XML and multimodal corpus design: experiences with multi-layered stand-off annotations in the GeM corpus

John Bateman*, Judy Delin†, Renate Henschel‡

*University of Bremen, Bremen, Germany
bateman@uni-bremen.de

†University of Stirling, Stirling, Scotland
and Enterprise Information Design Unit, Newport Pagnell, Bucks, England
j.l.delin@stir.ac.uk *and* judy.delin@enterpriseidu.com

‡University of Stirling, Stirling, Scotland
rhenschel@uni-bremen.de

Abstract

Current views of multimodal language resources have not yet sufficiently captured the complex interrelationships within page-based information delivery. This is critical for development of multimodal corpora and language resources suitable for large-scale empirical investigation. Serious attempts to interrogate the nature of multimodal meaning-making in professionally-produced documents, both paper and electronic, require a clear understanding of the organisation of the layers into which meaning is organised. In this paper, we present the first multi-layered XML annotation scheme that meets these requirements, developed using a combination of expertise from computational linguists and designers from various sectors of the publishing industry.

1. Introduction

With current developments and goals involving multimodal documents in the widest sense—i.e., including highly interactive artifacts capable of responding to, and producing information in, input/output modes ranging across verbal, gesture, touch and so on, animated/video content, traditional texts, graphics, and so on—it is perhaps tempting to believe that the organization of 'simpler', more traditional document forms, such as two-dimensional presentations involving textual, graphical and diagrammatic information, has been 'solved'. Attention is then drawn away from the complexities of these document types, such as they are, and are to be picked up as a by-product of dealings with more complex artifacts. In our ongoing work on two-dimensional, non-animated information presentations—e.g., books, information leaflets, traditional websites, newspapers (in both print and online forms), and so on—we have found a wealth of complexity that raises serious doubts about such an approach. One aspect of the problem, and the challenge, can be seen in the large gap that exists between previous corpus encoding initiatives (e.g., TEI and the derived CES) which are text based and more recent proposals for capturing mixed media/mode presentations: Somewhere between these two extremes, much of the highly flexible and meaningful resources of two-dimensional information presentation traditionally and non-technically subsumed under 'layout' and graphic design go missing.

As a consequence of this, we have found it necessary to develop a new annotation scheme for describing the informational relationships employed in the area. Two-dimensional information presentation—whether on the page, screen, or whatever—still represents the overwhelming majority of users' contact with information, and so a revealing and empirically based understanding of the meaning-making resources of this area remains of crucial importance. Previous attempts to provide annotation schemes for setting up corpora for documents of this kind have not succeeded in covering very much of the range of phenomena encountered in natural documents however (Corio and Lapalme, 1998; Bouayad-Agha, 1999; Bouayad-Agha, 2000). In this paper, we describe the goals of our own annotation work, set out the basic levels of annotation we believe are required, describe the technical approach taken, and indicate what we see as the next immediate stages, problems and challenges of follow-up development.

2. Goals

We take the view that language, layout, image, and typography are all purposive forms of communication. Accordingly, in our research project GeM ("Genre and Multimodality", http://www.purl.org/net/gem), we aim to describe and analyse all these elements within a common framework, thereby providing a more complete understanding of meaning-making in visual artefacts. By analysing resources across visual and verbal modes, we can see the purpose of each in contributing to the message and structure of the communicative artefact as a whole.

One particular goal of the research is to formalise and model the role of *genre* in layout and typographical decisions. Through the analysis of sample types of multimodal document, the project aims to develop a theory of visual and textual page layout in electronic and paper documents that includes adequate attention to local and expert knowledge in information design. The model is being implemented in the form of a computer program that allows exploration of both existing and potential layout genres, generating alternative and novel layouts for evaluation by design profes-

sionals.

Our use of the term genre here is similar to Biber's (1989, pp5–6), who in his study of linguistic variation states that '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'. We adhere, too, to Biber's view that these categories of text also reflect distinctions in the author's purpose: the documents look different, and contain different language forms, because they are intended to do different things.

