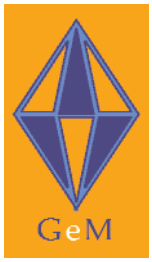

Genre and Multimodality (GeM): a computer model of genre and document layout

Multimodality and empiricism: methodological issues in the study of multimodal meaning-making

GeM project report 2002/01

authors: John Bateman
Judy Delin
Renate Henschel

email: bateman@uni-bremen.de
j.l.delin@stir.ac.uk
rhenschel@uni-bremen.de

website: www.purl.org/net/gem

Funded by ESRC

University of Bremen

University of Stirling

Multimodality and empiricism: methodological issues in the study of multimodal meaning-making

John Bateman / Judy Delin / Renate Henschel

Universities of Bremen, Nottingham Trent and Stirling

Abstract

Following the 'visual turn' in many areas of communication, investigators are increasingly considering explicitly both the presentation of information in forms such as photographs, diagrams, graphics, icons and so on, and interrogating their relationships with linguistically presented information. The majority of analyses currently proposed, however, remain impressionistic and difficult to verify. In this paper, we argue that the study of multimodal meaning-making needs to be placed on a more solid empirical basis in order to move on to detailed theory construction. We describe the state of the art in corpus preparation and show how this can be expanded to be of value for supporting investigative work in the area of multimodality.

1. Introduction

Following the so-called 'visual turn' in many areas of communication, it has become increasingly usual for investigators both to consider explicitly the presentation of information in forms such as photographs, diagrams, graphics, icons and so on and to place such information in combination with linguistically presented information. This is being done by investigators who are practically motivated (e.g., document design: Schriver (1997), web design: Nielsen (2000), among others), those concerned with text types which are explicitly and unavoidably multimodal (e.g., advertisements: Cook (2001)), philosophers and semioticians (Barthes, 1977), specialists in media communications and film (e.g., Philo (1999), Hall and Critcher and Jefferson and Clarke and Roberts (1999)), and the new breed of 'generalized' linguists who consider all presentation modes as fair game for linguistically-inspired analytic methods (e.g., Kress and van Leeuwen (1996), Royce (1998), O'Halloran (1999), Kress, Jewitt, Ogborn and Tsatsarelis (2000)). Naturally these categories can (and often do) overlap.

Given Kress and van Leeuwen (2001)'s very suggestive account of the 'fall' of monomodality over the past century, it is certain that the attention to multimodal meaning-making will gain in importance and centrality in the analysis of communication generally. Indeed, the very existence of 'monomodal' meaning-making has been revealed as a social fabrication—one which has in part reflected and been formed by the social and technological conditions of production available. On this view, every linguistic act, spoken or written, takes place over more than one 'mode' or channel of communication: spoken language involves gesture, for example, while written language always involves other visual elements, such as the most basic choices of typeface, margins, and headings, whether the text is hand-written or not, and so on. All of these can be made to carry meaning. The study of 'monomodal' texts traditional in linguistics and other text-based investigations is thus revealed as an abstraction whose validity is questionable.

One of the corollaries of the broadening in the area of concern is that we are forced to deal with systems which are manifestly meaning-making (e.g., photographs, diagrams) but for which we lack the rich battery of investigative tools that we now have for linguistic entities. To talk of the 'grammar of images' is to move immediately into difficult and contentious areas. The very considerable consolidations in linguistic knowledge and methodology that have been achieved over the last 40 years have not yet found widespread application here.

The move towards employing a linguistic mode of analytic discourse is one of the defining features of the 'generalised linguists' mentioned above. A principal justification is not that all semiotic modes are to be subordinated to language but that our *tools* for analysing complex social semiotic systems have now been sharpened sufficiently to allow application to any such semiotic, not just the traditionally linguistic. We share this belief: it is indeed now time to apply linguistic constructs and methodologies in an investigation of meaning-making generally. We also believe, however, that one of the tenets upon which modern linguists is building—now more than ever—and which has contributed considerably to its success in unravelling many aspects of linguistic organisation, is its emphasis on a sufficiently strong empirical basis.

Linguistic research is being driven furthest where a close coupling of theory and data is achieved. When the data-basis is too weak, or is not sufficiently strongly linked to the claims that a theory makes, then theory can run wild and the claims it makes are not clearly, if at all, testable; when there is a clear link between theoretical claim and data then theory-construction is more constrained and, consequently, more productive. Precisely this close coupling between data and theory-construction is not yet a strong feature of 'multimodal linguistics'.

This is an old concern—as we see in the following proposal of Descartes in his 'Rules for the Direction of the Mind' (1701) raised by Kay O'Halloran during the Salzburg Symposium:

> " There is need of a method for investigating the truth about things (Rule IV). ... The method consists entirely in an orderly arrangement of the objects upon which we must turn our mental vision in order to discover some truth. (Rule V)."

To which we can add:

> "But if we arrange these items in the ideal order, then as a rule they will be reduced to certain classes; and it may be enough to have an exact view of one class, or of some member of each class, or of some classes rather than others; at any rate, we shall not ever go futilely over and over the same point. This is a great help; a proper arrangement often enables us to deal rapidly and easily with an apparently unmanageable multitude of details." (Descartes, 1701, Rule VII).[1]

Or, as Guenther Kress rather more succinctly during the Symposium phrased the problematic we wish to address here: we need to consider how to "turn stuff into data".

But what even constitutes the data in the setting of multimodal documents is not yet clear. For speech analysis, we have detailed models of phonetics and phonology that can organise the phenomena at hand, and for written text analysis, the grammatical units that are to be found are also relatively straightforward: our 'data' can thus be prepared and classified in many useful ways prior to further investigation. For multimodal documents, however, such models and techniques are still largely lacking. A substantial set of problems is raised by the fact that the object of study is not linear, either temporally or in terms of the principles for its consumption; moreover, its multichannel nature makes it difficult to reconcile and peg together the methods of recording, transcription, analysis and annotation that have been developed separately for each mode. This makes empirical study and validation of theory particularly problematic. We do not have the means for the "orderly handling" of our objects of study and discussion and analysis easily remains impressionistic.

In this paper, we address this concern. We give two examples where informal, interpretative claims have been made about aspects of multimodal discourse and argue that the claims demand a much more rigorous empirical basis to be taken further. We then briefly introduce our own attempt to place multimodal study on a firmer empirical basis. Here we have

---

[1] Descartes quotations taken from: Descartes, Rene. "Rules For The Direction Of The Mind." Descartes Philosophical Writings. Translated by Elizabeth Anscombe and Peter Thomas Geach. Thomas Nelson and Sons Ltd. London. 1954. p. 153-180.

developed an initial classification scheme for organising multimodal information presentations and are constructing a corpus of such presentations to allow further closer empirical investigation. Here we are focusing on page-based documents; it is our intention that our approach complement current attempts to provide guidelines for the provision of corpora involving other kinds of multimodal information presentations; we mention more of this area of investigation in Section 3.3 below.

2. Examples of interpretative analyses

In this section, we consider two very different examples of interpretative analyses, or 'claims' about meaning-making in multimodal documents. However, before we start, it is necessary to emphasise that we do this not to present 'straw-men' to argue against—indeed, quite the contrary is the case. Both examples present either important milestone results or substantial claims in multimodal analysis that we are more concerned to use and build upon than to cast aside. Our only concern in the present paper is to argue that we now need to improve the foundation for these analyses and their conclusions by adopting a linguistically-oriented empirical basis.

2.1 Given/New, Ideal/Real, and Centre/Margin on the Front Page

Our first example draws on a particular aspect of the social semiotic interpretation of layout argued by Kress and van Leeuwen (1996). Kress and van Leeuwen suggest that illustrated documents of a variety of kinds can meaningfully by analysed in terms of the following 'signifying systems' that structure the information on the page:

Salience	Elements are made to attract attention to different degrees, articulated through foreground/background placement; size, contrasts in tone and colour, difference in sharpness
Framing	Framing devices such as boxes and lines disconnect and distance elements from one another; connective devices have the opposite effect
Information value	Placement of elements in particular 'zones' (given, new; upper, lower; right-left; centre, margin) in the visual space endows them with particular meanings

Kress and van Leeuwen's observations on salience and framing seem entirely plausible and find much agreement in the field of document design—for example, *salience* appears well embedded in the visual processing system involved and *framing* has antecedents in Gestalt thinking on layout, which suggests that items that are physically similar or close to one another are seen as a group (for a summary, see Bruce and Green (1985); see also Sarkar and Boyer (1993) for a discussion of the use of grouping in computer vision, and Schriver 1997:303ff in relation to document design). The notion of information value is, however, not so clearly motivated. It proposes that particular semiotic values are realised by particular configurational and locational properties of the elements that carry those values. The values distribute themselves across the page as suggested in Figure 1 below.

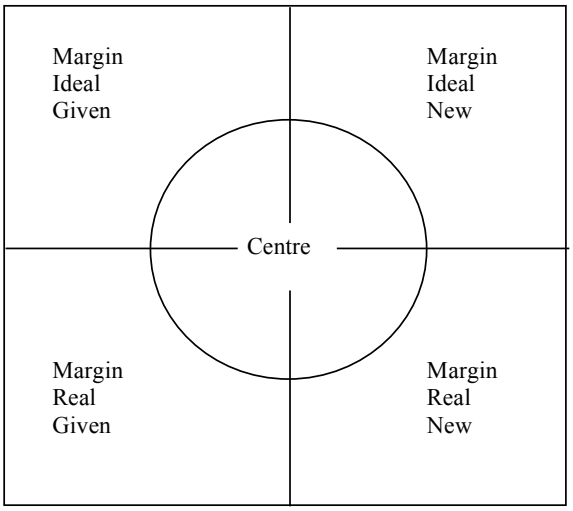

Figure 1: Dimensions of page layout. From van Leeuwen and Kress (1995:31)

Figure 1 expresses the 'zones' of information layout that Kress and van Leeuwen suggest, each of which, they claim, 'accords specific values to the elements placed within it' (Kress and van Leeuwen, 1998:188). We can summarise the 'specific values' of these zones as follows. The 'Given/New' distinction, drawing as it does on generalisations about information units in language offered by Halliday (e.g. Halliday (1985)) and relating specifically to points of intonational prominence, is used by Kress and van Leeuwen to describe oppositions between elements placed on the left of a page or image, and those placed on the right (again by analogy with linguistic information units, 'Given' is left, and 'New' is right):

Given — Presented as material the reader already knows; 'common sense and self-evident…presented as established' (1998: 189);

New — Presented as material as yet unknown to the reader; 'the crucial point of the message…problematic, contestable, the information at issue' (1998:189).

The notion of 'Real' and 'Ideal' is suggested by Kress and van Leeuwen as a means of theorizing the opposition between the top (Ideal) and bottom (Real) of the page or image:

Ideal — Presented as the 'idealized or generalized essence of the information… general, abstract, theoretical' (1998:193-4);

Real — 'More specific information (e.g. details) and/or more 'down to earth information…evidence…practical information ' (1998:193).

Finally, the orientation of a layout around the Centre-Margin distinction is reported by Kress and van Leeuwen to be 'relatively less common' in Western layouts (1998:196), but, when it does occur, the Centre is suggested to present information as 'the nucleus of the information to which all the other elements are in some sense subservient'.

The analysis is appealing in that its concern with values such as Given/New, Ideal/Real, etc. provides a ready vocabulary for reading more out of page design than would otherwise be possible. Just as the analysis of English clauses into a Theme/Rheme structure, in which the element(s) placed at the beginning of the clause have been shown to participate with high regularity in larger text-structuring patterns (cf. Fries (1995)), the Given/New, Ideal/Real, Margin/Centre patterning appears to offer a similar analytic win for the page.

But to what extent is the claim supported? Indeed, how would it be supported? The use of given/new is very much more abstract than that generally found in clause (or intonational unit) analyses; for Kress and van Leeuwen the given/new involved revolves more around problematised breaks in the social norms expected. The analytic procedures for establishing to

what extent this could be a reliable property of layout rather than an occasionally plausible account are unclear. Nevertheless, following on the initial presentation of the analytic scheme in van Leeuwen and Kress (1995), it has been presented again in Kress and van Leeuwen (1996, 1998) and is now itself being adopted as unproblematic, or 'given', in some systemically-based research on multimodality (see, for example, Royce (1998), Martin (2002)). Unfortunately, we have not so far found it to be supported by designers and layout professionals in practice. It is certainly not used as a design criterion in layout. What, then, is its status?

We can see this problem particularly well in the area of newspaper design, the area within which Kress and van Leeuwen's proposal was first couched. In one of their analyses, in which they deal with a Guardian front page, they attribute the positioning of a prominent photograph as contributing the 'Ideal' with respect to the 'Real' provided by the related story appearing beneath. Similarly, in an analysis of a *Daily Mirror* front page (1998:190ff) , they attribute the 'opposition' between an article about a murder on the left hand side (their 'Given' position: because we 'expect' newsstories about murders and other violent activities) with a story about Michelle Pfeiffer adopting a baby on the right hand side ('New' position: famous film star acts like a mother) to Given-New organisation:

> 'Given, then, is the bad news: an instance of discord between lovers, with dramatic results. This is what we are exposed to every day in press reports about everyday 'private' relationships: infidelity, breakups, abuse. New is the good news…' (1998:190)

This is a good example of what we are referring to as 'impressionistic interpretative' analyses. The story told is an appealing social interpretation of a multimodal product—but it has not yet been established whether such an analysis is actually any more than a post hoc rationalisation of design decisions that occur on a page for quite other reasons.

