Layout Annotation in a Corpus of Patient Information Leaflets

Nadjet Bouayad-Agha

Information Technology Research Institute
University of Brighton
Lewes Road
Brighton BN2 4GJ, UK
nadjet@itri.bton.ac.uk

Abstract

We discuss the problems and issues that arised during the development of a procedure for annotating layout in a corpus of Patient Information Leaflets. We show how the genre of the corpus as well as the aim of the annotation influenced the annotation scheme. We also describe the automatic annotation procedure.

1. Introduction

We discuss the problems and issues that arised during the development of a procedure for annotating layout. The corpus is the ABPI (ABPI, 1997) Compendium which comprises about 500 Patient Information Leaflets (PILS).

Section 2 presents the aim of this annotation and explains why it plays a major role in deciding on the annotation scheme. Section 3 describes the different possible analyses of the layout of a document while section 4 presents some arguments for choosing between a purely descriptive structure and a purely interpretive structure of layout. Finally, section 5 describes the automatic annotation procedure.

2. Why annotate layout?

The goal of this annotation is to derive rules concerning the interaction between layout and language. These rules should account for orthographic, syntactic, referential and rhetorical behaviour of language in a specific layout context. Examples include the syntactic and orthographic behaviour of items in a list and their relation to their introductory sentence; the role of text segmentation (division in paragraphs, sentences and clauses) in discourse structure; the use of referring expressions accross paragraphs. These relations between layout and language have scarcely been studied in the linguistic community (see however (Bernhardt, 1985; Nunberg, 1990)). However, we need to understand them since our goal is to automatically generate structured documents. [1] Thus, the idea is to annotate the PILS corpus with layout, and then use this annotated corpus for our study matters.

In certain genres such as the *academic paper* genre, layout is highly conventional, often enforced with style sheets and guidelines. In such cases, it is easy to identify layout segments such as headings and paragraphs and then perform a linguistic analysis on them. Furthermore, the language within these layout segments is often highly constrained. For example, textbooks about expository prose teach students that paragraphs are a unit of meaning which should be emphasised by the use cohesive ties within them (Ostorm and Cook, 1988). However, the "patient information leaflet" genre lacks rigid layout conventions, especially across different companies (i.e., the compendium comprises leaflets about 50 pharmaceutical companies).[2] Thus, there are many possible renderings of the same textual segment, not all of which will lead to an optimal document, and therefore, it is hard for the analyst to identify these segments. As DeRose (1995) puts it, "[w]hen dealing with pre-existing information we do not have the luxury of being the author: we can only do our best to discern structure and meaning from what we have".

This difficulty means that layout might not be sufficient for identifying these segments and the analyst might also require other data such as linguistic information. Since our aim of annotating layout segments is to perform a linguistic analysis on those segments, there is a danger of circularity in this procedure which needs to be avoided.

3. How to annotate layout?

Markup schemes such as SGML, XML and DSSSL emphasise the separation between the *form* and *content* of documents (Goldfarb, 1998; Clark, 1996). An author while composing his/her text only specifies the logical structure via SGML/XML tags which are then converted to the desired visual display by a DSSSL processor. Similarly, guidelines such as the Text Encoding Initiative (TEI-P3, 1997) provide analysts with a means of encoding the logical structure of a document as SGML elements with a renditional attribute describing the discriminating physical properties of each visual element (see (WWP, 1999) for an expansion of this renditional attribute). All these frameworks assume that there are at least three levels of layout in a document:

Physical structure. A document is an artifact consisting of volumes, pages, front and back cover, column, lines, dimensions, orientation, margin, typefaces with different values for each of their features (size, weight, slant, case, etc), etc.

Visual structure. Principles of visual organisation which are described in theories such as Gestalt (Köhler,

[1] This work was carried out within the ICONOCLAST project, http://www.itri.brighton.ac.uk/research.html#ICONOCLAST.

[2] The only common denominator of all PILS is a list of particulars which are required by law to be included in every leaflet, such as ingredients and side-effects of the medicine.

1947) explain how the *recipient* of the message groups or separate information according to principles of proximity and similarity via size, symmetry or intensity of information (Campbell, 1995). This is done via the *meaningful* features of the physical structure which "mark, organize or modify" the text (Gilreath, 1993).

Logical structure. This level corresponds to what of the argument (content and discourse) is made *visually observable* by the author (Summers, 1995). From the recipient's point of view, it corresponds to attributing a functional role to the visual elements of the document.

4. Choosing the right level of annotation

4.1. paragraphs versus blocks

As mentioned in section 2, there is a many-to-many mapping between the physical attributes and the logical elements in the PILS corpus, which means that the identification of the functional role of some visual elements is ambiguous unless one relies on linguistic properties. For example, in the corpus, the blank line is used to set off a warning, a list or a paragraph. This problem is prevalent since the corpus contains 2050 lists, 1210 of which are followed by a block of text.

