

A corpus for graphic analysis of texts

Martin Thomas, Judy Delin & Rob Waller April 2011

To cite this paper:

Thomas M, Delin J, & Waller R (2010). A framework for corpus-based analysis of the graphic signalling of discourse structure. Paper presented at Multidisciplinary Approaches to Discourse, Moissac, France, 2010. Technical paper 3, Simplification Centre, University of Reading.

Linguists study how language is actually used by collecting together real examples in computer databases known as corpora (Latin for 'bodies'). They can search each corpus to see, for example, how frequently particular words are used, and which words tend to be used together, or in which kinds of text. This research paper describes a project to extend this approach to cover graphic aspects of text. It is based on the first author's PhD work, extended to cover kind of financial documents that are central to Simplification Centre's research programme.

```
xref="lay-05.15.6.1.1" styles="P93 Default_20_Paragra  
nt-style="" font-family="sans-serif" font-size="11" ba  
ffffff" font-weight="bold" color="#28437e"/>
```

```
xref="lay-05.15.7.1.1" styles="P113 Default_20_Paragr  
t-style="" font-family="sans-serif" font-size="8" back
```

A framework for corpus-based analysis of the graphic signalling of discourse structure

Martin Thomas, Judy Delin, Rob Waller

Simplification Centre, University of Reading, UK

m.thomas@leeds.ac.uk, judy.delin@roedelin.com, r.waller@reading.ac.uk

Abstract This paper describes a corpus-based approach to the analysis of graphic text signalling in complex information documents. To make the task of populating the corpus tractable, we have developed software to automate as much of the annotation process as possible. OCR output is first obtained in OpenDocument format. This is post-processed semi-automatically to generate stand-off XML annotations following the GeM model (Henschel, 2003). These generated layers describe the content and layout of the document. This information is augmented with functionally-oriented descriptions and RST analyses (Mann and Thompson, 1987). Together these annotations support empirical research into the relationship between the things that are said in documents and the linguistic and graphic resources used to express them. Such research might inform the evaluation of existing documents and the design of new ones.

Keywords: corpus linguistics, discourse organisation, document design

1 Background and motivation

This paper describes a corpus-based approach to the analysis of graphic text signalling in complex information documents from the financial services industry. These documents are products of corporate and regulatory processes, and include content from marketing, customer experience, product technical, administration, legal and compliance specialists. Some are produced by automated systems that dynamically merge customer data with other content. They are influenced by external factors such as government regulation, industry codes, and corporate branding. They sometimes have multiple audiences (for example, customers, intermediaries, and administrators). This complex provenance means that these documents are particularly challenging to analyze, but it also means that a sound analysis will provide valuable insight for the organisations concerned, in terms informing both the evaluation and improvement of existing documents and the design of new ones.

To make the task of populating the corpus tractable, we have developed software to automate as much of the annotation process as possible. We do this by post-processing the output of Optical Character Recognition (OCR). As we have yet to populate the corpus, the focus here will be to document our approach, highlighting our aims and motivating the corpus design. We conclude by giving a taste of the kinds of phenomena which might be investigated on the basis of this design. In parallel with designing the corpus, we have been developing a web-based interface

to support querying the data and to present results. Describing this in detail falls beyond the scope of the present paper.

While linguistic accounts of discourse structure differ in important ways, they essentially involve the segmentation of text and the assignment of discourse relations between the resulting segments (see, for example, Grosz and Sidner, 1986; Mann and Thompson, 1987; Martin, 1992). In some instances, surface features may signal these relations. However, as Sporleder and Lascarides (2006, p.371) point out: ‘While rhetorical relations are sometimes signalled lexically by discourse markers [...] such as *but*, *since* or *consequently*, these are often ambiguous, either between two or more discourse relations or between discourse and non-discourse usage.’ They note that ‘*since* can indicate either a temporal or an explanation relation’ and that *yet* may signal a CONTRAST relation or else be used as a synonym of *so far*, ‘with no function with respect to discourse structure whatsoever’ (2006, p.371). Moreover, in many cases, relations are not signalled: ‘roughly half the sentences in the British National Corpus lack a discourse marker entirely’ (2006, p.371). Similarly, in a study of two corpora, one containing task-oriented dialogues, the other a collection of articles from the Wall Street Journal, Taboada (2006) found that, on average, between 60 and 70% of rhetorical relations were not signalled. Interestingly, she also found that some relations are more often signalled than others: in the case of the newspaper corpus, the ‘signalling level’ for different relations ranged from 4 to 90% (Taboada, 2006, p.587).