For example, we find that the practical workflow of newspaper production would most often mitigate against a reliable allocation of the areas of the page so as to conform with the semiotic values that Kress and van Leeuwen have proposed. First of all, the relevant unit of analysis from the *production* point of view is not the page in its entirety: it is what has been termed the 'newshole' (Lie, 1991)—i.e., the area that is available once advertising, mastheads, and other fixed elements have been allocated[2]. Advertising in particular is produced independently in newspapers, and (again, from a production point of view) it seems to over-read the results if an attempt is made to explain advertisements as part of the 'Real' when they occur at the bottom of the page when they most often need to be placed here in any case so that when the newspaper is folded in half and placed for sale at the newsagents enough of the newshole remains visible to sell the newspaper. Second, while a principle certainly exists to place the top story at the top of the paper, this is also part of its function to sell the newspaper—and therefore again has to be visible when folded; it also often runs from top to bottom—which would take it from the 'Ideal' to the 'Real', which would be a novel analysis for the structure of hard news items.

It may be the case that there is a 'good-bad' opposition working between the stories discussed by Kress and van Leeuwen, and this may have influenced their selection by the editor or editorial team for the front page, but we do not immediately see how the 'Given-New' nomenclature can be justified in explaining the assumed opposition between the left and right of the page. And, again, an important reason that newspapers place photographs above the fold on the front page is so that potential buyers are attracted to the paper when they see it folded on the newsstand.

It is unlikely that what is 'ideal' is what will geneally sell the newspaper. The masthead, as generically the topmost element on the page, can also be read as 'ideal' in that it names the

---

[2] In other genres, the area conceived of as available for layout may not be a page at all: it may be a spread, a run of pages, or a screenful.

newspaper and sometimes what it stands for—but how much this adds to an analysis is again unclear. Given the number of other reasons we can find for the masthead being where it is, the analytical 'bite' of attributing it to the 'ideal' is considerably weakened. Moreover, we have little doubt that had the design of the concrete pages discussed been reversed, an equally 'convincing' story could nevertheless have been told—again employing terms such as given/new, ideal/real.

The fact that news editing and the concrete practice of newspaper production do not involve explicit conceptualisations in terms of given/new or ideal/real does not, of course, mean that these categories are not employed by readers. The historical process of development in layout design may well have brought about a situation in which the semiotic values proposed by Kress and van Leeuwen hold regardless of the intentions of layout designers. But in this case, we must, on the one hand, be able to investigate readers' responses to layouts in order to provide support (or otherwise) for this interpretation and, on the other, be able to trace the development of the semiotic practise over time to see how it arose. Both investigations are scarcely possible without a tighter hold on the data that is being questioned.

We need then to ask the questions concerning the semiotic values and their realisation in layout that have been proposed by Kress and van Leeuwen more precisely. Is the entire scheme to be dismissed as a suggestive idea that did not work? Or, does the scheme apply to certain kinds of documents and not to others? Or to certain kinds of page layouts and not to others? Or, is it not to be seen in the same way as a structuring device analogous to Theme and Rheme in the grammar of English at all: that is, its function is suggestive of analyses and is not something that is 'falsifiable' by negative instances? All of these issues need to be addressed and answered as multimodal document analysis moves away from the suggestive and towards the analytic. Methods need to be adopted and documented whereby suggestive frames of analysis can be expressed as predictive and falsifiable claims about document design and meaning making.

2.2 The rhetorical motivation of layout design

Our second example is drawn from a very different area—that of practical document design as described by Schriver (1997) . We consider this text in particular to be a central one in setting out many aspects that are crucial for the analysis of meaning-making in page-based documents. In contrast to Kress and van Leeuwen's more theoretical analysis, Schriver's concern is to present usable guidelines for practical and effective design.

One of her recurring recommendations echoes a claim that is now increasingly heard in modern information and document design—that is that the layout should fit the communicative goals of a document. This is described by Schriver in terms of the 'rhetorical cluster', which she defines as:

> "a group of text elements designed to work together as a functional unit within a document. Rhetorical clusters act as reader-oriented modules of purposeful and related content. They are comprised of visual and/or verbal elements that need to be grouped (or put in proximal relation) because together they help the reader interpret the content in a certain way." (Schriver, 1997:343)
>
> "The key feature of a rhetorical cluster is that its elements interact as a Gestalt; the elements have structure and the parts are interdependent. Every element in a rhetorical cluster influences the interpretation of others (e.g., a caption for a picture significantly constrains the world of intepretation for the picture)." (Schriver, 1997:344)

This notion is also very appealing. It establishes a clear connection between high-level communication functions and precise decisions made in the presentation of information to fulfil those functions. It is based on Schriver's extensive experience with both designing documents and evaluating how well documents designs 'work'. Appealing to notions of rhetorical organisation provides a powerful tool for document critique: document layout that fails to support some distinction in meaning that the document is meant to be drawing may be

considered inferior. In our desire for order, this seems entirely convincing: it is only logical that a layout that 'supports' the logical development of an argument or content should be preferred over one that does not.

But, for all its appeal, the account has some holes theoretically. First, how does one know what the rhetorical or communicative goals of a document are in order to analyse whether, or how well, the layout conforms to them? Second, to what extent is this proposed ordering of layout according to rhetoric a *necessary* feature of design and to what extent is it a fashion that applies to some times (and to some genres) more than others? While this is not so much of an immediate problem for Schriver—she can rely on designers having a good informal sense both of what their communicative goals are and of current fashions—it is a problem for us as we try to push our theoretical grasp of the phenomena further. Unless we can construct theoretically motivated and repeatable rhetorical analyses of documents, we can say little about the meaning-making potential of layout and its possible relations with rhetorical or communicative function.

This is important because it is quite unclear *to what extent* a 'good' layout must conform to the rhetorical purposes or clusters of a document. The results of our own analyses of documents, to which we will return below, show that it is very common for document design to deviate from fine-grained rhetorical analyses of those documents. It does not appear to be the case that an absolute relationship between rhetorical distinctions and layout design decisions can be upheld as a general rule. A detailed rhetorical analysis, if carried out, soon produces distinctions that seem too fine-grained to be carried by layout distinctions. If the mismatch between layout decisions and rhetorical distinctions occurs at this level of detail, it is unclear whether it makes any measurable difference to the perceived meaning of the document. But where the line should be drawn here, and why, is unknown. An entire constellation of questions concerning how closely design decisions should reflect rhetorical purposes, what might influence the required closeness of fit, whether this is subject to fashion or genre, and many others therefore remains to be addressed.

2.3 Interim lessons drawn

The examples of this section could be expanded to include almost all proposed analyses of multimodal meaning. In the best cases, to which these examples belong, we have very plausible and/or interesting claims, which now need refinement and more extensive testing to stand as solid results. To what extent are we revealing details of multimodal meaning making and to what extent are we 'simply' providing sometimes plausible stories about particular kinds of documents (or even of individual documents)?[3] We believe that to make further progress, we need to take the step of making multimodal analysis more strictly corpus based. We need to subject the plausible stories to more detailed and systematic investigation, varying types of documents, types of consumers, types of presentation medium, purposes so that we can get a finer grip on the meaning-making possibilities of the various semiotics in play.

We do not think that this particular direction of 'empirical foundation' is in any way at odds with the approaches of Kress and van Leeuwen or Schriver. Kress and van Leeuwen also base their proposals on their extensive experience of document design; the proposals did not arise out of pure data-free introspection. Moreover, Schriver already presents extensive empirical support for her claims—support principally drawn from psychological investigation of visual perception and, more directly relevant to document design, from empirical investigations of readers' responses to documents, usability tests (in the case of, for example, instructional texts or websites), and so on. This allows reliable quantitative results concerning whether a changed

[3] Note again that we are *not* denying that plausible stories can be useful. Schriver's very plausible account of how documents work is arguably exactly right for formulating practical design guidelines. A designer is not going to set out a detailed rhetorical analysis whenever a document is to be produced. When we wish to theorise the notions of meaning-making concerned, however, we are compelled to go further.

design improves the usability of a document for some purpose or not. This is valuable support for the guidelines that Schriver gives.

Our point is then simply that we believe this empirical basis must be given a more prominent role. After establishing likely hypotheses by informal analyses of selected documents, the next methodological step must be an automatic extension to consider systematically how the hypothesis fares with an extended data set. This can only be achieved as more data becomes available; just as linguistic research is now greatly aided by the ready availability of prepared corpora, the task of providing such corpora for aiding multimodal analysis now also needs to be pursued with a high priority. Furthermore, we consider it an essential additional step that such corpora go beyond being simply collections of material. As we will describe in the next section, linguistic corpora gain considerably in their usefulness when they are structured and augmented to support investigation. We also need methods and techniques for achieving this for multimodal corpora.

In fact, it appears to us that the existence of appropriate corpora is even more critical for multimodal analysis than it is for linguistic analysis; consider one of the early criticisms of corpus linguistics (cf. that of Chomsky in the early 60s as discussed in McEnery and Wilson (2001)) that one does not need to use a corpus because one has access to the intuitions of native speakers and that, the argument went, must be more comprehensive than any of the data contingently making up any given corpus. This criticism may have some relevance and needs to be borne in mind when evaluating results of corpus analyses. However, where are the "native speaker intuitions" in multimodal meaning making? Here we lack convenient cut-off criteria such as 'grammaticality' that can be interrogated. Questions of good or bad do not suffice for detailed analysis and theory construction. We need to inspect more broadly just what occurs in multimodal documents in order to map out the dimensions of variation that actually occur.

And finally, as a further, possibly rather controversial, distinction between linguistic analysis and multimodal analysis, it appears that simply collecting a larger data set on the basis of documents found in the world will not automatically mean we have access to what is 'correct' or valid. The fact that there are no 'native speaker intuitions' has a down side: there is a substantial body of documents exhibiting what any design expert would consider as 'substandard' (or even incorrect) design. What are we to do with these data? As an analogy, it would seem that we are faced with a 'linguistic community' in which what we might term 'adult linguistic competence' is rarely achieved! Or perhaps we must recalibrate our standards: basic multimodal meaning-making competence may be very much more restricted (or of a different nature) than 'sophisticated' design suggests. This leads to important questions of education and changes in awareness concerning the semiotics involved. And again, we need access to a far more representative range of multimodal meaning making practices even to begin sorting these issues out.

3. Multimodal annotated corpora

In order to provide a solid empirical basis for investigating questions of meaning making in multimodal documents, we need to construct extensive collections of data organised in a manner that supports this inquiry. Here we can draw on the experiences gained with traditional linguistic corpora. With the extensive collections of texts now available, it is fast becoming part of everyday linguistic work to collect corpus instances of phenomena or patterns of concern in order to guarantee a broader and more objective basis for hypothesis formation, theory construction and verification.

In this section, we introduce the important notion of *annotated* corpora—that is, collections of texts that are augmented structurally so as to support investigation of linguistic questions more readily than would simple collections of text.

3.1 The origin and representation of annotated corpora

Linguistic corpora containing collections of several million words are fast becoming the norm (the British National Corpus, for example, contains 100 million words). With this mass of available 'data', it is increasingly important that the data be organised so as to support, rather than hinder, scientific inquiry. Standard introductions to corpus linguistics provide many examples of just why this is essential and how it is achieved (cf., e.g., Biber, Conrad and Reppen (1998), McEnery and Wilson (2001)).

One simple illustration involves the problem of variant linguistic forms that do not play any role in an inquiry being pursued but which make the posing of questions to a corpus more complex. If, for example, we are seeking all occurrences of the verb 'buy' in order to see what complementation patterns it occurs in, or which collocations it supports, we cannot just ask a text collection to print out all occurrences of the string of characters 'b-u-y'. We cannot even ask it to print out all occurrences of the word "buy"—because in both cases we then do not get forms such as 'bought' and in the second case we miss forms such as "buys", "buying", etc. While relatively straightforward to avoid, such minor problems reoccur with every inquiry that one wishes to make of a corpus and easily lead to error or incomplete results.

A further illustration, a little more complex, is how to deal with a linguistic inquiry concerning uses of the modal 'can'. We can ask to retrieve all instances of the word 'can' from a corpus—but then how do we avoid all the (for this particular question irrelevant) instances of the noun 'can'. Again, we can do this by hand, ruling out the irrelevant cases, but this work reduces the effectiveness of using a corpus and represents a considerable overhead. More sophisticated still, if we wish to investigate the contexts in which some grammatical construction is used rather than another, then we need to be able to search for such constructions rather than particular words or sequences of words and this can be quite a difficult undertaking.

In all these cases, modern corpora provide direct support for investigation by *annotating* their contained data to include additional information that may be employed when formulating questions. That is, not only will a corpus contain the bare textual information, it will also contain information about the root form of the words used (thus enabling a single question about all occurrences of the word 'buy' in *any* of its forms), their word classes (thus enabling a question exclusively about modal 'can') , and possibly some grammatical structures or other information in addition (as supported by the growing availability of 'treebanks': cf. McEnery and Wilson (2001)). The provision of corpora viewed as collections of texts is thus giving way to *annotated corpora*, which contain additional information for the asking of more exact linguistic questions.

In the earlier days of corpus construction annotation was achieved on a case-by-case basis. Contrasting such corpora shows very different representational techniques in use. This is illustrated in the contrast between Figure 2 and Figure 3, which show extracts from the Lancaster-Oslo-Bergen (LOB) corpus and the SUSANNE corpus (Sampson, 1995) respectively. The LOB corpus is annotated to show part of speech information by means of a so-called **tagset.** This is a more or less complex classification scheme that identifies the classes of the individual words indicated. These tags are attached to the words as they occur in the corpus. Clearly, in order to find either the original text or to use the part of speech information, particular support software is necessary.