The example below illustrates this problem. The text consists of six visual blocks B1 to B6, two of which (B2 and B5) are lists. One can group together, using syntax, B1 and B2, and also B4 and B5. Now B3 topically belongs to the same group as B1-2. On the other hand B6 if grouped with B4-5 makes the text appears in two visually balanced groups B1-2-3 and B4-5-6. However, when interpreting B6, the referring expression *this list of possible events* may be ambiguous between referring to the last physical list B5 or the conceptual list composed of B2 and B5.

> As with all medicines undesirable events are sometimes experienced. With 'Sorbichew' these may include:
>
> headache
> flushing of the face
> dizziness
> weakness
>
> These may occur at the start of treatment but tend to become less as treatment continues.
>
> Other effects which may occur less frequently include:
>
> nausea and vomiting
> dizziness on standing up
> rash
>
> Do not be alarmed by this list of possible events. You may not have any of them.
>
> *(Sorbichew, Zeneca)*

For our purposes, one can see that there is a risk of circularity because we want to annotate layout for subsequent analysis of its relation with language but are using language together with layout in order to determine the logical structure of the document.

In addition, this analysis means that we assume some defining linguistic properties of logical elements. For example, we can decide that the block of text following a list is part of the same paragraph if it uses pronouns and demonstratives referring back to the list. However, this might hinder the understanding of the purpose of inserting a blank line between two blocks of text which have an anaphoric reference between them. The following example should illustrate this point.

> AFTER USING YOUR PATCHES
>
> These patches sometimes cause unwanted effects in some people:
>
> - Headaches, nausea or breast tenderness
> - Cramping pains in the calf.
> - Feeling slightly bloated.
> - Slight redness and itching of the skin where a patch has been [..]
>
> These effects are often mild and may wear off after a few weeks' treatment.
>
> If they are very troublesome and do not improve tell your doctor. *(Estraderm TTS, CIBA)*

The last block is topically related to the previous one and this is emphasised by the use of a pronoun which most likely refers to an entity mentioned in the previous block. However, the communicative function of this paragraph break is made clear if one puts the last two blocks together. Indeed, a concessive seems required at the beginning of the second block as illustrated below.

> These effects are often mild and may wear off after a few weeks' treatment. *However,* If they are very troublesome and do not improve tell your doctor.

This example clearly illustrates the rhetorical function of layout. It shows, together with the previous point, that paragraphs should only be analysed as visual blocks.

4.2. headings versus labels

Logical elements in the PILS corpus are on a continuum between more layout- to more content- oriented. Thus, the continuum of visual informativeness established by Bernhardt (Bernhardt, 1985) across different genres occurs within the same genre. For our purposes, it means that we sometimes cannot decide whether a visual element belongs to a certain logical class, because it does not have the prototypical features of that class.

The example below should illustrate the point. It shows the same type of information presented in four different ways on a continuum from heading-like to less heading-like segments. The first example illustrates a prototypical heading, distinguished typographically (capitals), spatially (centered in a separate line) and syntactically (a noun phrase). In the middle examples, the label is typographically emphasised and orthographically separated from the body (with a colon) and could be annotated as emphasised text. The last case is not even typographically distinguished. Nevertheless, all these cases could still be annotated as headings because they point, more or less visibly, to the kind of information the reader will be looking for in a patient information leaflet. On the other hand, all these labels do not reflect the structural hierarchy of the document, but rather, simply provide the reader with *reading access points* (Waller, 1988).

PRODUCT LICENCE HOLDER

Elixir Limited, Manchester, UK.

Product licence holder: Elixir Limited, Manchester, UK.

Product licence held by: Elixir Limited, Manchester, UK.

Product licence held by: Elixir Limited, Manchester, UK.

One solution to this naming problem would be to annotate only the physical attributes which make one string of text stand out from its surroundings. In SGML, this translates as an element, called for instance *visual-segment*, with its physical attributes.

5. The annotation process

The automatic derivation of SGML structure from the PILS documents abstracts itself from language. Thus, only prototypical headings are marked as such and only visual blocks (rather than paragraphs) are annotated. This also means that some desirable groupings are not made, for example, between a list and its introductory sentence.

Figure 1 illustrates the process we went through for converting the PILS Compendium into an electronic corpus marked-up with layout.

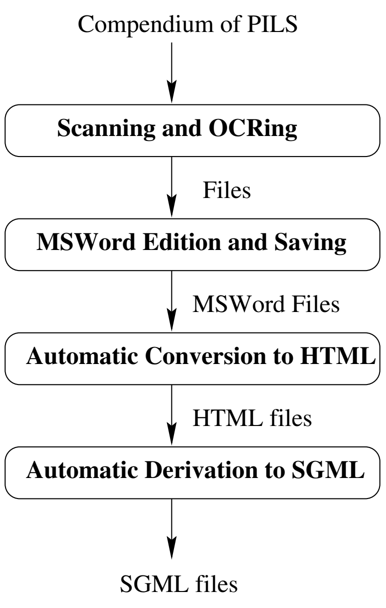

Figure 1: Conversion to Logical Structure Representation

This process consists of the following stages:

Scanning and OCRing. Each page of the compendium was scanned and OCRed.[3]. At this stage, some decisions had to be made about what is text and what is picture because pictures are not (obviously) OCRed. This is not a trivial decisions because some figures are intermingled with text that can be part of the main text. Furthermore, one needs to decide whether to analyse a figure into multiple ones or not.