Although there are many attempts to categorise the kinds of language that occur in different genres of texts in linguistics, there are few attempts to extend genre analysis into other aspects of visual meaning: Twyman (1982) and Bernhardt (1985), for example, provide preliminary schemes for categorising documents according to the interrelationships between images and text, while Kress and Van Leeuwen (2001) have now also explicitly begun to relate multimodality and genre. Waller (1987), however, is the only attempt extant, to our knowledge, that has attempted to describe the role of language, document content, practical production context and visual appearance in the formation of document genre within the same framework. Our work draws upon and extends Waller's in several ways, as we shall make clear below.

For this, or any project addressing the communicative strategies involved in two-dimensional visual artefacts, the provision of suitable corpus materials is fundamental. Furthermore, since such materials are not currently available, the development of such a corpus has been adopted as an additional explicit goal of the GeM project. The purpose of the corpus development within GeM is to investigate systematic connections between a rich characterisation of the context of use of multimodal documents and their linguistic, graphical, and layout realisations. Within the GeM project itself, four broad document genres have been selected for initial treatment: traditional paper-based newspapers, online web-based newspaper sites, instructional documents, and wildlife books; in each area we have secured a collection of documents and have established contact with designers either expert in these respective fields or, in several cases, actually responsible for the documents gathered. We focus here on the annotation scheme that we have found necessary for structuring the corpus developed.

3. Basic levels of annotation

Waller (1987, pp178ff) represents the constraints on the typographer in producing a graphical document as emerging from three sources:

- Topic structure: 'typographic effects whose purpose is to display information about the author's argument—the purpose of the discourse';
- Artefact structure: 'those features of a typographic display that result from the physical nature of the document or display and its production technology';
- Access structure: 'those features that serve to make the document usable by readers and the status of its components clear'.

Waller did not produce detailed text analyses based on his model but, grounded as it is in the very practical concerns of document design, his view that document appearance results from satisfying goals at different levels is persuasive. We have particularly taken the force of his point that the physical nature of the document and its method of production play a major role in its appearance. In this way, the 'ideal' layout of information on a page may never occur: it must be 'folded in' to the structures afforded by the artefact, and labelled and arranged according to the structures required for access. Document design is therefore never 'free', in the sense that it is never motivated solely by the dictates of the subject matter. We therefore have required a place for these kinds of constraints in our annotation.

In our revision of Waller's model, we suggest that there is an advantage to be gained in uncollapsing his 'topic structure' into a separation between content and rhetorical presentation. We view content to be the 'raw' data out of which documents are constructed. What Waller describes as 'the author's argument' is not solely or completely dictated by content: many rhetorical presentations are compatible with the same content. In terms more familiar from natural language generation, we separate out the 'what-to-say' from rhetorically structured text plans for expressing that content. Secondly, we take what Waller terms 'artefact structure' to be not a structure in the sense of some set of ideas that are to be incorporated in the document, but rather as a constraint on the combination of all the other elements into a finished form.

The levels we propose as minimally necessary for revealing accounts of the operation of the kinds of visual artifacts being gathered in our corpus are, then, as follows:

- Content structure: the structure of the information to be communicated;
- 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, and in terms of the constraints that operate on the information at each level in the generation of a document. Document design, then, arises out of the necessity to satisfy communicative goals at the five levels presented above, while also addressing a number of potentially competing and/or overlapping constraints:

- 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.

Following Waller (1987), then, we claim that not only is it possible to find systematic correspondences between these layers, but also that those correspondences themselves will depend on specifiable aspects of their context of use. In particular, they will depend on 'canvas constraints' set by the nature of the realizational medium (paper, screen-based browser, palmtop, screen resolution) and 'production constraints' imposed by available technology and design choices (allowable cost, number of pages, available printing or rendering techniques, etc.). A model of multimodal genre 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.

Our provision of a corpus of multimodal documents serves as the empirical basis for more thorough investigations of this claim. So far our work has identified widespread mismatches between rhetorical purposes and layout structures even among professionally produced documents; this offers a useful basis for constructive critique. We see the collection of extensive corpora of multimodal documents of this kind, annotated according to the levels of description that we have here briefly motivated, as an essential research and direction for the next five years.

