the main theme of a text by looking quickly at an associated image. By seeing pictures of people that accompany eighty online news stories but not the text, twenty-five subjects could predict very accurately whether the story was about the specific person/people depicted in the image or about a more general theme. Human performance on this task was then emulated with automatic classification based on low-level image features. Using a face detection algorithm set to detect large full-frontal faces, a measure of variation in image sharpness across the image and certain features intended to correlate to image modality, it was possible to correctly classify photographs into "Specific" or "General" categories in 82.5 percent of eighty online news stories. This result is the first sign that visual features relevant to multimodality can indeed be extracted automatically from image data.

The currently available evidence for frequent and regular surface forms in image data is significantly weaker than that for text data discussed in the previous section. The detailed set of image–text relations discussed here was conceived on the basis of manual investigation, albeit with one eye on realizations for each relation in terms of features that are machine processable. What is certain is that there is a good variety of text features, image features, and page-layout features that can be analyzed automatically, and that these fit well with the explanations of multimodality initiated by Kress and van Leeuwen (1996), and developed by Martinec and Salway (2005). One important factor here is that in order for image–text relations to be recognized automatically, they need to be realized consistently in the same ways, as with online news stories. However, it remains to be seen whether there are many other kinds of multimodal stories that are widely available in digital form, and that exhibit a sufficiently high degree of conventionality in how visual and verbal information are combined. One way to answer this question in the spirit of the current chapter would be to attempt to induce a set of image–text relations based on text, image, and page-layout features extracted from a large sample of multimodal stories. Another direction for further research is to look at multimodality in feature films, perhaps using audio description and subtitles as surrogates for the moving image and dialogue to facilitate the analysis of logico-semantic relations between the two; or by fusing text and video analysis as did Salway, Lehane, and O'Connor (2007).

CLOSING REMARKS

To conclude, in broad terms first, the automatic identification of forms peculiar to certain kinds of multimodal narratives can be a starting point in developing data-driven theories that explain their functioning. The data could be used to validate or refute existing theories in a hypothesis-led fashion, but I would argue that the approach lends itself to stimulating new questions and to the development of new theories to account for aspects of narrative and multimodality that were previously unrecognized, or at least remain as yet to be established. Following the first path certainly, and most probably following the second too, this approach ought to be compatible with many of the diverse theoretical perspectives from which narrative and multimodality are viewed. In the simplest version of the method there would be one step of automatic analysis to identify idiosyncratic features in a corpus of multimodal stories, and then one step to interpret and explain the observed data. However, most likely, to be effective this method will need to be cyclical, so that results from the early steps guide further steps of automated text annotation, the automated generation of databases and the forming of analytical concepts and hypotheses, which will then all feed back into corpus analysis.

For example, Herman directs narrative theory so that "the real target of narrative analysis is the process by which interpreters reconstruct the storyworlds encoded in narratives" (Herman 2002, 5), and so that narrative theory will develop through the charting of "constraints on the variable patterning of textual cues with the mental representations that make up storyworlds" (Herman 2002, 12). Previously this charting has started with researchers postulating mental representations and then looking for the textual cues to which they may correspond. Now we may be in a position to offer a complementary approach that reverses the process and starts with the identification of the textual cues—at least those corresponding with the most common mental representations that make up storyworlds. From our case studies, it seems that insights into the narrative functioning of multimodal documents can be gained from the statistical analysis of an unannotated corpus.[4]

However, a fundamental limitation of the approach is that whilst it might facilitate generalizations by drawing attention to frequently occurring forms, it will miss the rarer and, to some, the more intriguing aspects of narrative and multimodality. We must also note that the approach is currently more feasible when surrogate texts are used in place of image and video data. The lack of suitable surrogate texts may then limit the applicability of the approach. That said, ongoing developments in the very active fields of image and video analysis, along with new multimedia encoding standards like MPEG-4 that make the structure and content of multimedia documents more explicit in machine-processable representations, mean that we expect most kinds of multimedia data to become increasingly computationally tractable. Thus, through the interdisciplinary effort of narrative and multimodality scholars, and computer scientists, we should expect much more in the coming years from the computer-based analysis of narrative and multimodality than I have been able to demonstrate here.

