Multimodality and empiricism: preparing for a corpus-based approach to 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. 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. Whereas the application of a linguistic mode of analytic discourse is already showing significant benefits (cf. Kress and van Leeuwen, 2001), the strong coupling between data and theory-construction that forms a tenet of much of modern linguistics is not yet a strong feature of 'multimodal linguistics'.

In this paper, we address this concern. We give an example 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.

2. An example of interpretative analysis

Kress and van Leeuwen (1996) suggest that illustrated documents of a variety of kinds can meaningfully by analysed in terms of several 'signifying systems' that structure the information on the page. Of particular relevance here, is their discussion of *information value* in which they propose that the placement of elements in particular 'zones' in the visual space endows them with particular meanings. Each zone 'accords specific values to the elements placed within it' (Kress and van Leeuwen, 1998:188).

While suggestive, the notion of information value as used by Kress and van Leeuwen is still in need of further clarification. Kress and van Leeuwen use it to describe oppositions between elements placed on the left of a page or image, and those placed on the right. Those on the left are considered to be 'given'; those on the right 'new':

> 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 analysis is appealing in that it 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 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 here is very much more abstract than that generally found in clause (or intonational unit) analyses; for Kress and van Leeuwen the given/new in the page revolves 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 *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 an 'impressionistic interpretative' analysis. 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[1]. Much of the positioning on a newspaper front page is determined in advance, for example, by an a priori decision concerning where the advertising is to be and by considerations such as the need to place suitable material in the top-half of the page 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.

The fact that news editing and the concrete practice of newspaper production do not involve explicit conceptualisations in terms of given/new 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,

[1] 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.

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? 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. To do this, we need to subject analyses to more detailed and systematic investigation, varying types of documents, types of consumers, types of presentation medium, and purposes so that we can get a finer grip on the meaning-making possibilities of the various semiotics in play. And to do this, we need to turn to multimodal corpora specially designed for supporting the investigation of multimodal meaning.

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 draw on the experiences gained with traditional linguistic corpora. It is now 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. Moreover, the model we need for useful multimodal corpora draws particularly on *annotated* corpora—that is, collections of texts that are augmented structurally so as to support investigation of linguistic questions more readily than do simple text collections.

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.

One simple illustration of the problem here involves the phenomenon of variant linguistic forms that do not play a 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. The provision of corpora viewed as collections of texts has largely given way to *annotated corpora*, which contain additional information for the asking of more exact linguistic questions; standard introductions to corpus linguistics describing this development in detail include: Biber *et al* (1998) and McEnery and Wilson (2001).

In modern annotated corpora, it is 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 standardised formats. This separates very clearly *data* from information *about that data*—which makes the information as a whole considerably easier to process and manipulate. The currently most accepted and well developed standardised formats are based on the 'Standard Generalized Markup Language': (SGML: Bryan, 1988) developed in the publishing industry and, most recently, its particular instantiation for wide-scale electronic

information representation XML (the 'eXtensible Markup Language', XML). Standards for corpus annotation adopting these frameworks are also now available (cf. XCES: CES, 2000).

Both SGML and XML recommend the definition of *Document Type Descriptions* (DTDs), which specify precisely the structures that are possible in documents 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 automatically *checked for conformity with their intended structure*. This process is called document **validation**. It is by no means straightforward to guarantee that any sizeable collection of information is structurally correct and consistent; this is the kind of service that a DTD provides. Widely available DTD-parsers check documents for conformity with their specified DTD so that at least formal errors may be avoided.

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 the hypertext markup language 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 the XCES corpus standard as a single paragraph, we use the 'p' tag. The opening 'p' tag is written as and the closing 'p' tag as . To support a richer variety of information in the annotation, tags may also specify **attributes**. Thus, we might, for example, not only want to specify that some particular element in a corpus is a word—perhaps using the XCES tag 'w'—we may also want to give it a unique identification number, specify its part of speech information, and its root form. Providing this information makes the kinds of query problems

mentioned above simple to handle. We can express this information with a complex XML mark-up such as the following:

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

The precise attributes that are allowed and the kinds of values that they may take is specified formally in the Document Type Description, and this allows this information to be formally validated for misspellings, missing brackets, wrong values of attributes, etc. Validation is already a significant reason for providing information in this structured form; we will see below, however, that many more advantages accrue from the adoption of XML.