4. Technical implementation

As we have seen, the two communication modes of visual and verbal information presentation are the main perspectives to be captured in the GeM annotation scheme. The scheme accordingly identifies textual elements (verbal mode) and layout elements (visual mode) in a multi-layered annotation, and specifies how these elements are grouped into hierarchical structures (primarily: the rhetorical structure for textual elements, the layout structure formed by the layout elements, and a page model formed by an 'area model': see below). The alignment between these intersecting hierarchies is achieved by specification of the 'GeM base'—a list of the basic units out of which the document is constructed. In accordance with the goal of the

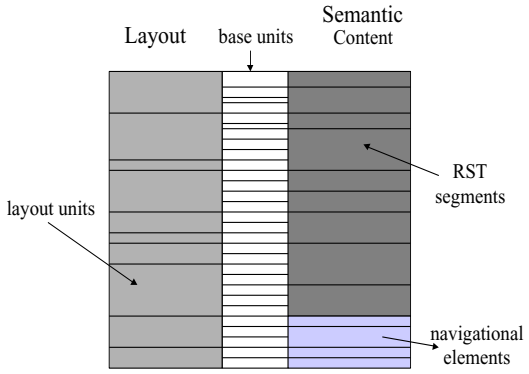

Figure 1: The distribution of base elements to layout, rhetorical and navigational elements

GeM project, the granularity of the linguistic basic units employed in the annotation is approximately the sentence level—this does not preclude providing correspondences with other levels of granularity that might be required for other purposes of course.

Each layer in the GeM model is represented formally as a structured XML specification, whose precise informational content and form is in turn defined by an appropriate Document Type Description (DTD).[1] The markup for one document then consists generally of the following four inter-related layers:

Name	content
GeM base	base units
RST base	rhetorical structure
Layout base	layout properties and structure
Navigation base	navigation elements and structure

All information apart from that of the base level is expressed in terms of pointers to the relevant units of the base level. This stand-off approach to annotation readily supports the necessary range of non-isomorphic, overlapping hierarchical structures commonly found even in the sim-

tion operates at a less delicate level and uses bigger chunks (mostly sentences and graphical page elements) as the bases of the markup. Everything which can be seen on each page of the document has to be included. How the material on each page is broken up into basic units is given by the following list, each is marked as a base unit:, 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, emphasized text, horizontal or vertical lines which function as delimiters between columns or rows, lines, arrows, and polylines which connect other base units.

Everything on a page should belong to one base unit. The base annotation has a flat structure, i.e. it consists of a list of base units.[2] Generally any text portion which is differentiated from its environment by its layout (e.g. typographically, background, border) should be marked as a base unit. The list of base units needs to comprise everything which can be seen on the page/pages of the document. The tag used to mark base units is the <**unit**>. Each base unit has the attribute **id**, which carries an identifying symbol. If the base unit consists of text, the start and end of this text is marked by the <**unit**> tag. Illustrations, however, are not copied into the GeM base. Thus, base units which represent an illustration or another graphical page element are empty XML-elements but can optionally be equipped with an **scr** and/or an **alt** attribute to show, indicate or access the source of an illustration.

4.2. Layout base

The layout base consists of three main parts: (a) layout segmentation – identification of the minimal layout units, (b) realization information – typographical and other layout properties of the basic layout units, and (c) the layout structure information – the grouping of the layout units into more complex layout entities. We explain these three components in more detail below.

In typography, the minimal layout element (in text) is the glyph. In GeM, however, we are primarily concerned with typographical and formatting effects at a more global level for a page; therefore we do not go into such detail, instead considering the paragraph as minimal layout element. That means, a sequence of sentences with the same typographical characteristica which makes up one paragraph is marked as one layout unit. In addition to that we mark all graphically realized elements from the GeM base as layout units. Also highlighted text pieces in sentences, or text pieces within illustrations are marked as layout units. Hence the same list which has been given for the markup of the base units applies here, but with paragraphs instead of orthographic sentences. The tag for a layout unit is <**layout-unit**>. Each layout-unit has the attribute **id**, which carries an identifying symbol, and the attribute **xref** which points to the base units which belong to this layout unit.