The form of the SUSANNE corpus is more complex. It is organised in tabular form with the original text maintained only in the third column of the information given. The other columns provide information that makes the corpus more easily usable for a wide range of linguistic investigations. The first column identifies each word uniquely within the corpus, the second specifies the part of speech of the word, again by means of a complex tagset, the fourth column gives the root form of the word (e.g., in line 3 'becoming' is identified as a form of the verb *become*), and the fifth indicates, again via a complex classification and distributed phrase

structure bracketting, aspects of the grammatical structure involved. These two examples show something both of the diversity of representations adopted and of the additional information maintained.

```
P05   32  ^ Joanna_NP stubbed_VBD out_RP her_PP$ cigarette_NN with_IN
P05   32  unnecessary_JJ fierceness_NN ._.
P05   33  ^ her_PP$ lovely_JJ eyes_NNS were_BED defiant_JJ above_IN cheeks_NNS
P05   33  whose_WP$ colour_NN had_HVD deepened_VBN
P05   34  at_IN Noreen's_NP$ remark_NN ._.
```

Figure 2: Extract from the LOB corpus

```
J04:0230c-  NN1u  resonance     resonance   .Ns:s102]
J04:0230d-  VBZ   is            be          [Vzu.
J04:0230e-  VVGv  becoming      become      .Vzu]
J04:0230f-  YG    -             -           [Tn:e<S102.S102>
J04:0230g-  RR    increasingly  increasingly [R:h.R:h]
J04:0230h-  VVNt  used          use         [Vn.Vn]
J04:0230i-  II    in            in          [P:p.
J04:0230j-  NN2   investigations            investigation [Np.
J04:0240a-  IO    of            of          [Po.
J04:0240b-  NN1n  structure     structure   .Po]Np]P:p]Tn:e]Fa:c]S]
J04:0240c-  YF    +.            -           .
```

Figure 3: Extract from the SUSANNE corpus

This diversity meant that corpus designers of individual corpora might make very different decisions in the construction of their corpora and that tools developed for asking queries and examining results for one corpus would not be re-usable for others. In addition, each corpus would commonly adopt its own tagsets, requiring users to be familiar with a range of incompatible classification schemes.

In more recent annotated corpora, it is now usual to employ some kind of explicit *markup language* in order to capture the extra information they contain. That is, the basic textual information is 'marked-up' with the additional information to be represented drawing on a standardised format and increasingly standardised proposals for specific tagsets. This separates very clearly *data* from information *about that data*—which makes the information as a whole considerably easier to process and manipulate. We can see that this is part of the problem involved in the examples of Figure 2 and Figure 3: it is not clear to the reader, and hence to any particular computational tool for using the data, what aspects of the corpora correspond to the original data and which are added information. Extensive details concerning current approaches to such linguistic mark-up are available in the introductions to corpus linguistics mentioned above; we will not be concerned with this kind of mark-up further here however.

This general task of 'adding' information to a textual basis in a way that maintains the clear distinction between data and annotation has been a concern of the publishing industry for a long time. There, various kinds of information such as formatting or layout attributes, or notes of differences between editions, or keys providing clues for search engines so as to be able to retrieve certain specified classes of documents are all generally regarded as useful value-adding components of any body of textual information. For this purpose, a standard form for mark-up was developed, called the Standard Generalised Mark-up Language, or SGML (cf., e.g., Bryan (1988)). SGML provided a very powerful way of specifying such additional information in a commonly agreed and standardised format so that generic tools could be employed for a variety of purposes and by a variety of users. SGML was quickly established as the sensible target for electronically stored textual information.

SGML is 'generalised' in the sense that it supports the definition of specific mark-up languages for particular purposes: a user can define the particular classification scheme to be used for

annotating data (again in a standardly defined format) and generic tools can interpret this definition in order to 'understand' the corresponding annotated data. The most commonly known specialisation of SGML is the World-Wide Web hypertext definition language: HTML, where annotations provide information about basic text structure, desired appearance, and hypertext links to other HTML documents. Such specialisations of SGML are expressed in terms of *Document Type Descriptions* (DTDs), which specify precisely the structures that are possible and the kinds of entities that can fill slots in those structures. One of the reasons for managing things in this way is that it allows documents to be *checked for conformity*. This process is called document **validation**. It is by no means straightforward to guarantee that any sizeable collection of information in fact contains no errors even of a formal, structural kind: this is the kind of service that a DTD provides. Standard DTD-parsers check documents for conformity with their specified DTD so that at least formal errors (such as misspelled tags, inconsistent or incorrect structures, missing tags, etc.) may be avoided. As we will see below, corpus annotation results in complex webs of interlocked information and it is virtually impossible to maintain by hand such bodies of information without error or inconsistency.

Since corpora as used for linguistic research are also collections of electronically stored texts or text fragments, and it is clear that additional mark-up information is essential for their use, it was logical to consider the application of SGML to this task. This was undertaken as part of the *Text Encoding Initiative* (TEI) of the Association for Computational Linguistics, the Association for Literary and Linguistic Computing, and the Association for Computers and the Humanities (Sperberg-McQueen and Burnard, 1994). The TEI defines guidelines that specify the desired form of machine-readable texts. This has subsequently been built upon by one of the EU Expert Advisory Groups on Language Engineering Standards (EAGLES) in order to provide specialised guidelines for corpus construction. This enables corpus designers to prepare and construct corpora more systematically and in ways that supports their re-use within a wider research community. The result is called the *Corpus Encoding Standard* (CES: Corpus Encoding Standard, 2000), which sets out several layers of increasingly demanding mark-up that corpora should adopt for expressing their additional information. Thus, whereas TEI specifies the form of machine-readable documents in general, the CES provides particular details about the content of the additional information that needs to be represented when building corpora. Software now being designed for corpus processing can be specifically tailored to the more precise specifications given in the CES while still maintaining very broad applicability—i.e., applicability to any corpus that is (provably, via the given DTDs) *CES-conformant*.

While SGML was an enormous step forward in the preparation of electronic documents, it suffered, and still suffers, from one significant drawback: it is extremely complicated. The computational tools available for processing SGML are restricted to expensive high-end publishing applications. And yet the aim of SGML, to structure data in meaningful ways so as to support the use of that data for a variety of purposes and tasks, is even more important nowadays than it was before. This is one reason why use of the very much simpler HTML used for marking-up documents for the World-Wide Web was, in contrast to SGML, so quick to establish itself. In opposition, however, to the goals of SGML by which data and information about that data would be kept separate, HTML was soon forced to carry a far greater load. Since HTML was the only means of publishing information so that it could be understood by web-browsers, web designers started using it for a variety of purposes—including explicit statements of how the information should appear when presented on the screen. This means that it now commonly the case that, if we examine the source HTML code of a document displayed on the web, it is very difficult to distinguish the 'content' presented in that document from the specification of how that content is to be displayed. The desired separation of data and information about the data has been compromised.

This mixture of data and display information has a number of drawbacks that are well known in the information presentation industry. It is also clearly reminiscent of the problem with annotated corpora seen above—if the tags and the data are mixed, then it is difficult to know what to search for. However, significant for us here is only the fact that now a further step has been taken precisely to correct the situation. A new specialisation of SGML, called the 'extensible mark-up language' (XML), now provides enough of the generality of SGML to define richly structured annotation schemes for all kinds of data, while at the same time avoiding some of the over-complexities of SGML that prevented its widespread use. Moreover, very careful attention is being paid to mechanisms for presenting information structured with XML in far more flexible ways than ever possible with HTML; almost as an unintended side-effect, we will see below how some of these mechanisms are now extremely relevant and useful for the linguistic analysis of XML-annotated corpora. The signs now are that most of the content on the World-Wide Web will move in the short-to-medium term to involve XML-represented data; there are accordingly already many computational tools for processing XML data and presenting it in a variety of forms via the World-Wide Web or other delivery mediums.

This increasing acceptance of XML has motivated the latest development of the Corpus Encoding Standard, which is itself now represented fully in the terms defined by XML and, in this version, is called XCES. Modern corpora all now generally employ SGML/XML-compatible forms of mark-up for their additional linguistic annotations. The formal correctness of very sizeable bodies of data can be guaranteed via the conformity checking provided by DTD-parsers, since XML, like SGML, encourages DTD specifications of the form of annotations that are to be used in a document or set of documents. The adoption of XML also allows them to develop and employ state-of-the-art display and data manipulation tools as they become available. As we shall see below, as the information content of a corpus expands, then the less the corpus documents come to look like the original texts. But this is not allowed to become a problem precisely because the corpus is *not intended* to be read directly by a human reader—the point of providing additional information is to make it easier for computational tools to be developed that strongly support linguists in their search for linguistic patterns. The view of the corpus is always to be mediated by appropriate tools and XML makes the provision of such tools very much easier than hitherto. Increasingly powerful concordance and corpus-search programs that are based on XML will be applicable to any corpus represented in an XML-conformant fashion. The benefits of standardisation in this area are therefore very tangible and are set to significantly improve the design and use of linguistic corpora in all areas over the coming decade.

3.2 Annotation problems with complex data

The basic organisation of a document written in XML is very simple. Information is structured by means of **tags** in the same way as information for web pages in HTML. A piece of information is marked with a certain tag by enclosing it within an opening tag and a closing tag. If, for example, we are marking a body of text according to XCES as a single paragraph, we use the 'p' tag. The opening 'p' tag is written as and the closing 'p' tag as . Unlike HTML, in XML the closing tags are compulsory; this makes complicated documents easier to validate against their specified document type descriptions. An example of an annotated fragment of text taken from the CES documentation is given in Figure 4. This shows use of the paragraph tag and the quotation tag ('q') and gives an indication how explicit structure is imposed on raw text.

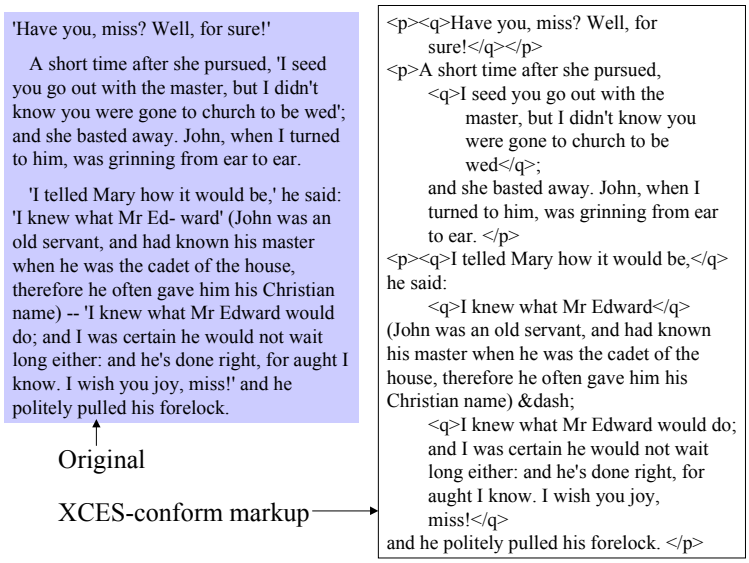

Figure 4: Example of XCES-conformant annotation

We can then begin to query such text collections directly in terms of their logical structure rather than contingent features of the original: for example, if we were using a collection of texts where some editions used a single quotation mark for direct quoted speech whereas others used a double quotation mark, perhaps with distinguished open and closing quotes, then this would need to be known and considered in search for quotations: the CES specification allows us simply to look for quotations no matter how they are represented in particular texts. Texts maintained in this fashion become increasingly useful the more annotation that they contain. The CES suggests several levels of annotation—the first, and minimal level, is as suggested in Figure 4, while the higher levels require basic linguistic information to be marked-up as well, which brings us closer to the information content of linguistic corpora motivated above.

To support a richer variety of information in the annotation, tags may also specify **attributes**. For example, not only may we specify that a particular element in the corpus is a word—perhaps using the CES tag 'w'—we may also give it a unique identification number, specify its part of speech information, and its root form with a complex mark-up such as the following:

```
<w id="J04:0230e" pos="WGv" lemma="become">becoming</w>
```

This is then the XML equivalent to the third line of Figure 3 above drawn from the SUSANNE corpus. The precise attributes that are allowed and the kinds of values that they may take is again specified formally in a Document Type Description, which allows formal validation for this information also.

Whereas the use of SGML/XML represents a major advance for the design of large-scale corpora, there are some problems with capturing the required linguistic (or other) information in ways which remain faithful to the requirements imposed by XML. As long as these (largely formal) requirements are met, then standard tools can be used for processing the data; this is extremely beneficial because of the already very large and growing community of XML users and developers. When data does not conform to the XML specification, it can no longer be processed by the available generic tools and its use is restricted. An appropriate analogy is the use of HTML for web-pages: as long as someone uses standard HTML, then they know that anyone, using a standard browser, will be able to see their information offering; but, as soon as they depart from HTML, then their potential audience is cut dramatically. It is therefore highly desirable to find representation solutions that do not go beyond the standard.

The main problem encountered when attempting more sophisticated linguistic annotation is that of *intersecting hierarchies*. One of the basic formal requirements of XML is that tags must 'nest properly'. That is, when representing structured data, the structures must properly fit

inside each other—there can be no overlapping or intersecting boundaries. This is not only a problem with linguistic annotation, there are many text annotation applications where it is not possible to force the structures that are to be represented to nest one within the other.