When attempting more sophisticated linguistic annotation, the most significant problem is that of *intersecting hierarchies*. A good example of this from the area of annotation for literary editions is given by Durusau and O'Donnell (submitted).[2] One simple XCES-conformant markup 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 XML-conformant markup 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. Thus the simplest way of capturing this information, which might appear to be something like the following:

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

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

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 a document's conformance with its DTD .

A solution for this problem that has now established itself is that of *standoff annotation* (Thompson and McKelvie, 1997). Standoff annotation recognises the independence of 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 1, where we have two annotation layers that show how a single text document is divided according to 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 a text where there is a page break immediately following the text: "... Have you, miss? Well,".

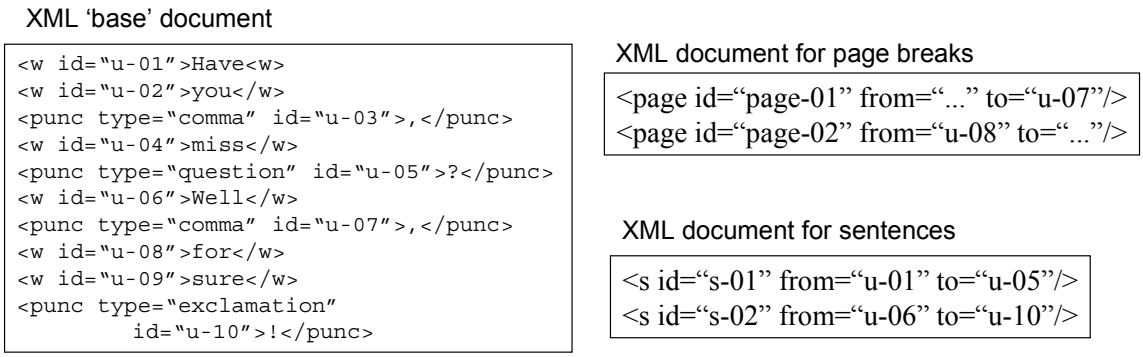

Figure 1: Example of standoff annotation

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. This mechanism 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 richly structured information of this kind. 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 would previously 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 XML-annotated corpora to be navigated and manipulated using widely available and familiar tools rather than complex, corpus-specific schemes and software.

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 linguistic corpora 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).[3] 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. More detailed introductions to the GeM model and its motivation can be found in Delin *et al* (2002) and Delin and Bateman (2002).

4.1 The GeM Model

Our starting point for considering genre draws primarily on linguistic uses, such as, for example, that evident in Biber (1989) or Swales (1990).

We also emphasise and build on the social 'embeddedness' of genres: texts look different because they are to function in different social contexts (cf. Halliday, 1978; Martin, 1992). Moreover, 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 (cf. Kress and van Leeuwen, 2001).

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 was that of Waller (1987). Our own work draws upon and extends this framework by examining the interdependencies between possible characterisations of genre on the one hand and of the various functional constraints on the other. The basic levels of analysis that the project has defined 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.

---

[3] 'Genre and Multimodality: a computer model of genre in document layout'. Funded by the British ESRC, grant no. R000238063. Project website: 'http://www.purl.org/net/gem'.

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; e.g.: paper or screen size; fold geometry such as for a leaflet; number of pages available for a particular topic;

Production constraints: Constraints arising out of the production technology: e.g., limit on page numbers, colours, size of included graphics, availability of photographs; 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, etc.

Particular genres are 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 1. 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 we have just sketched. These are set out in full in the technical documentation for the corpus design (cf. Henschel, 2002). Here we focus on just one layer of annotation, the layout structure, which has been developed within the GeM project. 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. With GeM, 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. The base units we define for GeM include: orthographic sentences, sentence fragments initiating a list, headings, photos, drawings, figures (without caption), captions of photos, text in pictures, icons, table cells, list headers, page numbers, footnotes (without footnote label), footnote labels, and so on. Each such element is marked as a base unit and receives a unique base unit identifier. The base units provide the basic vocabulary of the page—the units out of which all meaning-carrying configurations on the page must be constructed.