While the contribution of graphic resources to the structuring of discourse has long been acknowledged from the perspectives of typographers and information designers (e.g. Hartley and Burnhill, 1977; Waller, 1987), linguists engaged in discourse analysis (e.g. Bernhardt, 1985; Lemke, 1998) and those working on natural language generation (e.g. Bouayad-Agha, Scott and Power, 2001), there is a gap between this acknowledgement and the kind of corpus-based research that has yielded insights such as those cited above. This is not to say that the need for such work has gone unnoticed. Bateman, Delin and Henschel (2002a, p.8) point out that, in the case of multimodal meaning making, ‘we lack convenient cut-off criteria such as “grammaticality” that can be interrogated’. As such, in purely descriptive terms, empirical work is arguably even more critical for building the conceptual scaffolding for multimodal analysis than is the case for linguistic analysis. Moreover, as with language, corpus data can help distinguish between those graphic phenomena that are at all frequent and those which constitute anomalies. While, in some contexts, such anomalies may be of interest in themselves, there is a risk that according them undue significance, as may happen when interpreting hand-picked examples, skews the overall picture. In terms of prescription, while the very lack of a grammar might mean we lack criteria for establishing whether a given instance is well-formed and, as such, run the risk of learning from bad examples, it would nonetheless seem useful to be able to identify the different options selected by document designers in realizing a common set of rhetorical structures.

We take graphic signals to include resources such as layout and spacing, the typographic realization of verbal elements, and non-verbal devices such as bullets, ticks and crosses. In addition to these features of the surface realization of document elements, texts in the corpus are annotated in terms of Rhetorical Structure Theory (RST) (Mann and Thompson, 1987) and a number of other functional labels are applied to segments.

The separation of graphic and linguistic realization from functional descriptions allows investigation of phenomena such as graphic complexity and pacing. We can ask whether variation in graphic realization at a given point is motivated, perhaps by signalling structure or a change in voice. Identifying unmotivated complexity can be seen as a step towards finding opportunities to simplify graphic communications. By supporting the retrieval of instances in which rhetorical structure does, or does not, correspond with artefact structure (Waller, 1987, p.179), it also enables us to consider the chunking of information in documents.

Apart from establishing, in broad terms, the ways in which graphic signalling is used, a corpus-based approach allows us to ask more specific questions. For example, we might test the hypothesis that graphic signalling is ambiguous, albeit in ways which differ from lexical signalling. Similarly, we might also investigate whether, as seems likely, the distribution of graphic signals differs, in terms of discourse relations, from the distribution of lexical signals. For example, intuitively, bullets might support LIST or JOINT relations, combinations of ticks and crosses might signal ANTITHESIS or CONTRAST relations, and enumeration might be a natural cue for SEQUENCE. Our corpus currently remains too small to make general claims about the role of graphic signalling of discourse relations. However, as we will suggest in section 4, analysis of documents through the framework of the multi-layered annotation described here seems to offer a means of detecting consistency, or its absence, in the design of a given document.

2 Corpus design

In the corpus, each document will be represented at three levels: 1) metadata about the document; 2) transcribed and annotated document content; and 3) a high-resolution facsimile of each page.

Document metadata will facilitate comparison of sets of documents across various dimensions. These will include the sector and brand from which each document has been taken, the genre to which it belongs (e.g. bill, brochure, form, letter), the topic covered (e.g. insurance, pensions, tax) and the date it was produced. Where relevant, we will identify the relationship between each document and others included in the corpus (in terms of document versions and in relation to steps in business processes). Where available, we will also record data relating to business process errors associated with the document and any existing expert evaluations.

The annotation of specific document elements will be discussed in greater detail in the next section. In brief, it allows us to compare what is said, in rhetorical and functional terms, with how it is said, in terms of graphic and linguistic realization. Genre, for which a characterization is given to a whole document, can also be associated with document parts: thus instances of embedded genres, such as a form within a brochure, may be described and retrieved.

3 Corpus annotation

The corpus annotation is based on the GeM scheme described comprehensively by Henschel (2003). The scheme implements stand-off annotation in XML layers. These include *base*, *layout* and *rhetorical structure* (henceforth *RST*). We have converted the original GeM DTDs into RelaxNG¹ schemas.