A good example of this problem from the area of annotation for literary editions is suggested by Durusau and O'Donnell (submitted).[4] One simple TEI-conformant mark-up of the linguistic content might break a document down into a number of identified sentences; this would use a sequence of <S> and matching closing </S> tags. Another simple TEI-conformant mark-up might want to indicate the division into pages that an edition employed—here we would use a sequence of <page> ... </page> tags. Now consider an annotation for a machine-readable version of the literary work that wants to capture the page breaks *and* the linguistic divisions simultaneously. This is not straightforward simply because the linguistic division into sentences and the division into pages have no necessary relationship to one another: there is no reason why the structures imposed by the two kinds of division should embed one within the other. The simplest way of capturing this information might appear to be something like the following:

> <page> ... <S> This is a sentence </page> <page> that goes over two pages. </S> <S> Then there are more sentences on the page ... </page>

But this is not 'legal' XML: the structures defined by the <S>-tags and the <page>-tags do not 'properly nest'. The first sentence tag is not 'closed' before its enclosing page tag is closed. Allowing such non-nesting structures would vastly complicate the machinery necessary for checking document conformance.

But this is perhaps also a good illustration of the valuable role of formalisation: the fact that allowing non-nesting structures would result in very much more complex machinery gives us a hint that perhaps this is not the right way of doing things. Indeed, why should a single document contain such incompatible kinds of information? Imagine that we take the example further and wish to add in other information: for example, the pages breaks of differing editions (each of which may have an arbitrary relationship to the page breaks of others), or the intended (or actually used) intonational phrasings of the sentences if the text is to be spoken (as in a play). Each of these structures is compatible and simple when considered in isolation, but may relate in complex ways to the others.

A solution for this problem that has now established itself is that of *standoff annotation* proposed by Thompson and McKelvie (1997). The idea is straightforward, although its realisation in standard XML involves some more technical machinery. Essentially, standoff annotation recognises the independence of the differing layers of annotation and separates these both from the original data and from each other. Thus, instead of having a single marked-up document where the annotations are buried within the data, the annotation information is separated off into independent annotation layers—hence the phrase 'stand off'. Each individual layer is a well-formed XML document. Contact is made with the original data *indirectly* by referring to particular elements. This solves the problem of intersecting hierarchies because within any single XML document there is no intersecting hierarchy; there is only the single hierarchy of the particular annotation layer that the document represents.

The additional technical complexity involved is that we need to be able to access the individual elements of the data in order to bind them into a variety of annotation structures. This can be achieved most simply within XML by giving each element a unique identifying label and employing *cross-references*. This is shown in a simplified example in Figure 5 below, where we have two annotation layers that show how a single text document is divided according to

---

[4] Durusau and O'Donnell's example is actually rather more complicated. They also give an excellent overview of possible approaches and problems.

sentences and according to pages. This accepts the fact that the linguistic division into sentences and the print division into pages have no natural relationship with one another, making it inappropriate to insist that such mark-up nest properly into well-formed recursive structures simply to fulfil the SGML/XML formal restrictions. The situation illustrated takes the first pair of sentences from the document fragment used as an example in Figure 4 above and assumes further that there is a page break immediately following the text: "... Have you, miss? Well,".

This information is captured by breaking the original document into a set of 'basic level' annotation units—shown here on the left of the figure and consisting of words ('w' tags) and punctuation ('punc' tags)—each of which receives a unique identifying label as the value of their 'id' attributes. The two layers of standoff annotation shown on the right of the figure then refer to these labels. Thus, the first page—given its own identifying label of 'page-01'—is shown as running from some base unit that we have not shown in our figure up until the unit labelled 'u-07'. The second page then runs from unit 'u-08' onwards. In a complete annotation all of the units would have received identifying labels and so the cross-references would be complete. The other standoff layer shows precisely the same kind of information but for sentences. Each individual layer is a well-formed XML document and, because of the cross-references, there is now no problem when the distinct hierarchies fail to respect one another.

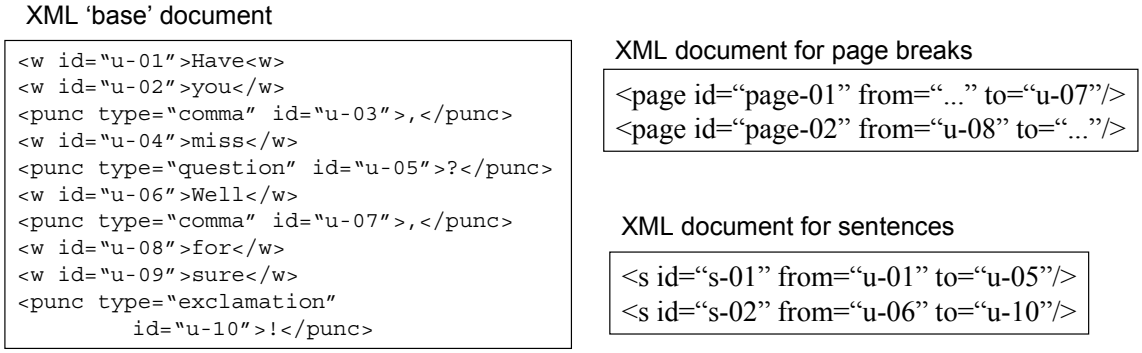

Figure 5: Example of standoff annotation

This example provides the basis for an open-ended set of annotation layers, each of which adds in further information to the base material. The utility of this method relies crucially on the effectiveness of the computational software for dealing with this kind of richly structured information. The fact that the entire framework is XML-conformant is very important. The tools for writing inquiries that interrogate data structured in this way are now being refined and extended extremely quickly. This is because the main users of XML structured data are not linguists, but standard commercial providers of information that previously would have been maintained in databases, such as sales catalogues of online companies, stock-lists, personnel data, and so on. Because of this very practical and economic demand, methods for using such data are already finding their way into the standardly available web-browsers—this virtually guarantees that it will soon be possible for annotated corpora to be navigated and manipulated using widely available and familiar tools rather than complex, corpus-specific schemes and software.

3.3 The state of the art: complex annotation for speech and video data

To conclude this section, we draw the strands that we have introduced together and set the scene for our return to our main concern in this paper: the provision of multimodal corpora suitable for supporting multimodal analysis. We consider XML as the standard form that should now be adopted for representing structured data in electronic form. Linguistic corpora that are not so represented simply cut themselves off from a rapidly developing set of tools and

techniques for maintaining consistence and for pulling out information under specified conditions for inspection. In fact most linguistic corpora now already respect the XML guidelines for structuring—either directly in that they are represented in XML or by specifying explicit mappings by means of which their data can be converted into XML and TEI/XCES-compatible forms.

Two of the most active areas for the further specification of guidelines for 'linguistic' corpora involve extensions beyond corpora as commonly maintained. The first goes back to a very early concern of corpus-builders: the provision of speech data. A number of guidelines are being developed for annotating live speech data in real communicative situations where there may be several speakers involved. The second is, in contrast, new and is itself a product of the fall of monomodality: this area is formulating guidelines for the mark-up and annotation of video data—including both 'constructed' data, such as film, and natural data, such as video-taped, multi-party interactions. A useful review of tools and techniques is given in Dybkjaer et al. (2001); some further approaches and applications are described in Baldry (2000).

The guidelines being constructed in these areas have some properties in common. In particular, it is natural for both kinds of annotation to employ a set of annotation 'tracks', in which different aspects of the data being annotated are maintained. For speech data, we might require a syntax track, an intonational track and, if the interaction is multi-party, particular tracks for each speaker. For video data, for example, there might be a music track, a gestural track, a movement track, as well as a language track that itself involves all of the tracks that we might expect for speech data. For all of these demands, stand-off annotation an introduced in this section is particularly well suited and is one of the most commonly used techniques.

A further demand for tools inspecting speech and video data is that the various tracks can be appropriately *synchronised.* That is, we need to be sure, for example, that particular gesture annotations can be placed in time alongside any simultaneous speech or other behaviour that is annotated. This is typically addressed by employing *time-stamps* whereby each annotated unit receives additional mark-up specifying when it starts and when it finishes. Since all the behaviour annotated unfolds linearly in time, this is an obvious and effective approach to take. Tools using the annotated data can then pick out information from all tracks that overlap in time and display these as required. When we consider applying such techniques to multimodal document analysis, however, we have a problem: the information on a page does *not* unfold linearly in time. In our further development of an annotation scheme for multimodal corpora, we cannot use the existing approaches for speech and video data. We have needed to develop other ways of structuring the data and it is to this that we now turn.

4. The Gem Model: layering for classification.

In this section, we set out how we are approaching the design of multimodal corpora drawing on the state of the art for annotated corpora as described in the previous section. We have been pursuing these aims in the context of a research project, the 'Genre and Multimodality' project GeM (http://www.purl.org/net/gem).[5] The basic aim of GeM is to investigate the appropriateness of a multimodal view of 'genre': that is, we are seeking to establish empirically the extent to which there is a systematic and regular relationship between different **document genres** and their potential realizational forms in combinations of text, layout, graphics, pictures and diagrams. This is to take traditional views of genre, as developed in literary studies and linguistics, and to extend the concept to include documents that are not restricted to monomodal linguistic products. We begin this section with a brief introduction to the approach adopted within the GeM project and its motivation, and then move on to how this is concretely

[5] 'Genre and Multimodality: a computer model of genre in document layout'. Funded by the British ESRC, grant no. R000238063.

used in the design of a multimodal corpus appropriate for empirical research. More detailed introductions to the GeM model and its motivation can be found in Delin, Bateman and Allen (2002) and Delin and Bateman (2002).

4.1 The GeM Model

Our starting point for considering genre draws on primarily on linguistic uses, such as, for example, that evident in Biber (1989) or Swales (1990). In general, we are investigating 'text categorizations readily distinguished by mature speakers of a language; for example…novels, newspaper articles, editorials, academic articles, public speeches, radio broadcasts, and everyday conversations…categories defined primarily on the basis of external format' (Biber, 1989:5-6). We take it as a basic premise that these categories of text also reflect distinctions in the author's *purpose*: the texts look different, and contain different language forms, because they are intended to do different things. We also emphasise and build on the social 'embeddedness' of genres. Texts also look different because they are to function in different social contexts (cf. Halliday (1978), Martin (1992)). And as a final step, we then reconnect this notion to the practical contexts of production and consumption of the discussed genres—that is, genres also are partially defined by their 'rituals of use' and the application of various technologies in the construction of their members.

The question of the identification of genres has always been a problematic one. It is difficult to locate particular formal features that are sufficient in their own right to define an instance of language as belonging to one genre rather than another. This is one reason why some proponents of genre theory—probably most well-known Swales (1990)—insist on an important definitional role for standard characterisations adopted by a language community. We see part of the difficulty of this problem as arising out of the complexity of the realizational relationship that appears to exist between the very abstract notion of genre and the particular linguistic phenomena that might accompany an instance of a genre's occurrence. As a framework we build on Martin's (1992) proposal of a stratified approach in which genre is realized in register configurations which are themselves realized in linguistic (semantic, lexicogrammatical, etc.) configurations.[6] The framework is in many respects 'variationist' in that we view a specification of genre as a description of a **space** of genre possibilities—movements within this space then pull the accompanying register configurations in various, systematically specifiable, directions. A similar relationship is to be posited between the space of register possibilities and their realization in linguistic patterns. For some of the theoretical modelling possibilities proposed for this and related views of genre, see Matthiessen (1993).

Recently attention has been drawn, particularly by 'generalized' linguists, to the possible role of genre within multimodal accounts. Here we extend the notion of genre space to include documents that employ a variety of semiotic systems within their realizations. Thus, for example, Twyman (1982) provides a preliminary scheme for categorising documents according to the interrelationships between images and text, while Kress and van Leeuwen have studied particular 'genres', such as the newspaper front pages mentioned in Section 2.1 above. Genre in this multimodal sense appears to be taking up a stronger role in their more recent work, too (cf. Kress and van Leeuwen, 2001). The GeM project is to be seen very much in this new 'tradition'. The basic goal of the project is then to contribute to an identification of the relevant **dimensions** for describing the genre space of multimodal documents.

One of the ways in which the study of multimodal genres appears to be developing differently to the study of linguistic or literary genres is in its focus on the practical contexts of production and consumption of the documents analysed. This is perhaps a natural consequence of

[6] Any directionality inherent in this formulation is not a commitment of the model and is not intended.

examining what are obviously 'social and cultural products' rather than disembodied texts and inherits directly from the kind of semiotics that emphasises the cultural and social situatedness of signs. The role, for example, of 'production' and 'distribution' is now particularly emphasised in Kress and van Leeuwen's (2001) proposals for multimodal discourse analysis. The first attempt that we are aware of that provided a detailed model of multimodal genre taking into consideration the vital contributions of language, document content, and visual appearance as well as practical conditions of production and consumption is that of Waller (1987). Our own work draws upon and extends this framework.

Waller's model is summarised in the diagram in Figure 6. The arrows indicate dependencies between the various areas of concern in document design. There is a task for the document to fulfil, there are generic constraints imposed by the culture, and there are various functional constraints on what would be an appropriate design solution—constraints including what is to be communicated (Topic constraints), what material form the document is to take (Artefact constraints) and the ways in which readers are to use the document (Access structures). Much of the complexity of document design arises from the interplay of these vary different considerations.