The second part of the layout base is the realization. Each layout unit specified in the layout segmentation has a visual realization. The most apparent difference is which mode has been used – the verbal or the visual mode. Following this distinction, the layout base differentiates between two kinds of elements: textual elements and graphical elements marked with the tags <**text**> and <**graphics**> respectively. These two elements have a differing sets of attributes describing their layout properties. The attributes are generally consistent with the layout attributes defined for XSL formatting object and CSS layout models.

Some of the layout units identified in the segmentation part of the layout base can be grouped into larger layout chunks. For instance, the heading and its belonging text form together a larger layout unit, or the cells of a table form the larger layout unit "table". The criterion for grouping layout elements into chunks is that the chunk should consist of elements of the same visual realization (font-family, font-size, ...), or the chunk is differentiated as a whole from its environment *visually* (e.g. by background colour or a surrounding box). In Reichenberger et al. (1995), the authors propose identifying layout chunks by applying a decreasing resolution to the document. The grouping into chunks usually can be applied in several steps, thus forming larger and larger layout chunks out of the basic layout units up to the entire document. Note that one chunk can consist of layout elements of different realizations (text and graphics). The third part of the layout base then serves to represent this hierarchical layout structure. Generally we assume that the layout structure of a document is tree-like with the entire document being the root. Each layout chunk is a node in the tree, and the basic layout units, which have been identified in the segmentation part of the layout base, are the terminal nodes of that tree.

Area model. Each page usually partitions its space into sub-areas. For instance, a page is often designed in three rows – the area for the running head (row-1), the area for the page body (row-2), and the area for the page number (row-3) – which are arranged vertically. The page body space can itself consist of two columns arranged horizontally. These rows/columns need not to be of equal size. For the present, we restrict ourselves to rectangular areas and sub-areas, and allow recursive area subdivision. The partitioning of the space of the entire document is defined in the **area-root**, which structures the document (page) into rectangular subareas in a table-like fashion.[3]

The tag to represent the area root is <**area-root**> The tag to represent the division of a sub-area into smaller rectangles is <**sub-area**>, this shares the attributes of the root but adds a **location** attribute so that subareas are positioned relative to their parent. Locations are indicated with respect to a logical grid defining rows and columns. If, for example, we were considering a page made up of a running

[2] In certain cases, we diverge from the flat structure of the base file. See the technical documentation for further details.

[3] Note that the area-root need not to be a page; if the document to be annotated is a book or brochure, then it can also be the entire book or brochure.

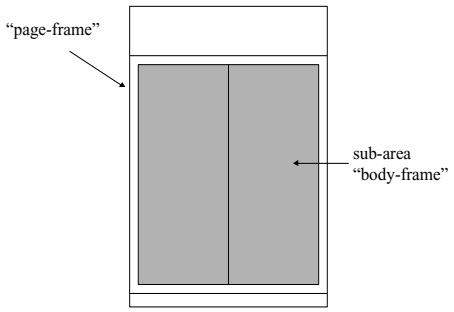

Figure 2: Visualized area model

head, a page body, and a footer for the page number, and in which the page body itself is divided into two columns, then the following annotation would define a corresponding area model. Here, the example's area model consists of a specification of the area-root (called "page-frame"), and the specification of one particular sub-area located in row-2 (called "body-frame"):

```
<area-root id="page-frame" cols="1" rows="3"
      hspacing="100" vspacing="10 85 5"
      height="16cm" width="14cm">
  <sub-area id="body-frame" location="row-2"
      cols="2" rows="1" hspacing="50 50"
      vspacing="100"/>
</area-root>
```

The attribute `vspacing=''10 85 5''` means that the running head takes 10% of the entire page height, the page body 85% and the page number 5%. The page body consisting of two columns is indicated by the hspacing attribute value "50 50", i.e., that both columns are equal in width and take half of the parent unit's width.[4] This area model is visualized in Figure 2.