This stand-off approach provides an elegant way in which to accommodate the annotation of cross-cutting and overlapping phenomena. It also supports the straightforward addition of new annotation layers, which allows us to record additional descriptions of document segments, such as speaker, intended audience, or communicative intention.

3.1 Capturing data

Leech (1991, p.10) notes that, during the 1980s, OCR freed ‘corpus compilation from the logjam of manual input’. Approaches which seek to represent the graphic as well as linguistic realization of texts inevitably face an even more formidable logjam if manual input is the only means of capturing data. In addition to making the transcription and annotation task tractable, exploiting recent developments in OCR technology to capture information about graphic realization has the additional benefit of excluding judgements about rhetorical relatedness from consideration when describing the visual segmentation of document elements. Such influence, to which human annotators may unavoidably be susceptible, risks introducing circularity into subsequent analysis of information chunking.

Given that we wish to retain as much information as possible about the graphic realization of each document, including the layout and placement of elements as well as specifics such as the size, weight and colour of individual characters, the format in which we choose to save the OCR output should be adequately descriptive. The processes and tools we have implemented to support corpus development will be available for the ongoing expansion of the corpus. The format chosen should therefore be sustainable: it should not be liable to withdrawal or obsolescence. As such, it should be a documented open standard. Finally, the format must afford automated processing downstream. Given that our target corpus annotation scheme is XML-based, using an XML format to store the raw OCR output offers a particularly good fit. The format which currently best meets these three criteria is OpenDocument², which is published as an ISO standard and can be used with a wide variety of office suites, including OpenOffice and Microsoft Office.

Support for OpenDocument in OCR software is increasing. Where it is not available for saving OCR output, this may first be saved as a Microsoft Word document and then converted without loss into OpenDocument using either Microsoft Word or OpenOffice.

¹<http://www.oasis-open.org/committees/relax-ng/spec-20011203.html>

²http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=office

3.2 Post-processing OCR output with XSLT

Using XSLT we can transform OCR output in OpenDocument format automatically into something similar to the *base* and *layout* layers defined in the GeM scheme. Essentially, this involves parsing the source document to assign each element a unique identifier and then reverse engineering the style information carried by each element to conform to the GeM formalism. The other automated intervention is to segment running text at sentence boundaries. This supports RST annotation and the future addition of new layers accommodating other linguistic analyses.

At this point, some rather fundamental differences between the underlying motivations behind OpenDocument and the GeM scheme become evident. OpenDocument is intended to specify document rendering from an authoring perspective, GeM describes the presentation of documents as end products. While it records the layout of elements in relation to a grid-like structure, OpenDocument is not required to record the placement of elements in relation to vertical space. Thus we do not know the extent of each row in the grid. Neither is there an explicit indication of the page on which a given element is placed. Finally, while column widths are retrievable, the placement of individual elements on a given line is not specified: it seems the text is poured into a container and the editing software decides its precise placement. For our purposes, the most significant consequence of this is that, while we can automatically retrieve information about the column and row in which an element is placed, if we wish to indicate page boundaries these must be added manually.

In order to establish the robustness of our approach, we have tested our transformations on a suite of documents containing various graphic features, including complex layouts, tables and lists, for which various routes were taken to arrive at their OpenDocument form.³ Success varies in terms of representation of layout segmentation, document flow and the capture of features such as lists. However, this variation is a function of the OCR process and the representation of OCR output in OpenDocument, rather than of our post-processing. One advantage of using a pivot format, such as OpenDocument, is that manual tweaks to layout recognition during OCR or improvements in OCR technology can feed directly into our processing model without the need for modification to our tools. Our tools coped well with the range of scenarios with which we have presented them, consistently allocating unique identifiers to elements and retrieving available information about their graphic realization.

3.3 Base and layout annotation layers

The *base* layer carries a transcription of the verbal content of the document. In the case of non-verbal graphics, a verbal description of the element is provided. The other purpose of the base layer is to segment the document into *base units*. Each unit is assigned a unique identifier (see Figure 1). Using this identifier, the base units are cross-referenced by the other layers.