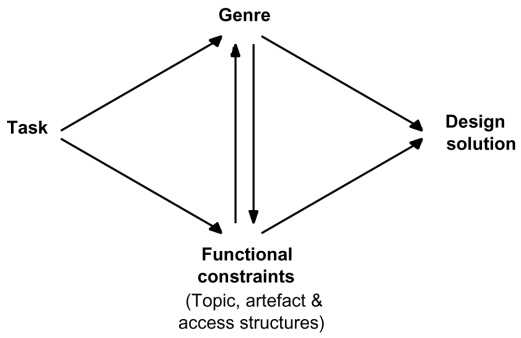

Figure 6: Model of document design developed in Waller (1987)

Within the GeM project we are seeking to take this further by examining the interdependencies between possible characterisations of genre on the one hand and of the various functional constraints on the other. We have found it necessary to extend and revise the functional constraints proposed by Waller in order to come more in line with current understandings of linguistic analysis. For example, Waller's Topic constraints, which he describes as 'the author's argument', need to be broken up into several related layers of constraint: 'the author's argument' is not solely or completely dictated by content: many rhetorical presentations are compatible with the same content. We also explicitly divide Access structure into particular aspects of the visual presentation making up the document and certain conditions of consumption. Finally, we take what Waller terms 'artefact structure' to be not a structure in the sense that it is a set of ideas to be incorporated in the document, but a constraint on the combination of all the other elements into a finished form. The basic levels of analysis that the project proposes are then as follows:

Content structure	the 'raw' data out of which documents are constructed;
Rhetorical structure	the rhetorical relationships between content elements; how the content is 'argued';
Layout structure	the nature, appearance and position of communicative elements on the page;
Navigation structure	the ways in which the intended mode(s) of consumption of the document is/are supported; and

Linguistic structure the structure of the language used to realise the layout elements.

We suggest that document genre is constituted both in terms of levels of description such as these, and in terms of constraints that operate during the creation of a document. Document design, then, arises out of the necessity to satisfy communicative goals at the five levels presented above, while simultaneously addressing a number of potentially competing and/or overlapping constraints drawn from:

Canvas constraints Constraints arising out of the physical nature of the object being produced: paper or screen size; fold geometry such as for a leaflet; number of pages available for a particular topic, for example;

Production constraints Constraints arising out of the production technology: limit on page numbers, colours, size of included graphics, availability of photographs; for example, and constraints arising from the micro- and macro-economy of time or materials: e.g. deadlines; expense of using colour; necessity of incorporating advertising;

Consumption constraints Constraints arising out of the time, place, and manner of acquiring and consuming the document, such as method of selection at purchase point, or web browser sophistication and the changes it will make on downloading; also constraints arising out of the degree to which the document must be easy to read, understand, or otherwise use; fitness in relation to task (read straight through? Quick reference?); assumptions of expertise of reader, for example.

A model of genre, therefore, must begin by expressing adequately the above five levels of description as well as finding the most appropriate way of satisfying the three sets of constraints. Particular genres are then constituted by regularly recurrent and stable selections and particular sets of constraint satisfactions. And these can only be ascertained *empirically* by the investigation of a range of document types.

4.2 Designing and populating a multimodal corpus

We have already seen the basic technological requirements sufficient for constructing a multimodal corpus. When we adopt the GeM layers of analysis, it is possible to consider each one as a single layer of standoff annotation just as was illustrated for the simple page and sentence example of Figure 5. This has now been done with Document Type Descriptions specified in XML-form for each layer. As usual with formalisation, the demand for complete specification has resulted in a considerable number of refinements to the original model sketched above. These are set out in full in the technical documentation for the corpus design (Henschel, 2002). Here we focus on just two layers of annotation: the layout structure and the rhetorical structure. The layout structure has been developed new within the GeM project; the rhetorical structure draws on a slightly extended form of Mann and Thompson (1988)'s Rhetorical Structure Theory (RST). For the purposes of this paper, we will also concentrate on the addition of *pages* involving multimodal content rather than go into the details of considering entire documents.

As we have seen, a precondition for standoff annotation is to establish a single document containing the marked-up 'basic units' of any document being added to the corpus. These base level units range over textual, graphical and layout elements and give a comprehensive account of the material on the page, i.e. they comprise everything which can be seen on the page/pages of the document. For the purposes of GeM, we have defined the base units as: orthographic sentences, sentence fragments initiating a list, headings, titles, headlines, photos, drawings, diagrams, figures (without caption), captions of photos, drawings, diagrams, tables, text in

photos, drawings, diagrams, icons, tables cells, list headers, list items, list labels (itemizers), items in a menu, page numbers, footnotes (without footnote label), footnote labels, running heads, emphasised text, horizontal or vertical lines which function as delimiters between columns or rows, lines, arrows, and polylines which connect other base units. Each such element is marked as a base unit and receives a unique base unit identifier. Details concerning the form and content of each base unit are not represented here—it is the job of the other layers of annotation to capture these details, including font types, sizes, colours, pictorial modalities, positioning and so on. The base units provide the basic vocabulary of the page—the units out of which all meanings on the page must be constructed.

All information apart from that of the base level is then expressed in terms of pointers to the relevant units of the base level. As suggested above, this standoff approach to annotation readily supports the necessary range of intersecting, overlapping hierarchical structures commonly found in even the simplest documents. The relationships of the differing annotation levels to the base level units is depicted graphically in Figure 7. This shows that base units (the central column) provide the basic vocabulary for all other kinds of units and can, further, be cross-classified as required to capture their multifunctionality—for example, units can contribute to a visually realised layout element as well as simultaneously functioning as a component of a rhetorical argument. This usage ensures that we can maintain the logical independence of the layers considered.

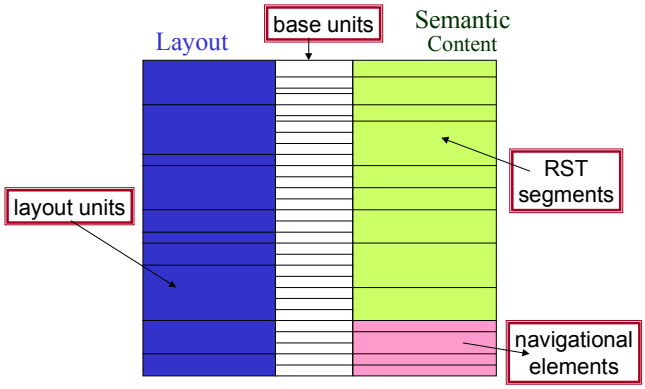

Figure 7: Relation of base units and standoff annotation layers

Thus, to take a relatively simple example, if we were adding the part of a page shown to the left of Figure 8 below, we would construct a base document along the lines of the XML annotation shown to the right of the figure.[7] Each typographically distinct element on the page is allocated to a different base unit; this means that collections of such units can, if necessary, be picked out by the other layers of annotation as carrying differing functions. The first unit (identified by the label 'u-01') corresponds to the headline at the top of the page extract; here we can see that the only information captured here is the raw text "£10m top of the range sale"—typographical information, placement on the page, rhetorical function (if any), etc. are not represented. The second unit does the same job for the large photograph—the 'raw picture' is represented indirectly by a link to a source file containing the image ('cuillins-pic.jpg') just as is done in HTML files for web presentation. The next five units describe the caption(s) underneath the picture; 'u-03' is an introductory label for the caption "Sea view:", 'u-04' and 'u-05' are two 'sentence'-like units making up the body of the caption, and 'u-06' and 'u-07' give information about the photographer. Again, the only role played by this division into units is to provide labels that subsequent layers of annotation can call on by cross-references when describing

[7] This page extract is selected from the front page of an edition of the Scottish daily newspaper, *The Herald*. It is reproduced by permission.

their functions on the page. Even the fact that the units are approximately ordered following their vertical ordering on the page is not significant—they could in fact be written in any order.

In contrast to the simplicity of the base layer, the others annotation layers are more complex. The layout layer of annotation is moreover itself internally quite complex. It has a number of tasks to perform in terms of capturing the layout decisions taken in a page. These may be summarised as follows. The layout structure must:

- capture all the particular typographical distinctions drawn on the page—such as, for example, the fact that certain elements are entirely in capitals, others are in bold, some are in one type face and others in another, and so on;
- represent the visually expressed hierarchy of related 'blocks' on the page—such as, for example, the relative grouping of a picture with its caption as a unit with respect to some other visual element for which the picture-plus-caption functions as elaborating material;
- relate the visual hierarchy of layout blocks to their concrete positions on a page of information.

Each of these kinds of information are managed as a locally complete XML structure. We show all three briefly for the selected newspaper fragment.

£10m top of the range sale

Sea view: the Black Cuillins, on the market for £10m. The clan chief has pledged the rugged range on Skye will stay in the public domain Picture: PETER JOLLY

Clan chief puts Black Cuillins on the market

RAYMOND DUNCAN

ONE of Scotland's most famous mountain ranges has been offered for sale at more than £10m by a clan chief.

The aim is to raise funds to repair an 800-year-old castle which is the seat of a leading clan.

Skye's Black Cuillins, one of the country's most spectacular landscapes and an area of major international importance, has been put on the market by the 29th Chief of the Clan MacLeod.

John MacLeod of MacLeod yesterday described the decision to sell as the most difficult of his life and "an extremely painful experience".

The move brought strong criticism from Western Isles MP Calum MacDonald. He said: "The idea of selling the Cuillins is ridiculous but at such a price it's obscene. MacLeod should hang his head in shame for trying to exploit what God has given the people of Skye. It goes to show that one thing worse than a thieving foreign landlord is a greedy Scottish one."

He also said it demonstrated the need to get land reform through Parliament at top speed to stop "the obscenity of landlords trying to cash in on public concern for areas like the Cuillins".

The 23,300 acres include more than 14 miles of coastline, two salmon rivers, a sheep farm and farmhouse.

News of the sale prompted a call from Highland Council for the mountains to be bought for the nation. It wants the "world treasure" bought by a partnership of public bodies, tenant farmers and local communities accessing Heritage Lottery Funding at District Valuer's valuation.

Potential partners could include the council, Highlands and Islands Enterprise, Scottish Natural Heritage and the National Trust for Scotland.

Councillor Michael Foxley, chair of the land and environment select committee, said: "We will be seeking a meeting in the very near future to set the ball rolling."

The 64-year-old owner of Dunvegan Castle said the move to sell the 35 square miles of rugged landscape was principally economic. Mr MacLeod called the mountain range "part of my soul". He vowed the Cuillins, once a temporary refuge for Bonnie Prince Charlie and now a paradise for geologists, botanists and mountaineers,

Continued on Page 2

INSIDE

The chief's castle – Page 2

```
<unit id="u-01">£10m top of the
      range sale</unit>
<unit id="u-02" src="cuillins-
      pic.jpg" />
<unit id="u-03">Sea view:</unit>
<unit id="u-04">The Black
      Cuillins, on the market for
      £10m. </unit>
<unit id="u-05">The clan has
      pledged the rugged range on
      Skye will stay in the public
      domain </unit>
<unit id="u-06">Picture:</unit>
<unit id="u-07">Peter Jolly</unit>
<unit id="u-08">Clan chief puts
      Black Cuillins on the
      market</unit>
<unit id="u-09">Raymond
      Duncan</unit>
<unit id="u-10">One of Scotland's
      ...</unit>

...

<unit id="u-70">Inside</unit>
<unit id="u-71">The Chief's
      Castle</unit>
<unit id="u-90">Page 2</unit>
<unit id="u-91" alt="line" />

...

<unit id="u-99">Continued on page
      2</unit>
```

Figure 8: Page extract from a newspaper and corresponding base unit annotation

The 'backbone' of the layout annotation is provided by the second of these kinds of information: the visually oriented hierarchy of layout elements. This is determined by a set of methodological heuristics for decomposing the information on the page. One such heuristic is

for the analyst to consider the relative visual prominence or salience of the blocks on the page. This can be supported by a range of 'tricks': for example, by progressively reducing the resolution of the image when displayed. The blocks which dissolve first are the lowest in the layout unit hierarchy (e.g., the smallest typographically displayed letters and words), those that dissolve into each other last are the highest level units of the hierarchy. A second heuristic is to consider which chunks of information 'belong together'—i.e., if one block were to be 'moved' on the page, which others are 'drawn along' with it. For example, if we were to move the photograph on the page, then it is natural that the caption would be drawn with it, and less likely that the body of the text or the headline immediately move: although there would be limits to this in the context of the page as a whole as the individual units making up this 'story' would not like to be separated. General proximity is thus to be maintained, which is itself an argument for maintaining all the units shown as a single higher-level layout unit. Furthermore, within this, the block in the middle of the second column of text stating that more information (of a particular kind: i.e., 'the Chief's castle') exists and providing navigation information about where that information is located ('inside' and 'Page 2') can also be moved relatively freely within its enclosing text block, arguing for its treatment as a distinct layout unit at an intermediate level in the overall hierarchy. Further examples of this kind of argumentation from page to layout hierarchy are given in Reichenberger, Rondhuis, Kleinz and Bateman (1995), which was the origin of our general approach to layout structure. We give an indication of this process for the example fragment in Figure 9 below.