The area model then provides logical names for the precise positioning of the layout units identified in the layout structure proper.

4.3. RST base

The RST base presents the rhetorical structure of the document. The rhetorical structure is annotated following the Rhetorical Structure Theory (RST) of Mann and Thompson (1988). In RST, a **span** is a continuous text fragment consisting either of a nucleus and one or more satellites (mononuclear relation), or of a number of nuclei which stand in a multinuclear relation (joint, sequence, ...) Some characteristics of RST vary between different research traditions, especially the granularity of the segmentation, the assumed set of rhetorical relations and the branching style of the rhetorical structure tree. We have also needed to make some extensions for the particularities of dealing with mixed verbal and visual information; clearly, when one wants to apply RST to modern, often multimodal, documents, new issues arise. Previous generalizations of RST to multimodal documents have either added new relations to model the relations between graphics and text (Schriver, 1996; Barthes, 1977) or parameterize the existing relation set by a mode parameter (André, 1995). We favour the second approach. However, there are other problems when generalizing RST to multimodal documents, which have not been addressed previously:

- The prominence of graphics in multimodal documents makes it often difficult to decide upon nuclearity in multimodal relations.
- The linear order of the constituents of the document is lost.
- The minimal unit for RST segmentation cannot be restricted to a clause or clause-like phrase.

We address these concerns briefly in turn.

Nuclearity in multimodal relations. Although graphical illustrations are often used to *rephrase* a text passage, it is often difficult to decide which of the two segments – the illustration or the text passage – is in fact nulear and which is the satellite. This seems to be a particular problem of graphics-text relations. To model this problem, we use the multinuclear **restatement** relation. A similar relation can also be found in Barthes under the name **redundant**.

Linear order. Conventional RST builds on the sequentiality of text segments. Relations are only possible (with some minor exceptions) between subsequent segments/spans (sequentiality assumption). With multimodal documents, the mutual spatial relations between the segments changes (from relations in a string-like object to relations in a graph). Segments can have not only a left and a right, but also an upper and a lower neighbour segment. In general one can imagine neighbouring segments in any direction, not only the four which presuppose a rectangular-based page layout. In addition to this, there can be more than one neighbour in each direction. The simpliest solution to apply RST (with its sequentiality assumption) to such a document would be to introduce a reading order on the segments of the document, which is then used as the sequence behind the RST structure. However, this can easily fail to reflect the actual reading behavior. A better, more straightforward generalization of the sequentiality assumption, which we will adopt here, is to restrict RST relations to pairs (sets) of document parts (segments/spans) which are adjacent in any direction. But again, in real documents, one can sometimes find a layout where the rhetorical structure obviously is in conflict with this adjacency condition. Our hypothesis here is that this is generally possible, but that in such a case an explicit navigational element is required so as to indicate the intimate relation between two separated layout units.

Clause as segment. The clause usually serves as minimal unit in RST. There are also approaches, which allow prepositional phrases to be a segment on their own. This is straightforward because both approaches assume something which denotes an action, an event or a state – also called eventualities – as the basic unit. However, if we move to modern documents, particularly multimodal documents, it is questionable whether the clause/PP basis should be kept. Typical examples in multimodal documents are:

[4] For the time being, we ignore space for margins, at least as long as they do not contain footnotes or other text.

- a diagram picturing a certain object and a text label which identifies (puts a name to) this object
- a list with an initiating sentence fragment, as in:

In the box are: ◇ three cordless handsets ◇ the base unit ◇ a mains power lead with adapter ◇ a telephone line cable ◇ two charger pods

- an attribute-value table, as in:

Juvenile	Grey-brown, flecked becoming whiter, adult plumage after three years.
Nest	Mound of seaweed on bare rocky ledge.
Voice	Harsh honks and grating calls at colony.