Within the *layout* layer, there are four main sections: 1) layout *segmentation* in which each *layout unit* cross-references one or more base units; 2) *realization* information; 3) an abstract

³OCR output in various Microsoft Word formats was converted by Microsoft Word and OpenOffice into OpenDocument format. In one case, we reconstituted a document in a word processor by merging OCR output from scanned images made on different scanners.

```

<unit id="u-5.15.6">
  <unit id="u-5.15.6.1">
    <unit id="u-5.15.6.1.1">Your property</unit>
  </unit>
</unit>
<unit id="u-5.15.7">
  <unit id="u-5.15.7.1">
    <unit id="u-5.15.7.1.1">
      <unit id="u-5.15.7.1.1.01">You could be
sitting on a valuable asset, your home, given the
massive growth in house prices in recent years. </unit>
      <unit id="u-5.15.7.1.1.02">If the children
have flown the nest, you may not need such a big house.
</unit>
      <unit id="u-5.15.7.1.1.03">One option is to
move to something smaller. </unit>
      <unit id="u-5.15.7.1.1.04">As well as being
easier to maintain (and quicker to clean), selling a
big house to buy something smaller means you could
release some of the money tied up in your home.</unit>
    </unit>
  </unit>
</unit>

```

Figure 1: Fragment of base layer annotation

representation of the overall layout of the display, known as the *area model*; and 4) a more concrete description of the *layout structure* of the document, in which layout units are located within the *area model*. Taken together, these four components allow us to build a fairly comprehensive picture of the graphic realization of document elements.

```

<text xref="lay-05.15.6.1.1" styles="P93 Default_20_Paragraph_20_Font
T88" parent-style="" font-family="sans-serif" font-size="11" background-
color="#ffffff" font-weight="bold" color="#28437e"/>
<text xref="lay-05.15.7.1.1" styles="P113 Default_20_Paragraph_20_Font
T9" parent-style="" font-family="sans-serif" font-size="8" background-
color="#ffffff" line-height="0.452cm" color="#000000"/>

```

Figure 2: Fragment of layout layer realization annotation

The fragment shown in Figure 2 describes details of the typographic realization of two verbal document elements. The first is realized in 11 point bold type. The hexadecimal RGB value for *color* describes a shade of blue. The second element is smaller, at 8 points, and black. In both cases, the background colour is white. These two descriptions relate to the heading ‘Your property’ and subsequent paragraph at the bottom of the left-hand column in Figure 3. The realization information retrieved from the data in OpenDocument format differs slightly from that specified by GeM. Some additional information is present (e.g. *line-height*), while other parameters are missing (e.g. *justification* and *case*). Our RelaxNG schema for the layout layer defines additional new attributes and relaxes existing requirements as necessary. We do not believe that this involves a significant deviation from the original annotation model.

The fragment of the *area model* shown in Figure 4 describes the overall layout of the page on which our previous example is realized: in a slight deviation from the GeM specification, we allocate an *area-root* element to each page in the document. The model assumes a grid layout in which each *sub-area* represents a row. So, our example describes a display comprising three rows. The first of these consists of a single column. The second has two columns of approximately equal width. The third also has two columns, but in this the case the first column, on the left, is much wider than the second, which just contains the page number, which is barely visible on the far right.



Figure 3: Facsimile of the display described in Figure 4.

```
<area-root width="20.999cm" height="29.699cm">
  <sub-area id="14" style-name="P54" cols="1" hspacing="100"/>
  <sub-area id="15" style-name="Sect4" cols="2" hspacing="4775 4819"/>
  <sub-area id="17" style-name="Sect6" cols="2" hspacing="8272 1735"/>
</area-root>
```

Figure 4: Fragment of layout layer area model annotation

This rather abstract description is populated with layout units in the *layout-structure*. Putting the fragment shown in Figure 5 together with the *area model* reproduced in Figure 4, we see that our heading ‘Your property’ (*lay-5.15.6*) and the subsequent paragraph (*lay-5.15.7*) are placed last in the first column of the second row in the display (sub-area 15).