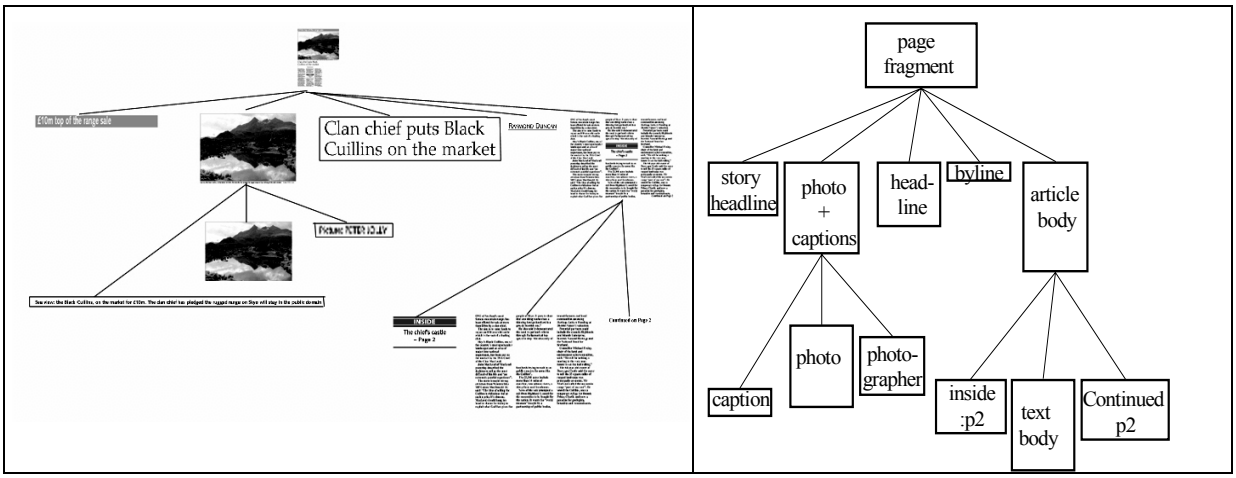

Figure 9: Derivation of hierarchical layout structure from the example page

In general, the hierarchical structures proposed should be conservative—that is, when there is no strong evidence in favour of a strict hierarchical relationship, we prefer to posit a flat structure rather than insisting on some particular hierarchicalisation. The layout hierarchy captures dependency relationships between visually discovered elements on the page but no longer includes information about the precise physical location of those elements on the page. It is therefore a significant abstraction away from the source document and generalises over a set of 'congruent' possible realisations.

A layout hierarchy is represented as a simple nested XML structure made up of 'layout chunks' and 'layout leaves'. Layout chunks can have further layout chunks embedded within them to set up the recursivity of the structures represented. Terminal elements in the structure are represented as layout leaves. Each such unit again receives its own unique identifying label and the entire structure is placed within a single enclosing XML tag called the 'layout root'. The contents of each layout unit, that is, the elements on the page that comprise them, are identified in the way standard for standoff annotation—i.e., the layout leaves contain cross-references to

the identifiers of the corresponding base units. The layout structure corresponding to the example in Figure 9 is then as follows:

```
<layout-root id="lay-01">
      <layout-leaf id="lay-02" xref="u-01"/>
      <layout-chunk id="lay-03">
            <layout-leaf id="lay-04" xref="u-03 u-04 u-05"/>
            <layout-leaf id="lay-05" xref="u-02" />
            <layout-leaf id="lay-06" xref="u-06 u-07" />
      </layout-chunk>
      <layout-leaf id="lay-07" xref="u-08" />
      <layout-leaf id="lay-08" xref="u-09" />
      <layout-chunk id="lay-09">
            <layout-leaf id="lay-10" xref="u-70 u-71 u-90"/>
            <layout-leaf id="lay-11" xref="u-10 ... " />
            <layout-leaf id="lay-12" xref="u-99" />
      </layout-chunk>
</layout-root>
```

The interested reader can following through the structure and the cross-references as identified in Figure 8 above to confirm that the hierarchical view thus created does indeed correspond to the hierarchy given in Figure 9. This should help make it clear why proper computational tools for checking the formal consistency (e.g., are all the identifying labels used actually defined somewhere?) are so important.

The representation of the orthographic and typographic information is then relatively simple. A set of XML-specifications state which *layout units* have which typographical features. In this way, it is straightforward to make generalisations over subhierarchies drawn from the layout structure: for example, all the layout units corresponding to a block of text that is realised uniformly in terms of its typography may be grouped as a single node in the layout structure and it is this node which has the corresponding typographic features associated with it. This allows information to be expressed concisely without repetition.

There are already very extensive vocabularies for describing typographical features: we adopt these for this aspect of the GeM annotation scheme rather than developing a further, ad hoc set of terms. Concretely, we use the typographical distinctions described as part of the XML formatting objects standard. An example of such a specification for the unit corresponding to the headline at the top of the page is then as follows:

```
<text xref="lay-01"
      font-family="sans-serif"
      font-size="18"
      font-style="normal"
      font-weight="bold"
      case="mixed"
      justification="left"
      color="white"
      background-color="grey"/>
```

The final component of the layout annotation layer adds in the information about precise placement within a page. We separate a general statement of the *potential* placement strategy employed on a page from that of the hierarchical layout structure for that page. Placement is then indicated by adding to the layout elements an ‘address’ given in terms of the general positions defined possible for their page. We have found this separation of information to be worthwhile for a number of reasons. First, it is quite possible that minor variations in the precise placement of layout elements can be undertaken for genre-specific reasons without altering the hierarchical relationships present. Second, the separation of placement information makes it possible to state generalisations over the physical placement that are inconveniently expressed at the level of individual layout elements: for example, it is common that pages use various alignments for their material—this alignment can hold over portions of the layout

structure that are not strongly related hierarchically. Good illustrations of the consequences of varying such alignments or non-alignments are given in, for example, Schriver (1997:314) for complex instructional texts.

In order to fully capture these possible dimensions of variation, then, we express within-page placement in terms of an **area model**. Area models divide the space on a page into a set of hierarchically nested **grids**, or tables. Since the grid technique is one that is commonly employed in professional design, it is often straightforward and useful to allow this information to be expressed directly in our annotation; this is particularly the case for newspapers, which are traditionally prepared and designed using pages divided into columns. However, in contrast to the basic column-structuring of newspapers, the function of the area model is more specific in that it provides particular physical reference points for the defined layout elements. Layout elements from the layout structure are then placed in correspondence with particular elements drawn from the page's grid structure. This is necessary because simply stating that some layout unit divides, for example, into three sub-elements still leaves very many options open for those sub-elements physical placement, both within the general space defined by their parent layout unit and with respect to one another.

The grid structure of the area model for our example page extract is shown in Figure 10. Here we can see that the main body of the page is annotated as having a 'row' structure rather than a full grid. Some of these rows are themselves subdivided into further row or column structures; the exact XML definition of the area model used for this page is given in Figure 12 below.

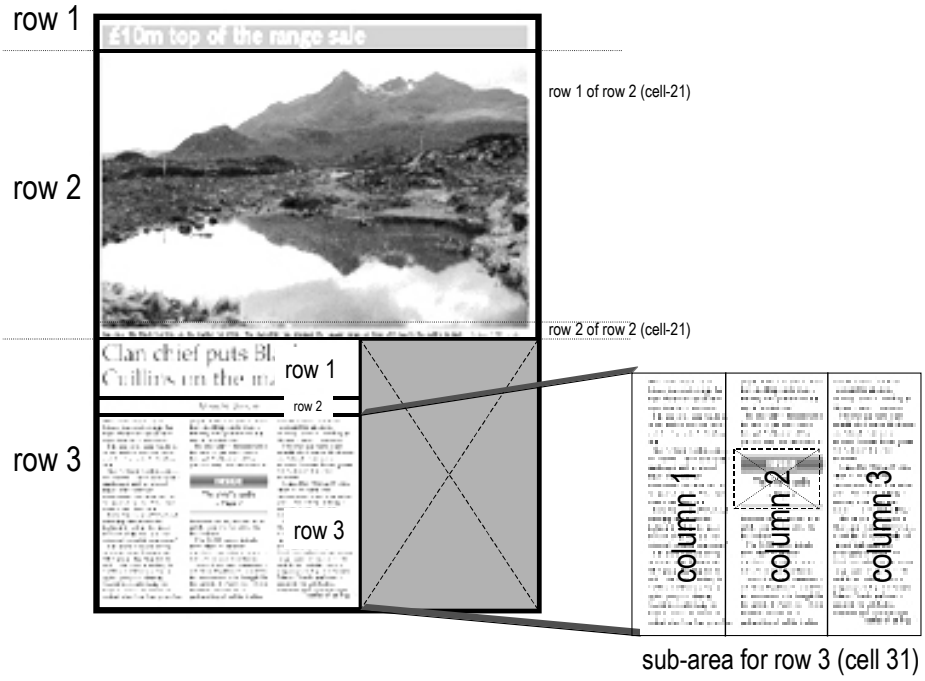

Figure 10: Area model represented as a grid structure for the page

This kind of area model is quite characteristic for newspapers both with respect to the use of an overall column structure, which is picked up as columns of various sub-areas, and with respect to the relatively frequent use of 'insets', which relatively arbitrarily 'cover' parts of the grid structure so that it is no longer available for some particular content. This is commonly the case for advertisements and other rhetorically distinct information such as the navigation elements in the middle of column 2 of the sub-area of row 5.[8]

[8] Note that to describe what is going on in the case of the newspaper page fully, we have an interesting interaction between several other layers of the GeM model. The fact that a newspaper page is organised throughout in terms

Although there are many interesting further issues that arise with this layer of annotation, space precludes their discussion here. Readers are referred to the GeM technical documentation for a more complete account. All of the pages of the documents being added to the GeM corpus are described in the general terms that have been set out here. We will return to some possible uses of this in the next section. First, however, we turn, rather more briefly, to the rhetorical layer of annotation.

As mentioned above, the basis for our rhetorical annotation is Mann and Thompsons's Rhetorical Structure Theory (RST: Mann and Thompson, 1988). RST provides a set of concepts and a notation to express the way in which segments of text are hierarchically and recursively related to one another in the presentation of a coherent text. It seeks to capture the way in which content is presented as an 'argument': that is, how the various elements of a text are interrelated textually so as to bring about the desired communicative purpose. To do this, RST decomposes a text into a hierarchical structure where each level consists of a nucleus, which is the main point of the argument at that point, and at least one satellite, which represents subsidiary information. Particular rhetorical relations then hold between the nucleus and satellite(s) at any level of structure. Current RST posits around 25 relations, which has been found sufficient for a broad range of texts and text types. An important point about RST is that it was conceived for the representation of text, rather than text and image. Within GeM, we treat RST structure as neutral regarding the mode of the 'text spans' it describes: these can be photographs, diagrams, or text. As a consequence, we have found it necessary to augment the relations in RST so as to capture some of the typical image-text relations that are found in texts of the various genres of study.

An RST-style analysis including the material shown in our example page extract is shown in Figure 11. The vertical lines mark the nuclei of the argument at each level of structure, the arcs represent the rhetorical relations, and the horizontal lines indicate the document elements that are being related. We additionally label some of the document elements so that we can refer to them below, although this is not a necessary part of the graphical representation adopted. From the analysis we can see, as is typical of newspaper stories (cf. Bell (1998), Ungerer (2000)), that the main body of the article is made up of a central premise (the news) and a sequence of elaborations of that central premise, generally surfacing as paragraphs in the final article.

The analysis raises a number of interesting methodological and technical issues. For example, whereas we have been describing the annotation process up until now as far as possible as if it were concerned with single pages, there are many document genres where this is not a sensible division. In such genres a more useful unit of analysis is the 'article' or 'story'. In the present case, the story is divided over two pages, the front page material illustrated above, and a continuation of that story on page 2 of the newspaper. Each of these segments of the story receives its own layout structure annotation as we have described it. In the RST analysis, however, we cannot make this division so readily. The 'place' where the page break occurs is indicated in the figure by the line of grey dots running vertically slightly right of the middle. This division is, from the point of view of the RST analysis, completely arbitrary. There is absolutely no rhetorical support for the division at this point.

---

of columns is nowadays one of the *canvas* constraints that hold for the genre: no matter how the individual articles are organised in terms of their own area models, they must be 'poured' into the mould provided by the canvas, which, for newspapers, consists of columns. In earlier times, when print technology was more restrictive, we can even imagine the 'column nature' of newspapers being a *production* constraint—i.e., one imposed by the technology of production and so not variable for different purposes. The GeM constraints form a natural hierarchy; for example, canvas constraints can only be varied within the range of possibilities that the production constraints provide for.