Details concerning the form and content of each base unit are not represented at this level. All such information is expressed in terms of pointers to the relevant units of the base level from the other layers of

annotation. 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. Single base units are commonly cross-classified to capture their multifunctionality and can, for example, contribute to a visually realised layout element as well as simultaneously functioning as a component of a rhetorical argument. This ensures that we can maintain the logical independence of the layers considered.

Thus, to take a relatively simple example, if we were annotating the part of a page shown to the left in Figure 2, we would construct a base document along the lines of the XML annotation shown to the right in the figure.[4] Each typographically distinct element on the page is allocated to a different base unit. 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 their functions on the page. Even the fact that the units are approximately

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

ordered following their vertical ordering on the page is not significant—they could in fact be written in any order. This means that any collection of such units can be picked out by the other layers of annotation in order to carry differing functions as necessary.

£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".

INSIDE

The chief's castle – Page 2

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

```
<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 2: Page extract from a newspaper and corresponding base unit annotation

In contrast to the simplicity of the base layer, the other annotation layers are rather more complex. The layout layer of annotation is probably the most complex, however, in that it has several different tasks to perform in 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.

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 our 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. An example of this kind of structuring is shown in Figure 3 below.

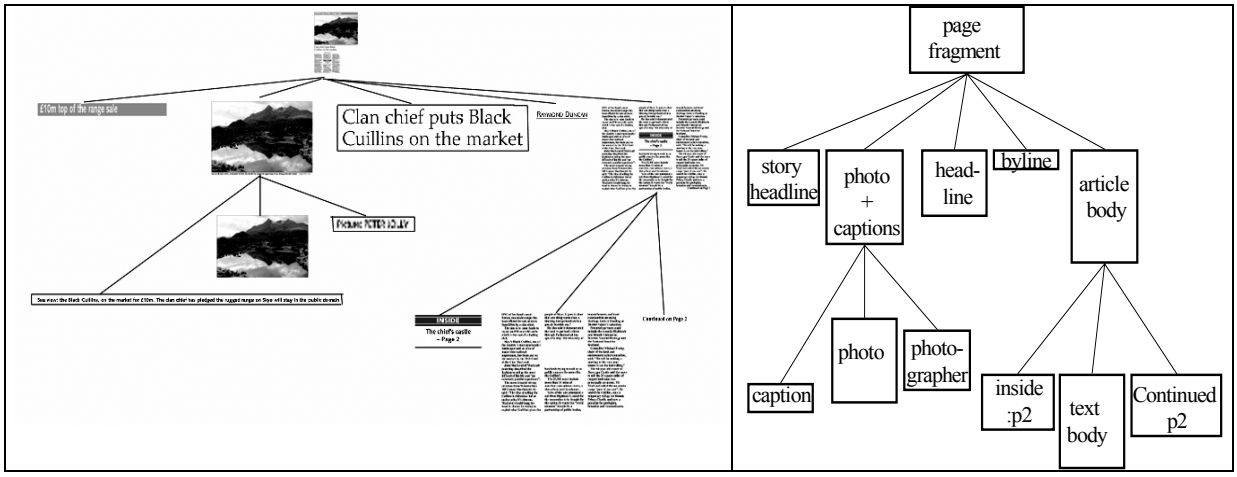

Figure 3: Derivation of hierarchical layout structure from the example page

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 3 is then as shown in Figure 4.