ACKNOWLEDGMENTS

My thanks go to colleagues with whom I have worked on narrative and multimodality, and who have helped to shape my thinking. Working with me on the TIWO project to investigate narrative and film from a computational perspective were Elia Tomadaki, Andrew Vassiliou, and Yan Xu; their PhD theses explore in depth some of the issues I touched on only briefly in this chapter. I am very grateful to Radan Martinec for a stimulating and enjoyable collaboration working on image–text relations. My appreciation for narrative has benefited enormously thanks to the help of David Herman and Alan Palmer. Ongoing work with researchers at the Centre for Digital Video Processing, Dublin City University, has helped me to understand a wide range of multimedia content analysis techniques. Needless to say, none of the aforementioned are responsible for flaws in this chapter.

NOTES

1. http://www.cyc.com/.
2. Audio description is produced for the benefit of blind and visually impaired television and film audiences. It is scripted before it is recorded and aligned with the film/television program: our corpus comprises audio description scripts. For more about audio description, see Diaz-Cintas, Orero, and Remael (2007).
3. Such as http://www.connexor.com/demo/tagger/.
4. These ideas are explored with reference to an analysis of monomodal narratives by Salway and Herman (forthcoming).

REFERENCES

Adams, B., C. Dorai, and S. Venkatesh. 2002. “Towards Automatic Extraction of Expressive Elements for Motion Pictures: Tempo.” *IEEE Transactions on Multimedia* 4 (4): 472–81.

Almas, Y., and K. Ahmad. 2006. “LoLo: A System Based on Terminology for Multilingual Extraction.” In *Procs. COLING Workshop on Information Extraction Beyond The Document*, 56–65. Sydney.

Baldry, A. 2004. “Phase and Transition, Type and Instance: Patterns in Media Texts as Seen through a Multimodal Concordancer” In *Multimodal Discourse Analysis: Systemic Functional Perspectives*, ed. K. O’Halloran, 83–108. London: Continuum.

Barthes, Roland. 1977. “Introduction to the Structural Analysis of Narratives.” In *Image Music Text*, trans. Stephen Heath, 79–124. New York: Hill and Wang.

Bateman, J. 2008. *Multimodality and Genre: A Foundation for the Systematic Analysis of Multimodal Documents*. Hampshire: Palgrave Macmillan.

Bateman, J., R. Henschel, and J. Delin. 2002. “A Brief Introduction to the GeM Annotation Scheme for Complex Document Layout.” In *Procs. 2nd Workshop on NLP and XML International Conference on Computational Linguistics. Association for Computational Linguistics*, 1–8.

Cai, D., et al. 2003. “Extracting Content Structure for Web Pages Based on Visual Representation.” In *Procs. of Web Technologies and Applications: 5th Asia-Pacific Web Conference, APWeb 2003*. Xian, China.

Dreyfus, H. 1992. *What Computers Still Can’t Do*. Cambridge and London: The MIT Press.

Diaz-Cintas, J., P. Orero, and A. Remael, eds. 2007. *Media for All: Subtitling for the Deaf, Audio Description, and Sign Language*. Amsterdam and New York: Rodopi.

Gaizauskas, R., and Y. Wilks. 1998. “Information Extraction: Beyond Information Retrieval.” *Journal of Documentation* 54 (1): 70–105.

Gonzalez, R., R. Woods, and S. Eddins. 2004. *Digital Image Processing Using MATLAB*. Upper Saddle River, NJ: Prentice-Hall.

Halliday, M. 1985. *An Introduction to Functional Grammar*. London: Arnold.

Harris, Z. 1988. *Language and Information*. New York: Columbia University Press.

Herman, D. 2002. *Story Logic: Problems and Possibilities of Narrative*. Lincoln and London: University of Nebraska Press.

Hughes, M., A. Salway, G. Jones, and N. O’Connor. 2007. “Analysing Image-Text Relations for Semantic Media Adaptation and Personalisation.” In *Procs. 2nd International Workshop on Semantic Multimedia Adaptation and Personalisation*. London.

Kress, G., and T. van Leeuwen. 1996. *Reading Images: The Grammar of Visual Design*. London and New York: Routledge.

Kozloff, S. 2000. *Overhearing Film Dialogue*. Berkeley and Los Angeles: University of California Press.

Lehane, B., and N. O’Connor. 2006. “Movie Indexing via Event Detection.” In *Procs. 7th International Workshop on Image Analysis for Interactive Multimedia Services*. Incheon, Korea.