The cited examples are all expressions of states, or of static relationships between two objects or between an object and a property such as: identification, location, possession, and predication relations. In a traditional linear text, such relations would have been expressed as *is*- and/or *has*-clauses. Each such clause would constitute **one** basic RST segment. In our examples above, however, the two constituents of such a static relation clause are broken out and printed as separate layout units—in the first example, they are even given in differing modes. It is their mutual arrangement on the page plus possible extra graphical devices that expresses the relation between them. This raises the question as to what counts as a minimal unit for an RST analysis in such documents. We solve this issue by introducing a new component for annotation distinct from RST: we analyse the object-object/property relations, if they are clearly separate layout units, according to a small set of relations based on Halliday (1985), which we term 'intraclausal-relations'.

The tag used to mark the basic RST units is <**segment**>. In order to find out which base units form segments, one has to filter out those base units which are in the document for navigational reasons only. These are, for example, page numbers, running heads, footnote labels, document deictic expressions. We also consider headings as navigational elements, and do not include them in the RST analysis. In addition to these segments, we compose other complex segments consisting of more than one base unit for the cases where a intraclausal-relation is expressed on the page by two (or more) separate layout units. Typical examples are diagram + label, table cell$_{i,1}$ + table cell$_{i,2}$ in a two-column table, list initiating sentence fragment + list items. And, finally, sentences disrupted into two base units by page/column breaks only form one segment in the RST base.

The GeM XML annotation for RST aims to overcome some drawbacks found in existing RST annotation approaches. The two standards common in the RST community are those provided by the annotation tools of Daniel Marcu and Mick O'Donnell (see, e.g., www.sil.org/~mannb/rst/toolnote.htm). In both these tools, the annotated output is primarily seen as the program-internal representation of RST structures to be visualized as graphical trees with the help of the tool, but not as output to be used for further XML processing; we describe the pros and cons of the alternatives more in the technical documentation.

4.4. Navigation base

Navigation in a document is performed with the help of pointers, text pieces which tell the reader where the current text, or 'document thread', is continued or which point to an alternative continuation or continuations. The addresses used by such pointers are either names of RST spans or names of layout chunks. For long-distance navigation, typical nodes in the RST structure and in the layout structure have been established for use in pointers; in particular, chapter/section headings are names for RST spans and page numbers are names for page-sized layout-chunks, which tend to be used for navigation. However, there can also be other name-carrying layout-chunks or RST spans such as, for example, figures, tables, enumerated formulas, and so on. The navigation base of a document lists all these "names" which have been defined in this document to be actually or potentially used in pointers. We call the names of RST spans **entries** because they are usually placed immediately before the text of this span. We call the name of a layout-chunk an **index**.

The tag for an entry definition is <**entry**>. We allow entries simultaneously to be segments. We annotate the definition of an index at the page where it is defined, and refer with xref to the base unit which serves as the identifier.

Beside the list of entries and indices, which just defines addresses, the most important part of the navigation base consists of all pointers occuring in the document. The surface realization of pointers are "document deictic expressions", a term coined by Paraboni and van Deemter (2002). Document deictic expressions occur either within sentences or as separate layout units. We have marked the first type as embedded base units and the second as main level base units in the GeM base. In the navigation base, we specify the semantic meaning of such a document deictic expression as **pointer**. We distinguish pointers which operate on the layout structure, and pointers which operate on the RST structure. A pointer (or link) operating on the RST structure points from the current segment (which entails the document deictic expression) to an RST span – the goal RST span – which is layouted at a different place and is not adjacent. A pointer operating on the layout structure points from the layout chunk (which entails the document deictic expression) to another layout chunk which is not adjacent. Another distinction is the pointer type, which indicates different pointing situations. A **continuation** pointer is used in the situation where the layout of an article is broken into two non-adjacent parts. The second part is often printed several pages later than the first part. Continuation pointers are typically layout-operating pointers. **Branching** pointers are used in the situation where a certain piece of information is with respect to its content appropriate at two (or

more) places in the same document. The designer has decided to put it at one of the possible places. In order to indicate the other possible place, a pointer is given at the other location. A third type of pointer is the **expansion pointer**. It is used when more information is available, but not central to the writer's goal. An expansion pointer points to this extra information. Coming along a branching or an expansion pointer, the reader has the choice between two alternatives to continue reading the document. With a continuation pointer there is only the choice between reading continuation or stopping.