```
<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>
```

Figure 4: Layout structure represented in XML according to the GeM scheme

The interested reader can following through the structure and the cross-references as identified in the base units of Figure 2 to confirm that the hierarchical view thus created does indeed correspond to the hierarchy given in Figure 3. 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-02"                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, 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 generic 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 5. 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.

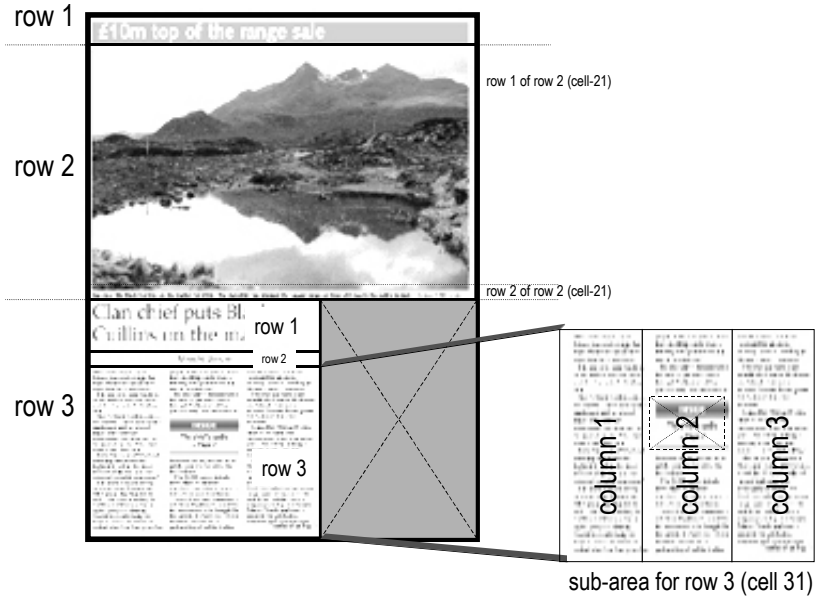

Figure 5: Area model represented as a grid structure for the page

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 3 within top-level row 3.[5]

[5] 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 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

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.

Providing annotation layers as described in this section for all of the GeM layers is then the main task involved in constructing a multimodal corpus of this sort. 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 using a GeM-annotated corpus

Space precludes anything here but a single brief example of using the GeM-annotated corpus for linguistic research—drawing on our example 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.

---

constraints can only be varied within the range of possibilities that the production constraints provide for.

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. One large-scale effort in the World-Wide Web community that is concerned with this task is the 'XPath' group. This 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—which is exactly what is needed for corpus investigation.

For example, the following applied to the representation for a linguistic corpus that we suggested in Section 3.2 above would return the contents of all elements tagged as w-elements without any annotation—i.e, just the words.

```
//w
```

More indicative of the power of the XPath mechanism is the following, which would give us all instances from the corpus where a word has been classified as having the part of speech designated "WGv" (by means of the value of the 'pos' attribute):

```
//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).

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 driving 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.

Returning then to the question of the distribution of given/new material on the page as analysed by Kress and van Leeuwen in Section 2, we can now design a series of empirical and corpus-based studies for its investigation. 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 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 happen 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 exactly the kind of manipulation for which the XML component XPath is being designed. 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 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 proposed here.

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.

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.

We believe the current layers of the GeM model to be the minimum necessary for capturing the basic semiotic meaning-making potential of multimodal pages. They are also, however, 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. This is one reason why the annotation scheme has been defined in a manner which is deliberately open-ended in terms of the information it covers. Further layers of annotation need to be considered. One obvious candidate for such a layer is the detailed analytic scheme proposed by Kress and van Leeuwen (1996). In addition, although 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 explicitly all the layers that constitute genre available to date.

Clearly, after 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 documents is something that only the future will tell.

Acknowledgements

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

References

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.

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

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

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

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

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

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.

Halliday, M.A.K. (1978) Language as social semiotic, 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')

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. ('http://www.bilkent.edu.tr/pub/WWW/People/howcome/TEB/www/hwl_th_1.html').

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.
McEnery, T. and Wilson, A. (2001) Corpus Linguistics, Edinburgh: Edinburgh University Press.
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 .
Schriver, K.A. (1997) Dynamics in document design: creating texts for readers, New York : John Wiley and Sons.
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 . .
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.

Author Details

John Bateman: FB10: Sprach- und Literaturwissenschaften,
Universität Bremen
Bremen D-28334.
bateman@uni-bremen.de

Judy Delin: Department of English and Media Studies
Nottingham Trent University
Clifton Lane
Nottingham
NG11 8NS Nottingham

judy.delin@ntu.ac.uk

and

Enterprise IDU, Newport Pagnell, UK.
judy.delin@enterpriseidu.com

Renate Henschel: Centre for Research in Communication and Language
University of Stirling
Stirling, FK9 4LA, Scotland.

rhenschel@uni-bremen.de

Contact author: Bateman