In terms of *layout-structure*, document elements may be related recursively: a column in a row in the grid representing the overall page may contain a table which in turn contains text arranged in columns and rows. In addition to recording tabular layouts, elements displayed as a list in the document are also identifiable as such from the *layout-structure*.


```

<layout-chunk area-ref="15">
<layout-leaf xref="lay-5.15.1" location="col-1"/>
<layout-leaf xref="lay-5.15.2" location="col-1"/>
<layout-leaf xref="lay-5.15.3" location="col-1"/>
<layout-leaf xref="lay-5.15.4" location="col-1"/>
<layout-leaf xref="lay-5.15.5" location="col-1"/>
<layout-leaf xref="lay-5.15.6" location="col-1"/>
<layout-leaf xref="lay-5.15.7" location="col-1"/>
<layout-leaf xref="lay-5.15.8" location="col-2"/>
<layout-leaf xref="lay-5.15.9" location="col-2"/>
<!-- ... -->
<layout-leaf xref="lay-5.15.19" location="col-2"/>
</layout-chunk>

```

Figure 5: Fragment of layout structure annotation

3.4 RST and other semantically-oriented layers

Adding functional labels to segments in the *RST* and other semantically-oriented layers must be done manually. This said, by automating the generation of skeleton files, the human annotator is largely relieved of the task of managing cross references across annotation layers.

Despite adopting RST, the GeM project identified several problems with implementing it in multimodal analysis (see, for example, Bateman, Delin and Henschel, 2002b; Bateman, 2008). Perhaps the most fundamental modification to RST as implemented in GeM is the generalization of the RST sequentiality assumption to allow relations to document parts which are adjacent in any direction (Henschel, 2003, p.15).

	✓ A good option for you might be...	✗ A poor option for you might be...
I want to make sure my wife has an income when I die	If you want to help your spouse or dependants after you die, a joint life annuity may be best for you.	A single life annuity will stop when you die, so it will not provide an income for your spouse or dependants.
I want my family to benefit from the money I've saved into my pension when I die	If drawdown is not a suitable option for you, you may want to choose an annuity with 'guaranteed payments' or 'Capital Protection'.	A standard annuity with no options will give you a higher income but it will stop when you die, so it won't provide for your family.
My family have a history of good health, and I'm worried about how my income will be affected over time	An escalating annuity will help lessen the effect of inflation.	The longer you live the more you will be affected by inflation, so a level annuity (where the income stays the same) may not be a good option for you.
I'm used to investing in the stock market and I understand the risks involved	If you want the potential for your income to grow, but drawdown is not a suitable option for you, an investment-linked annuity might appeal. Remember, your income can rise and fall with the market.	A standard annuity that normally pays a set income would not offer the potential for your income to grow.

Figure 6: Table describing good and poor options

One particular challenge presented by the material we have annotated is the presence of tables. Clearly, these support rhetorical relations between elements along more than one spatial dimension. To accommodate this, we allow segments to participate in more than one RST structure. Thus, considering the example in Figure 6, moving from left to right each of the rows of content presents a conditioning situation (CONDITION) followed by a pair of outcomes, which themselves form a CONTRAST relation. In terms of columns, the cells can be seen as constituting a set of situations in a JOINT relation. The column headings relate to the cells below as PREPARATION. This implementation, in which we allow segments to participate in multiple orthogonal relations, retains the concept of nuclearity central to RST.

Other layers are used to describe the *speaker*, *audience*, communicative *intention* and *genre* of document segments. These are formally less complex than the RST layer. They consist of a series of segments each of which has an *xref* attribute which cross references contiguous *base* units who share a given property or, in the case of audience, properties. Thus, the annotation

```

<speaker-unit id="s-13.56.2" xref="u-13.56.2 u-14.57.1 u-14.57.2 u-14.58.1 u-14.60.1
u-14.60.2 u-14.60.3 u-14.60.4 u-14.60.5 u-14.60.6 u-14.60.7 u-14.60.8 u-14.60.9
u-14.60.10 u-14.60.11 u-14.60.12 u-14.60.13 u-14.60.14" type="owner"/>
<speaker-unit id="s-14.60.15" xref="u-14.60.15" type="endorser"/>
<speaker-unit id="s-14.62.1" xref="u-14.62.1 u-14.62.2 u-15.63.1 u-15.63.2 u-15.64.1
u-15.65.1 u-15.66.1 u-15.66.2 u-15.67.1 u-15.68.1 u-15.68.2 u-15.68.3 u-15.68.4
u-15.68.5 u-15.68.6 u-15.68.7 u-15.68.8 u-15.68.9 u-15.68.10 u-15.69.1 u-15.70.1
u-15.71.1" type="owner"/>

```

Figure 7: Fragment of *speaker* layer annotation

fragment in Figure 7 shows that a segment spoken by a third-party *endorser* (*s-14.60.15*) is inserted within a context in which the document *owner* is speaking. In this instance, if we compared the *speaker* with the *layout* layer, we would find that this voice is differentiated typographically in terms of placement, size and colour.