4.5. Uses of the corpus

The main results found so far in use of the corpus have been local, in that we are uncovering the rather wide variation that exists between selected layout structures on the one hand and rhetorical organization on the other *within single documents*. In surprisingly many cases, this variation goes beyond what might be considered 'good' design: in fact, we would argue that such designs are flawed and would be improved by a more explicit attention to the rhetorical force communicated by particular layout decisions. This represents the use of the corpus for document critique and improvement (cf. Delin and Bateman (2002)); here further corpus collection is nevertheless essential in order to map further the limits of acceptable functional variation.

We are also exploring the formulation of constraints over collections of corpus entries—e.g., over the pages of a book, or over collections of books in a series, etc.—by means of further annotation levels in which values from the primary annotation levels are partially specified. These need to be hierarchically related. It is at these 'meta' levels that the role of Waller's production and canvas constraints become particularly clear. We are employing this information as an important source of input in a prototype automatic document generation system capable of producing the kinds of variation and layout forms seen in our corpus, thus extending the early generation work in this spirit presented in Bateman et al. (2001).

Finally, we are still searching for more effective means of interogating the corpus maintained in the GeM style. Queries expressed in the XML Xpath language allow simple retrieval of information maintained in the corpus, but are cumbersome for more complex queries. Whether further developments such as XQL or XQuery will bring benefits is not yet clear. Somewhat disappointing was the unsuitability of the previous generation of linguistic-oriented corpus tools, which, despite considerable investment, seem to have been outstripped by the very rapid developments seen in the mainstream XML community. Most of our current work is done directly with XMLSpy and XLST tools such as Xalan. We have found the non-linearity and the non-consecutive nature of the units grouped within our annotation scheme as presenting a major problem for annotation models that have been developed in the speech processing tradition where contiguity of units is the expected case.

5. Follow-up goals, challenges and requirements

We expect that the details of annotation will be refined further as we approach a wider range of documents. It is now a major challenge to produce workable annotation schemes and corresponding corpus collections that include the kind of information we have argued to be necessary in this paper. This information represents a crucial bridging between technicalities of document production and the real issues of design faced in the publishing industry. Corpora built in this way will face two-ways: both to further linguistic and computational plinguistic research and development and to practical issues of design and evaluation. We believe that this needs a firm place in any roadmap now envisaged for language resource construction.

With this in mind we are also exploring a second round of corpus collection and annotation; it is our conviction that only a thorough corpus-oriented study of documents will allow further motivated theoretical and practical statements to be made about the meaning resources that such documents offer. If language resources are to be constructed that include documents of the kind targetted within GeM, then information such as that captured in the GeM annotation scheme will be crucial.

Here there are several issues that require concerted effort. Theoretically, the acceptance of the value and role of rhetorical analyses as giving a fine-grained description of communicative intentions is not uncontroversial. There are attempts in progress to produce corpora of texts annotated rhetorically. We believe this is also essential for multimodal documents. However, as we have detailed above, there are also significant issues that need now to be faced when we move away from linear presentations even to two-dimensional page-based presentations.

More practically, there are issues concerning how much information can be obtained from existing annotation and industry-standard markups: for example, the information maintained in professional document preparations tools such as QuarkXpress or Adobe Framemaker, InDesign, etc. Providing conversion tools to the kinds of linguistically motivated corpus annotations described here would open up a huge area of data. The genre and design knowledge encoded implicitly in style sheets and templates needs also to be made available so that it may be subjected to the kinds of study described above.