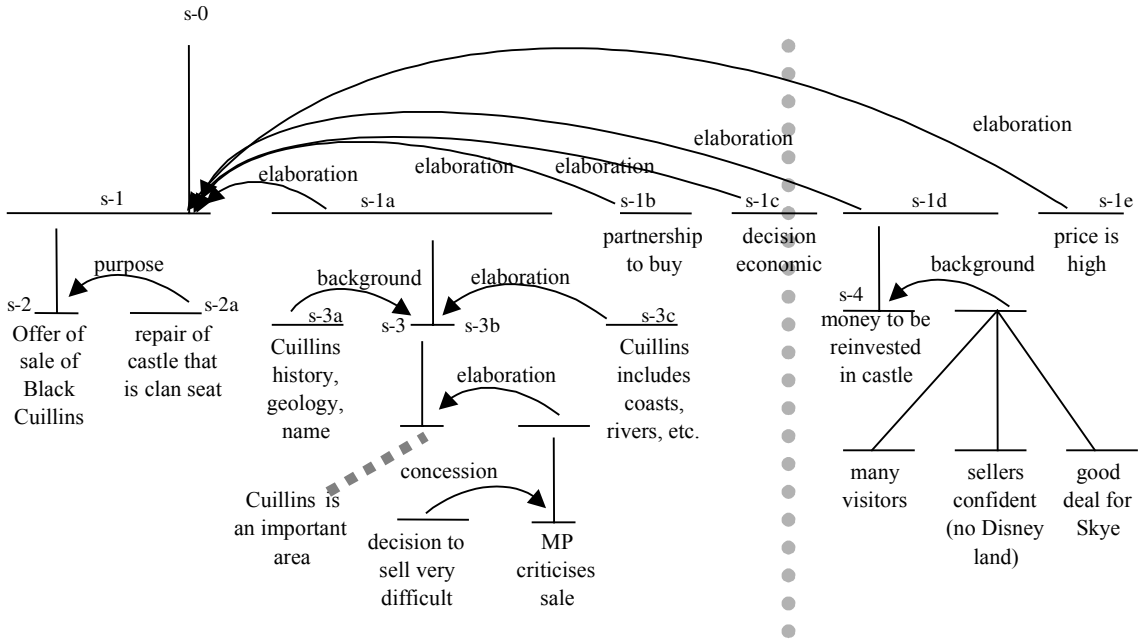

Figure 11: RST analysis of the Cuillins article

The reason for the page break is therefore to found in the other practicalities of newspaper design—for example, that only so much space is left in the newshole on the first page and that the *Herald* (like many other newspapers—i.e., those of a similar genre) wants to get approximately two to three front page stories into that space. The break here then resembles the break of a text across pages in a novel, but is different in that the break occurs not because of the canvas constraint that only so much is to be fit on a page but because of the consumption constraint that readers are to be presented with three articles for consumption. In Bateman, Delin and Allen (2000), we discuss the details of this breakdown of the Cuillins story in more detail; here, we simply note that this is another motivation for the strong separation in layers of annotation that we adopt for the GeM annotation scheme. When annotating the newspaper page, we have a single layout structure that represents that page as a visual unit of analysis. Considered from the perspective of the rhetorical structure, however, that single page may thread together components of rhetorically distinct units—i.e., the articles shown on the page—and these may require more than a single page for their physical presentation.

The RST analysis is captured in the GeM annotation again in the usual manner suggested by standoff annotations. We describe the RST structure as a single XML document, and the units of this layer refer to the base units described above. These base units can then be drawn from different 'documents' or 'pages'.[9] To express an RST-style analysis the main structural elements to consider are the relations, the nuclei and, for each relation, the satellites that are related. The annotation follows from this quite directly. Each nucleus-satellite-relation complex is given a single piece of XML annotation called a **span**. In addition, there are a small number of rhetorical relations that allow several nuclei (and no satellites)—examples of these are simple lists; these are identified by the XML annotation **multi-span.** The top levels of the RST structure of Figure 11 can then be given as follows:

[9] The precise mechanism for this again relies upon the XML standard. It is possible to use a special notation that 'follows' any given XML structure so as to find particular elements of that structure. The XML structure therefore serves to provide an exact 'address' for any element of the data that we are interested in. We use this further below when we turn to examples using the corpus.

```
<span id="s-0" nucleus="s-1"
         satellites="s-1a s-1b s-1c s-1d s-1e")
         relation="elaboration"/>
  <span id="s-1" nucleus="s-2" satellites="s-2a" relation="purpose"/>
  <span id="s-1a" nuclei="s-3" satellites="s-3a" relation="background"/>
  <span id="s-3" nuclei="s-3b" satellites="s-3c"
        relation="elaboration"/>
  <span id="s-1d" nuclei="s-4" relation="background"/>
    ...
```

The cross-references in the nucleus and satellite attributes that correspond to leaves of the RST structure, and which therefore are not identified by 'span' units of their own, are identified in a separate section of the analysis where collections of base units are grouped together into single rhetorical units. This allows the rhetorical analysis to impose structure on configurations of base units whose granularity is defined by the requirements of RST rather than by the requirements of decomposing the page visually. Again, if an annotation scheme were not to do this, then the information about rhetorical analysis would not be maintained modularly within a single annotation layer. In general, we cannot allow the requirements of one layer of annotation to influence the segmentation of other layers.

Providing annotation layers of the kind described in this section for all of the GeM layers is then the main task involved in constructing a multimodal corpus. We use XML so that we can rely on standard tools and techniques for storing the data, checking their integrity, and for presenting various views of the data when considering analysis. This then places multimodal corpus design for the kinds of documents that we are considering on a firm technological foundation. We also use XML, however, to be able to make use of the tools that are now emerging in the structured data representation industry for presenting queries and for searching for regularities in the data captured. And it is to this that we now turn.

5. Examples of empirical research using a GeM-annotated corpus

We give two brief examples of using the GeM-annotated corpus for linguistic research—drawing on our examples in Section 2. Although the corpus needs to be considerably extended in coverage before we can approach the kind of statements now possible in linguistic corpus analysis, we nevertheless believe that the approach outlined represents a sound methodological direction for eventually achieving this goal. Our discussion in this section must therefore be seen as merely suggestive of the possibilities that open up when multimodal corpora are available in the form we propose. Presentation of more solid results at greater length must wait for the future.

We have made much of the fact that we now have a method and framework for adding multimodal pages into a corpus of multimodal documents that is richly annotated and XML-conformant. A prime motivation for this direction is to be able to avail ourselves of another area of the emerging XML industry: that is the area of *searching and manipulating* XML documents. In essence, the only reason to put the effort into the highly structured forms of representation necessary for a representation such as XML is the promise of being able to get out more than one has put in. In the case of linguistic corpora, we are seeking the ability to ask questions of our corpus in sufficiently flexible and powerful ways as to promote theory construction and testing.

The components of the XML standard that are relevant here are those concerned with finding selected elements within a set of XML-structured data. Two large-scale efforts in the World-Wide Web community have concerned themselves with this task: that of the XPath group and that of the XQuery group.

The XPath group has formulated an approach to finding elements within an XML structure by specifying in a very general way 'paths' from the root of the XML structure to the element that is being sought. The path is similar to that used for files or folders on a computer system: as elements in XML may be recursively structured, and each structural element is identified by its tag, this provides a ready addressing mechanism to navigate around XML structures of arbitrary size and complexity. As a simple example, if we wanted to locate within a layout structure the top level layout chunks, then all we need write is an XPath specification such as:

```
/layout-root/layout-chunk
```

and the result, when passed to a standard XPath-processor, would be the set of layout-chunks immediately embedded within the layout-root. A variety of further constructions make the XPath specifications into a powerful way of locating sets of parts of XML documents that conform to given requirements. For example, the following applied to our base units document would give us the contents of all the base units (i.e., the raw text or pictures) directly without any of the mark-up:

```
//unit
```

A similar construction would extract just the text—i.e., would throw away the annotation—from the representation for the SUSANNE corpus suggested in Section 3.2 above:

```
//w
```

The result, when passed to a standard XPath-processor, would be all the elements marked by an 'w' tag, i.e., the words.

More indicative of the power of this mechanism is the following, which would give us all instances from the SUSANNE corpus where a word has been classified as having the part of speech designated "WGv":

```
//w[@pos="WGv"]
```

Further constructs allow us to sort these, again according to various criteria, or to impose further restrictions (e.g., all such words that follow or precede some other class). Similarly, and returning to the GeM annotation, the following would give us all the instances of the rhetorical relation 'elaboration' in our document:

```
//span[@relation="elaboration"]
```

The XPath language is being defined and implemented independently of any linguistic concerns—it is again subject to the primarily economic demands that are also serving to drive the development of XML. The fact that we can immediately use the results of this development for linguistic work is solely because of the XML-conformant nature of our annotation scheme.

The main task of XPath so far has been for writing transformation of XML documents—that is, procedures that take a particular XML structure and convert this into some other form; XPath is in fact an important part of a general transformation language for XML documents called XSLT. Under this perspective, extracting information from a linguistic corpus can be seen as a document transformation: the input document is the corpus or corpora under study, and the output document is the result of the corpus inquiry. One of the most commonly used areas of transformation, however, is the conversion of XML into HTML documents: this allows XML documents to be viewed with normal web browsers. For this to work, one needs to be able to find certain parts of the XML document (using the XPath scheme) and rewrite these as configurations of HTML. Here again, we can find very useful applications of this technique for our purposes: for example, given an XML specification of an area model for a page as described in the previous section, we can 'view' this graphically within a web-browser by converting the GeM annotation into appropriate HTML.

Below, in Figure 12, we show this for our example page fragment. On the left we have the actual area model as it is defined using the precise form specified in the GeM technical manual; on the right we show the corresponding visualisation of this model as seen in a normal modern web browser. This can naturally be used as a convenient method of checking that the formal XML area model specification is as intended. Deviations from the target page are generally very striking. Note that there are then other uses for this kind of conversion—in particular, we have a ready-built mechanism for providing a skeleton of a page for automatic layout generation. Filling in the cross-references to the layout model and the base units can then provide graphical and textual content for the areas distinguished by colour in the example in the figure; work taking this further is described in Henschel, Bateman and Delin (2002).

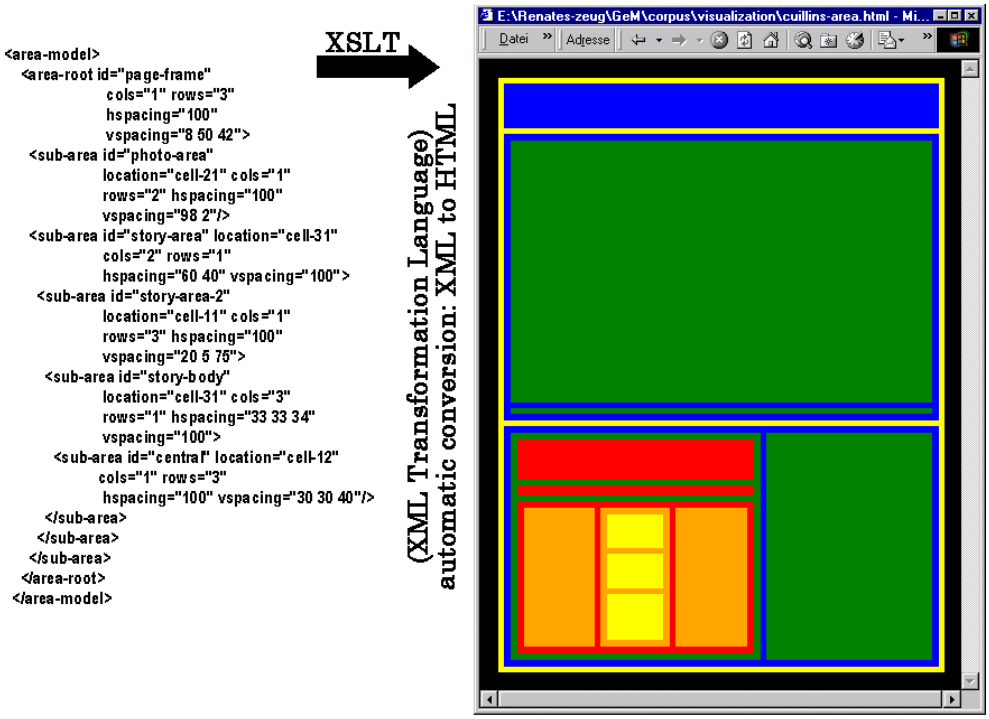

Figure 12: Original page and an automatic visualisation from the XML area model

The XQuery group has a related but differing set of concerns to the XPath group. Whereas the XPath usage is primarily for document transformations, the role of XQuery specification is intended to support the increasing use of XML as a replacement for databanks. Databanks provide more or less sophisticated methods of retrieving information via standard query languages (such as the well-known 'Structured Query Language': SQL). XQuery is to provide the same functionality, but directly in terms of XML paths. Functionality additional to XPath is therefore mostly concerned with collecting sets of results under certain conditions.

The current draft recommendations for the latest versions of XPath and XQuery (2.0 and 1.0 respectively at the time of writing) propose that the two specifications be largely merged—with additional XPath functionality for users concerned with document transformations, and additional XQuery functionality for users concerned with databank-like behaviour. Both of these areas will be of relevance for linguistic corpora: the former for visualisation and display of results, the latter for retrieval and analysis of the data held in corpora.

With this background, we now have all the tools to hand necessary for investigating our GeM-annotated corpora.

First, we consider the question of the distribution of given/new material on the page as analysed by Kress and van Leeuwen. If their framework were to be established as correct, then a news story placed on the left of the page is *by virtue of that placement* inherently 'given' with respect

to, or relative to, a story that is placed on the right of the page. Several experimental setups can be envisaged for investigating this claim. We might ask readers to rate the various stories and their pictures on a newspaper front page on a scale running from 'expected' to 'exceptional' and then see if there is any correlation with page placement. Alternatively, we might select articles that are on the 'left' of the page and those on the 'right' (allowing for area model and canvas perturbations) and have readers judge these with respect to one another. Then we might 're-generate' newspaper front pages (using something like our visualisation tool shown above) with the articles on the left and those on the right swapped to see if readers' judgements are effected.

For all of these tasks, we can profitably employ an appropriately annotated corpus of newspaper front pages. The selection of items on the left and those on the right probably needs to be made with some sensitivity to the generic layout of pages: it might be that we need to filter out the advertisements, or the table of contents, that regularly happens in some newspaper to occupy the leftmost (or rightmost) column. This can be pursued by following through the rhetorical structure annotation of the page, finding the main nuclear elements, following the cross-references back to the involved base-units, and selecting just those that are positioned in the layout structure to the right or to the left of the corresponding area models. This is just the kind of manipulation that the XML components XPath and XQuery are designed to facilitate. We might also need to separate out experimental runs involving pages with very different general layout schemes—for example, those which are predominantly vertically organised and those which show a horizontal organisation; again these kinds of properties can be calculated and made into an explicit selection criterion on the basis of the area model.