4 Conclusions and further work

The corpus is currently too small to support general claims about graphic signalling of discourse structure and, having put the processes and tools in place, the most pressing future work involves its population. This said, the pilot annotations performed so far have drawn attention to certain features and inconsistencies within individual documents.

All of the examples we have presented here have been taken from a brochure produced by a life insurance company for distribution to people approaching retirement. In many ways it is a successful design. The brochure is aligned with the reader, who it supports through a decision-making process. In terms of information structure, it is well-paced. In most cases, each display presents a topic coherently and topics do not run across pages. In terms of our annotation model, this informal judgement may be supported by comparing the RST and layout layers. The combination of good spacing and short line length maintains legibility, despite much of the type being relatively small.

<p>Step ①</p> <p>Do your research 6 months to 1 year before retiring</p> <p>Decide whether an annuity is right for you by looking at all the available options for funding your retirement and considering the advantages and disadvantages of each.</p>	<p>Step ②</p> <p>Find out the value of your pension fund 10–16 weeks before retiring</p> <p>Your pension provider will send you some information before your retirement date, but you can also contact them yourself ahead of this, to get a rough idea of your pension's value.</p> <p>You may have pensions in different places (eg with companies you no longer work for) so don't forget to dig around and trace them all. The Pension Schemes Registry can help (see 'Where can I get more help?').</p>	<p>Step ③</p> <p>Decide what kind of annuity you want 10–16 weeks before retiring</p> <p>Carefully consider the options available on an annuity, choosing those that best suit your immediate and long-term needs. You may want to speak to an adviser for some help.</p>
<p>Step ④</p> <p>Contact insurance companies for quotes 8–12 weeks before retiring</p> <p>Shop around for the best deal. Scour the market and compare what's on offer carefully. An adviser may be able to do this for you.</p>	<p>Step ⑤</p> <p>Choose your annuity 8–12 weeks before retiring</p> <p>Be careful to choose the right kind of annuity for you, not just the one that pays the highest income straight away. An adviser will give you the reasons for any recommendation he or she makes.</p>	<p>Step ⑥</p> <p>Enjoy your retirement! Once all the hard work is done, it's time to enjoy your free time, safe in the knowledge you'll receive a regular income that will last you for the rest of your life.</p>

Figure 8: Your next steps

However, with regard to the graphic signalling of specific rhetorical relations, a mixed picture emerges. On the one hand, we note that the one clear instance of a SEQUENCE relation is marked by enumeration and that each step is graphically framed (see Figure 8). On the other hand, a combination of ticks and crosses, and bullets of different shapes and colours are used to cue list items at various points in the document. The use of these resources seems to lack consistency.

Pros	Cons
<ul style="list-style-type: none"> ■ It's lower risk than any other retirement option. ■ It will pay you a regular income no matter how long you live. ■ You can choose to provide for your spouse or dependants when you die. ■ You can choose to protect your income against inflation. ■ If you die early, you could choose for your income to be paid for a set period of time, or get some of the remaining money you invested paid back to your estate. 	<ul style="list-style-type: none"> ■ For most annuities, once you've bought it you can't cash it in, swap it for something else or alter your annuity options. ■ The level of your income is not very flexible. ■ If you die early you may get back less than you paid in, although there are options you can choose that can help prevent this. ■ Unless you choose otherwise, your spouse or dependants will not be automatically protected. ■ Unless you choose otherwise, your income will not automatically be protected against inflation. ■ The options you choose affect the level of income you receive. Generally, the more options you add, the more it will cost you, so the lower your income will be.
<ul style="list-style-type: none"> ■ You have the flexibility to vary your income. In fact, you don't have to take anything at all if you don't want to. ■ You control where you invest the money in your pension fund. 	<ul style="list-style-type: none"> ■ You can only take advantage of this option until age 75. At 75 you need to buy an annuity or take an income through an Alternatively Secured Pension. ■ You need quite a lot of money in your pension to take advantage of this option. Typically, you need about £100,000.