Of particular interest to us at present are further extensions across languages so as to compare cultural variation in visual/verbal presentations and further, more detailed comparison of documents variants created by repurposing (e.g., print-to-web, web-to-palmtop, etc.). In both cases, we are concerned that quite ordinary, everyday documents be considered equally, such as bills, consumer letters, instruction manuals, newspapers—these are the documents which users encounter in their everyday lives and understanding how they can be best structured could have significant practical benefits. The acquisition of annotated data across genre and cultures should also therefore be a high priority task.

Finally, we also require that the GeM annotation should

be able to fit into broader annotation schemes. Thus any kind of artifact that includes two-dimensional presentations (for example, a video embedded in a webpage) may also receive a GeM annotation for that component of the information offering. Our claims concerning coherence and consistency of information presentation decisions across text, visuals and layout can then be investigated here also. In such cases, the GeM annotation offers an annotation slice consisting of several annotation levels contributing to more comprehensive annotations that take in other important aspects of the artifact's design beyond that considered within the GeM model. In this respect, we consider it a crucial design feature that such annotation slices be additive and open rather than excluding and closed.

6. References

Elisabeth André. 1995. *Ein planbasierter Ansatz zur Generierung multimedialer Präsentationen*, volume 108. Infix, St. Augustin.

Roland Barthes. 1977. *Image – Music – Text*. Hill and Wang, New York.

John A. Bateman, Thomas Kamps, Jörg Kleinz, and Klaus Reichenberger. 2001. Constructive text, diagram and layout generation for information presentation: the DArt$_{bio}$ system. *Computational Linguistics*, 27(3):409–449.

Stephen Bernhardt. 1985. Text structure and graphic design: the visible design. In James D. Benson and William S. Greaves, editors, *Systemic Perspectives on Discourse, Volume 1*, pages 18–38. Ablex, Norwood, New Jersey.

Douglas Biber. 1989. A typology of english texts. *Linguistics*, 27:3–43.

Nadjet Bouayad-Agha. 1999. Annotating a corpus with layout. In Richard Power and Donia Scott, editors, *Proceedings of the AAAI Fall Symposium on Using Layout for the Generation, Understanding or Retrieval of Documents*, pages 58–61, Cape Cod, Massachusetts, November. American Association for Artificial Intelligence.

Nadjet Bouayad-Agha. 2000. Layout annotation in a corpus of patient information leaflets. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, and G. Stainhaouer, editors, *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'2000*, Athens, Greece. European Language Resources Association (ELRA).

Marc Corio and Guy Lapalme. 1998. Integrated generation of graphics and text: a corpus study. In M. T. Maybury and J. Pustejovsky, editors, *Proceedings of the COLING-ACL Workshop on Content Visualization and Intermedia Representations (CVIR'98)*, pages 63–68, Montréal, August.

Judy L. Delin and John A. Bateman. 2002. Describing and critiquing multimodal documents. *Document Design*, 3(2). Amsterdam: John Benjamins.

Michael A. K. Halliday. 1985. *An Introduction to Functional Grammar*. Edward Arnold, London.

Renate Henschel. 2002. GeM annotation manual. Gem project report, University of Bremen and University of Stirling, Bremen and Stirling. Available at http://purl.org/net/gem.

Gunther Kress and Theo Van Leeuwen. 2001. *Multimodal discourse: the modes and media of contemporary communication*. Arnold, London.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Ivandré Paraboni and Kees van Deemter. 2002. Towards the generation of document deixis reference. In Kees van Deemter and Rodger Kibble, editors, *Information sharing: reference and presupposition in language generation and interpretation*, pages 333–358. CSLI.

Klaus Reichenberger, Klaas Jan Rondhuis, Jörg Kleinz, and John A. Bateman. 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. ACM.

Karen A. Schriver. 1996. *Dynamics in document design: creating texts for readers*. John Wiley and Sons, New York.

Michael Twyman. 1982. The graphic presentation of language. *Information Design Journal*, 3:2–22.

Robert Waller. 1987. *The typographical contribution to language: towards a model of typographic genres and their underlying structures*. Ph.D. thesis, Department of Typography and Graphic Communication, University of Reading, Reading, U.K.