Asking the readers to judge the articles for degrees of given/new can also be seen as an annotation task: and this can be supported by existing annotation tools for XML. To run our experiment, we might then define an additional 'experimental' layer of XML markup in which experimental subjects choose a rating for presented parts of a page or of selected articles shown independently of their position on a page. The selection of the articles is itself straightforward in that once we find the set of base units that constitute an article, we simply present these as a running text, or text with pictures, ignoring the other information given in the layout structure of the page. Our experimental layer of annotation then associates these articles with given/new ratings in senses hopefully including the very abstract ones intended by Kress and van Leeuwen. We then run over the resulting annotations, displaying the actual page placements of the articles with specific ratings. If the given/new claim of Kress and van Leeuwen is correct, then we should see clear preferences emerging. There may, however, be additional variables to take into consideration.

We do not yet know what the outcomes to experiments such as these would be, but given a sufficiently broad GeM-annotated corpus the experiments themselves will be far simpler to run since the preparation of experimental materials is considerably facilitated. The fact that we will probably obtain clues for further more refined hypotheses which will in turn require further experiments, with further materials to be prepared, is another strong motivation for automating as much of the materials preparation as we can. And this can only be done with a corpus annotated in a way similar to that argued for here.

Second and finally, we consider the question of the relationship between rhetorical organisation and layout. The basic question here can be phrased in terms of the degree of similarity that can be observed between a hierarchical layout structure and a hierarchical rhetorical structure for the material occupying that layout structure. That is: given a layout structure over a set of base units and a rhetorical structure over the same set of base units, what is the relationship between the two hierarchies involved? As noted above, it is now established that the two hierarchies are rarely identical; moreover, it is usually the case that the layout structure is some 'simplification', or even 'transformation', of the rhetorical structure (cf. Bateman,

Reichenberger, Kamps and Kleinz (2001), Bouayad-Agha, Power and Scott (2000)). But we still do not know exactly what dimensions of simplification are acceptable. Here again, we can start employing multi-layered annotated corpora to probe this question. Given a hierarchical representation of layout (our XML layout structure) and a hierarchical representation of rhetorical structure (our XML rhetorical structure), the question of their similarity or difference is one that can be formally pursued and investigated. We have begun investigations in this direction by locating positions where the structures do not match and simply displaying these as 'interesting' positions in the page to consider. We consider it as methodologically attractive that such an approach at first throws us many more questions than it answers. For example, when there is a large-scale difference between rhetorical analysis and layout, is our rhetorical analysis sound? Do we need to alter or refine the criteria on which the analysis is based? Do readers or users of such documents also encounter problems of interpretation corresponding to the mismatch? Can differences in structure be balanced by other layout or presentation strategies?

Again, we do not know what the answers in this area are going to be. But we are convinced that the ability now to ask the questions with significantly more precision than previously is a decisive step towards achieving a genuine multimodal, linguistically-inspired discourse analysis.

6. Conclusions and Directions for the Future

We have argued that it is essential that multimodal analysis that draws on linguistic methods of analysis adopt a more explicit orientation to corpora of organised data. Only in this way is there a hope of demonstrating that certain, currently more impressionistic styles of analysis in fact hold germs of truth (or otherwise). By presenting a first view of an analytic framework for organising multimodal (page-based) data, we have tried to show how this can be done. The availability of increasingly large-scale and inclusive bodies of such data should enable work on multimodal analysis to shift its *own* genre—we expect that the kinds of discourse adopted in analyses of this kind will be able to draw nearer to empirical linguistic discourse and to go beyond styles of discourse more closely allied with literary or cultural analysis. Such analyses are very valuable for forming hypotheses concerning the rich array of meaning-making to investigate, but are themselves inherently limited in precision and verifiability.

While it may turn out that the kinds of meaning-making involved in multimodal discourse are not amenable to analysis in this way, that the role of the interpretative subject is too great and the constraints on meaning brought by the products analysed too weak, we see it as at least methodologically desirable that we pursue this path before dismissing it. Questions not asked will not be answered. It would have been difficult to predict the role of corpus linguistics accurately before the advent of corpus methods and corpora to support them; we take a similar position now with respect to multimodal theorising.

The main content of this paper has been the introduction of an annotation scheme for organising multimodal data in a way that makes it amenable to corpus construction and investigation. The annotation scheme is, however, deliberately open-ended in terms of the information it covers. We argue that the current layers of the GeM model are minimally necessary to capture the basic semiotic meaning-making potential of multimodal pages. While we can begin to ask some detailed questions using these layers, however, they are also clearly not sufficient for all that one needs to ask—for example, we have deliberately left out the detailed annotation of the *contents* of pictorially realised elements of pages. That is, whereas we indicate that a picture is present, and may, if internal elements of that picture are picked up in the linguistically carried rhetorical organisation, recognise those elements as base units, we do *not* do this systematically or exhaustively. For this we would require another layer of annotation that is specifically responsible for this task. Here, an obvious candidate for such a layer is the detailed analytic scheme proposed by Kress and van Leeuwen (1996).

The use of additional layers as hypotheses for testing as suggested in the previous section is also a natural function of this open-endedness. When such hypotheses show themselves to be sufficiently robust (such as, for example, annotations of layout positions as given/new or ideal/real), then they can be added into the information maintained in a corpus and themselves used as the basis for further investigation. Also, we have said very little about those levels of meaning-making which are more usually of concern to linguists: i.e., the linguistic structure. We believe that the form of annotation presented here articulates well with the kind of linguistic analysis that is capable of representing the rich connections between language forms and their underlying functions, and that the model as a whole then forms the most sophisticated attempt to model all the layers that constitute genre available to date.

Clearly, after the setting out the motivation and methods for this approach to multimodal corpora construction, the main body of work remains to be done. Only when we have such corpora can we start putting the programmes of exploration sketched in the previous section into action. That is a considerable and long-term task; where it will take us in our understanding of the meaning-making potential of multimodal pages is something that only the future will tell.

7. Acknowledgements

The GeM project is funded by the British Economic and Social Research Council, whose support we gratefully acknowledge.

8. References

Baldry, A. (ed.) (2000) *Multimodality and multimediality in the distance learning age,* Campobasso, Italy : Palladino.

Barthes, R. (1977) *Image — Music — Text,* New York: Hill and Wang.

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

Bateman, J.A., Reichenberger, K., Kamps, T. and Kleinz, K. (2001) Constructive text, diagram and layout generation for information presentation: the $DArt_{bio}$ system. *Computational Linguistics* **27,** 409—449.

Bell, A. (1998) The discourse structure of news stories. In: Bell, A. and Garrett, P., (eds.) *Approaches to Media Discourse* , pp. 64—104. Oxford: Blackwell.

Biber, D. (1989) A typology of English texts. *Linguistics* **27,** 3—43.

Biber, D., Conrad, S. and Reppen, R. (1998) *Corpus Linguistics: investigating language structure and use,* Cambridge : Cambridge University Press.

Bouayad-Agha, N., Power, R. and Scott, D. (2000) Can text structure be incompatible with rhetorical structure? In, (.) *Proceedings of the International Natural Language Generation Conference (INLG-2000)* , pp. 194—200. Mitzpe Ramon, Israel.

Bruce, V. and Green, P. (1985) *Visual Perception: physiology,* London : Lawrence Erlbaum Associates.

Bryan, M. (1988) SGML: An author's guide to the Standard Generalized Markup Language. Addison-Wesley Publishing Company.

Cook, G. (2001) *The discourse of advertising,* (2nd. edition) London: Routledge.

Delin, J.L. and Bateman, J.A. (2002) Describing and critiquing multimodal documents. *Document Design* **3**(2).

Delin, J., Bateman, J. and Allen, P. (2002) A model of genre in document layout. *Information Design Journal* **11**(1).

Durusau, P. and O'Donnell, M.B. (submitted) Implementing concurrent markup in XML. *Markup Languages: Theory and Practice* .

Dybkjaer, L., Berman, S., Kipp, M., Olsen, M.W., Pirrelli, V., Reithinger, N. and Soria, C. (2001) *Survey of Existing Tools, Standards and User Needs for Annotation of Natural Interaction and Multimodal Data,* NISLab, Odense University, Denmark; IMS, Stuttgart University, Germany; ILC, Pisa, Italy; DFKI. ('http://isle.nis.sdu.dk/reports/wp11/')

Fries, P.H. (1995) Themes, methods of development, and texts. In: Hasan, R. and Fries, P., (eds.) *On Subject and Theme: a discourse functional perspective* , pp. 317—360. Amsterdam : Benjamins.

Hall, S., Critcher, C., Jefferson, T., Clarke, J. and Roberts, B. (1999) Policing the crisis (excerpt). In: Tumber, H., (ed.) *News: a reader* , pp. 249—256. Oxford: Oxford University Press.

Halliday, M.A.K. (1978) *Language as social semiotic,* London : Edward Arnold.

Halliday, M.A.K. (1985) *An Introduction to Functional Grammar,* London : Edward Arnold.

Henschel, R. (2002) *GeM annotation manual,* Bremen and Stirling: University of Bremen and University of Stirling. (Available at: 'http://www.purl.org/net/gem')

Henschel, R., Bateman, J. and Delin, J. (2002) Automatic genre-driven layout generation. In: *Proceedings of the 6. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2002)* . University of the Saarland, Saarbrücken .

Kress, G.R., Jewitt, C., Ogborn, J. and Tsatsarelis, C. (2000) (eds.) *Multimodal teaching and learning,* London : Continuum.

Kress, G. and van Leeuwen, T. (1996) *Reading Images: the grammar of visual design,* London and New York: Routledge.

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

Kress, G. and van Leeuwen, T. (2001) *Multimodal discourse: the modes and media of contemporary communication,* London : Arnold.

Lie, H.K. (1991) *The Electronic Broadsheet: All the news that fits the display,* Boston . Master's Thesis. School of Architecture and Planning, MIT. (Available at: 'http://www.bilkent.edu.tr/pub/WWW/People/howcome/TEB/www/hwl_th_1.html').

Mann, W.C. and Thompson, S.A. (1988) Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text* **8,** 243—281.

Martin, J.R. (1992) *English text: systems and structure,* Amsterdam : Benjamins.

Martin, J.R. (2002) Fair trade: negotiating meaning in multimodal texts. In: Coppock, P. (ed.) *The Semiotics of Writing: transdisciplinary perspectives on the technology of writing,* pp. 311-338. Brepols and Indiana University Press.

Matthiessen, C.M.I.M. (1993) Register in the round: diversity in a unified theory of register analysis. In: Ghadessy, M., (ed.) *Register analysis: theory and practice* . London : Pinter.

McEnery, T. and Wilson, A. (2001) *Corpus Linguistics,* Edinburgh: Edinburgh University Press.

Nielsen, J. (2000) *Designing web usability: the practice of simplicity,* Indianapolis, Indiana : New Riders Publishing.

O'Halloran, K.L. (1999) Interdependence, Interaction and Metaphor in Multisemiotic Texts. *Social Semiotics* **9,** 317—354 .

Philo, G. (1999) (ed.) *Message received: Glasgow Media Group research 1993-1998,* Harlow : Addison Wesley Longman.

Reichenberger, K., Rondhuis, K.J., Kleinz, J. and Bateman, J.A. (1995) Effective presentation of information through page layout: a linguistically-based approach. In: *Proceedings of ACM Workshop on Effective Abstractions in Multimedia, Layout and Interaction* . San Francisco, California. ('http://www.cs.tufts.edu/~isabel/mmwsproc.html').

Royce, T.D. (1998) Synergy on the page: exploring intersemiotic complementarity in page-based multimodal text. *Japan Association for Systemic Functional Linguistics (JASFL) Occasional Papers* **1,** 25—49 .

Sampson, G. (1995) *English for the computer,* Oxford : Oxford University Press.

Sarkar, S. and Boyer, L. (1993) Perceptual organization in computer vision: a review and a proposal for a classificatory structure. *IEEE Transactions on Systems, Man, and Cybernetics* **23,** 382-399 .

Schriver, K.A. (1997) *Dynamics in document design: creating texts for readers,* New York : John Wiley and Sons.

Sperberg-McQueen, C.M and Burnard, L. (1994) *Guidelines for text encoding and interchange (P3),* Chicago and Oxford : Text Encoding Initiative.

Corpus Encoding Standard (2000) Corpus Encoding Standard. Version 1.5. Available at: 'http://www.cs.vassar.edu/CES'.

Swales, J.M. (1990) *Genre Analysis: English in academic and research settings,* Cambridge : Cambridge University Press.

Thompson, H.S. and McKelvie, D. (1997) Hyperlink semantics for standoff markup of read-only documents. In: *Proceedings of SGML Europe '97* . .

Twyman, M. (1982) The graphic presentation of language. *Information Design Journal* **3,** 2—22 .

Ungerer, F. (2000) News stories and news events—a changing relationship. In: Ungerer, F., (ed.) *English Media Texts past and present: language and textual structure* , pp. 177—196. Amsterdam : Benjamins.

van Leeuwen, T. and Kress, G. (1995) Critical layout analysis. *Internationale Schulbuchforschung* **17,** 25—43 .

Waller, R. (1987) *The typographical contribution to language: towards a model of typographic genres and their underlying structures,* PhD. dissertation, Department of Typography and Graphic Communication, University of Reading, Reading, U.K. .