Lingabavan, V., and A. Salway. 2006. “What Are They Talking About? Information Extraction from Film Dialogue.” Dept. of Computing, Technical Report CS-06–07, University of Surrey.

Martinec, R., and A. Salway. 2005. “A System for Image-Text Relations in New (and Old) Media.” *Visual Communication* 4 (3): 337–71.

Mueller, Erik T. 2007. “Modelling Space and Time in Narratives about Restaurants.” *Literary and Linguistic Computing* 22 (1): 67–84.

O’Halloran, K., ed. 2004a. *Multimodal Discourse Analysis: Systemic Functional Perspectives*. London: Continuum.

———. 2004b. “Visual Semiosis in Film.” In *Multimodal Discourse Analysis: Systemic Functional Perspectives*, ed. K. O’Halloran, 109–30, London: Continuum.

———. 2009. “Multimodal Analysis and Digital Technology.” In *Interdisciplinary Perspectives on Multimodality: Theory and Practice. Proceedings of the Third International Conference on Multimodality*, ed. A. Baldry and E. Montagna. Campobasso: Palladino.

O’Halloran, K. L., S. Tan, B. A. Smith, and A. Podlasov. Forthcoming. “Challenges in Designing Digital Interfaces for the Study of Multimodal Phenomena.” *Information Design Journal*.

Palmer, A. 2004. *Fictional Minds*. Lincoln and London: University of Nebraska Press.

Royce, T., and W. Bowcher, eds. 2007. *New Directions in the Analysis of Multimodal Discourse*. Mawah and London: Lawrence Erlbaum Associates.

Salway, A. 2007. “A Corpus-Based Analysis of the Language of Audio Description.” In *Media for All: Subtitling for the Deaf, Audio Description, and Sign Language*, ed. J. Diaz-Cintas, P. Orero and A. Remael, 151–74. Amsterdam and New York: Rodopi.

Salway, A., and D. Herman. Forthcoming. “Digitized Corpora as Theory-Building Resource: New Foundations for Narrative Inquiry.” In *New Narratives: Theory and Practice*, ed. R. Page and B. Thomas. Lincoln and London: University of Nebraska Press.

Salway, A., B. Lehane, and N. O'Connor. 2007. "Associating Characters with Events in Films." In *Procs. of ACM Conference on Image and Video Retrieval—CIVR 2007.* Amsterdam.

Salway, A., and R. Martinec. 2005. "Some Ideas for Modelling Image-Text Combinations." Dept. of Computing, Technical Report CS-05–02, University of Surrey.

Salway, A., and A. Palmer. (2007). "Describing Actions and Thoughts." Paper presented at the Advanced Seminar Audiodescription for Visually Impaired People: Towards an Interdisciplinary Research Agenda, University of Surrey, June 27–28.

Salway, A., A. Vassiliou, and K. Ahmad. 2005. "What Happens in Films?" In *Procs. of IEEE Conference on Multimedia and Expo*, *ICME* 2005.

Smeaton, A. 2004. "Indexing, Browsing and Searching of Digital Video." In *ARIST—Annual Review of Information Science and Technology*, *Vol. 38*, ed. B. Cronin, 371–407. American Society for Information Science and Technology.

Smeulders, A., et al. 2000. "Content-Based Image Retrieval: The End of the Early Years." *IEEE Transactions Pattern Analysis and Machine Intelligence* 22 (12): 1349–80.

Song, R., et al. 2004. "Learning Block Importance Models for Web Pages." In *Procs of WWW 2004.* New York.

Tomadaki, E. 2006. "Cross-Document Coreference between Different Types of Collateral Texts for Films." PhD diss., University of Surrey.

Tomadaki, E., and A. Salway. 2006. "Cross-Document Coreference for Cross-Media Film Indexing." In *Procs. of LREC 2006 Workshop on Crossing Media for Improved Information Access.* Genoa, Italy.

Traboulsi, H., D. Cheng, and K. Ahmad. 2004. "Text Corpora, Local Grammars and Prediction." In *Procs. of 4th International Language Resources and Evaluation Conference*, vol. 3, 749–52. Lisbon.

Vassiliou, A. 2006. "Analysing Film Content—A Text-Based Approach." PhD diss., University of Surrey.