Figure 9: Pros and cons

It might be argued that the ticks in Figure 3 carry a positive value, which justifies their use in place of the circular bullets used elsewhere in the brochure. However, the lack of any graphic differentiation between the lists of pros and cons in Figure 9 might be seen as a missed opportunity. Finally, ticks cue both the positive and the negative values in the two lists shown in Figure 10. Here the rhetorical relation CONTRAST would appear not to be supported by the graphic signalling: indeed, one might argue that it is undermined by it. It is unclear why a graphic approach consistent with that already presented in the table reproduced in Figure 6 was not deployed in either of these latter two cases.

Is an annuity right for me?

Yes, it's right for me

- ✓ I want a regular income
- ✓ I want a guaranteed level of income
- ✓ I want an income for the rest of my life

No, it's not right for me

- ✓ I just want to withdraw money when I need it
- ✓ I want flexibility over the level of my income
- ✓ I'd rather have flexibility over security

Figure 10: Is an annuity right for me?

In sum, then, comparing the the layout and RST annotation layers reveals instances in which a common graphic realization serves different rhetorical functions and in which different realizations support a common rhetorical function. If this kind of analysis is revealing when applied to a single document, it seems likely that the application of this approach to a broader set of questions and much larger collections of comparable documents will yield significant new insights.

References

- Aijmer, K. and Altenberg, B. (eds) (1991). *English Corpus Linguistics: Studies in Honour of Jan Svartik*, Longman, Harlow.
- Bateman, J. A. (2008). *Multimodality and Genre: A Foundation for the Systematic Analysis of Multimodal Documents*, Palgrave Macmillan, Houndmills.
- Bateman, J., Delin, J. and Henschel, R. (2002a). Multimodality and empiricism: Methodological issues in the study of multimodal meaning-making. GeM project report.
URL: <http://www.purl.org/net/gem>
- Bateman, J., Delin, J. and Henschel, R. (2002b). XML and multimodal corpus design: experiences with multi-layered stand-off annotations in the GeM corpus, *Proceedings of the LREC'02 Workshop 'Towards a roadmap for multimodal language resources and evaluation'*.
- Benson, J. and Greaves, W. (eds) (1985). *Systemic perspectives on discourse*, Vol. 2, Ablex, Norwood, NJ.
- Bernhardt, S. A. (1985). Text structure and graphic design: the visible design, in Benson and Greaves (1985), pp. 18–38.
- Bouayad-Agha, N., Scott, D. and Power, R. (2001). The influence of layout on the interpretation of referring expressions, in Degand, Bestgen, Spooren and van Waes (2001), pp. 133–141.
- Degand, L., Bestgen, Y., Spooren, W. and van Waes, L. (eds) (2001). *Multidisciplinary Approaches to Discourse*, Stichting Neerlandistiek VU, Amsterdam.
- Grosz, B. J. and Sidner, C. L. (1986). Attention, intentions, and the structure of discourse, *Computational Linguistics* **12**(3): 175–204.
- Hartley, J. and Burnhill, P. (1977). Understanding instructional text: Typography, layout, and design, in M. Howe (ed.), *Adult Learning*, Wiley, London, pp. 223–247.
- Henschel, R. (2003). *GeM Annotation Manual Version 2*, GeM Project.
URL: <http://www.purl.org/net/gem>
- Leech, G. (1991). The state of the art in corpus linguistics, in Aijmer and Altenberg (1991), pp. 8–29.
- Lemke, J. (1998). Multiplying meaning: Visual and verbal semiotics in scientific text, in Martin and Veel (1998), pp. 87–113.
- Mann, W. and Thompson, S. A. (1987). Rhetorical structure theory: A theory of text organization, *Technical report*, Information Sciences Institute, Los Angeles.
- Martin, J. (1992). *English Text: System and structure*, John Benjamins, Philadelphia.
- Martin, J. and Veel, R. (eds) (1998). *Reading Science: Critical and functional perspectives on discourses of science*, Routledge, London.

Sporleder, C. and Lascarides, A. (2006). Using automatically labelled examples to classify rhetorical relations: an assessment, *Natural Language Engineering* **14**(3): 369–416.

Taboada, M. (2006). Discourse markers as signals (or not) of rhetorical relations, *Journal of Pragmatics* **38**: 567–592.

Waller, R. (1987). *The Typographic Contribution to Language*, PhD thesis, University of Reading.