Editing and saving to Microsoft Word. Since the OCR performs more or less well depending on the quality of the original and the properties of the typeface, the pages were then edited in Microsoft Word (Macintosh version 6.0.1). There were two constraints to respect in doing so. Firstly, the electronic version had to look like the original (multi-column texts, complex grid, pictures surrounded by text, etc). Secondly, the document had to be *explicitly* formatted in Word so that the Word to HTML conversion facility provided by Microsoft performs as good as possible. Thus, lists needed to be formatted as Word lists (and not as a series of disconnected paragraphs starting with a bullet for instance) and multi-column texts needed to be formatted using a table.

Conversion to HTML. A Word macro was written that merges the leaflet pages together and saves the result in HTML format. This operation converts only *standard* layout features such as bulleted lists which have been preselected in Word; it does not convert items which are marked with an asterisk or simply indented and have thus not been preselected in Word style. In addition, page and column breaks are represented as rows and columns in a table, breaking lists that run over them; headings are not distinguished from emphasised text; headings marked with a bullet are tagged as (one-item) lists; the structural hierarchy between sections is lost, etc.

Derivation to SGML. A program was written that derives the logical structure from the presentational markup. To do so, a number of rules and heuristics are used. For example, HTML lists or paragraphs broken into two segments because of page and column breaks are merged back together, series of paragraphs beginning with an asterisk are marked as a list, typographically salient paragraphs are marked as emphasised text, a bulleted single list-item which is typographically salient and does not end with a full stop or a colon is marked as a heading, etc. The division of the document into sections is done in the last phase, by comparing the current heading with the list of heading types that came before it.

The resulting SGML files conform to a Document Type Definition (DTD, see (Goldfarb, 1998)). The DTD was inspired mainly from the TEI Guidelines (TEI-P3, 1997). However, a set of layout attributes had to be developed since the current version of TEI does not provide for them.

Predictably, the program performs very well on documents with a distinguishable hierarchy and a simple grid and very badly on documents with a typographically non-distinguishable hierarchy and a complex grid. We manually evaluated ten randomly selected leaflets from the PILS corpus and found that the

[3] OCR stands for Optical Character Recognition.

program has a precision for nesting (determining the hierarchy) of about 50% and for sectioning (identifying the sections) of about 75%.

6. Conclusion

We have discussed the problems associated with the identification of the logical structure in patient information leaflets. We found available annotation schemes (such as the TEI's) describing layout-dependent structure in texts inadequate for our type of documents. This is because these schemes are mainly concerned with hierarchical models of documents which are typical of many prose texts such as of the expository genre. This illustrates a gap in the understanding of the role of layout in *visually-informative* documents such as PILS (Bernhardt, 1985).

Furthermore, we found that layout and language are intermingled in this genre of texts and opted for a less interpretive structure than logical structure for the annotation of paragraphs, that is, we only annotated them visually. Our SGML derivation process allowed us to do so.

7. References

ABPI (ed.), 1997. *Compendium of Patient Information Leaflets*. Association of British Pharmaceutical Industry.

Bernhardt, S.A., 1985. Text structure and graphic design: the visible design. In J.D.Benson & W. Greaves (ed.), *Systemic Perspectives on Discourse*, volume 2. NJ: Ablex, pages 18–388.

Campbell, K.S., 1995. *Coherence, continuity and cohesion — theoretical foundations* for *document design*. Lawrence Erlbaum Associates.

Clark, J., 1996. Document style semantics and specification language. http://www.jclark.com/dsssl/.

DeRose, S.J., 1995. Structured information — navigation, access, and control. In *Berkeley Finding Aid Conference*. Available at http://sunsite.berkeley.edu/FindingAids/EAD/derose.html.

Gilreath, C.T., 1993. Graphic Cuieing of Text: The Typographic and Diagraphic Dimensions. *Visible Language*, 3(27):336–361.

Goldfarb, C.F., 1998. *The SGML Handbook*. Oxford University Press.

Köhler, W., 1947. *Gestalt Psychology*. Mentor Book, New American Library.

Nunberg, G., 1990. *The Linguistics of Punctuation*. Number 18 in CSLI Lecture Notes. Stanford, CA: CSLI Publications.

Ostorm, J. and W. Cook, 1988. *Better Paragraph Plus*. London: HarperCollins, 6th edition.

Summers, K., 1995. Near-wordless document structure classification. In *Proceedings of the International Conference in Document Analysis and Retrieval (ICDAR–95)*. Montreal.

TEI-P3, 1997. *Guidelines for Electronic Text Encoding and Interchange*, volume I and II. The Association for Computers and Humanities (ACH), The Association for Computational Linguistics (ACL), The Association for Literary and Linguistic Computing (ALLC). C.M. Sperberg-McQueen and Lou Burnard (eds). Available from the Text Encoding Initiative Home Page at http://www.uic.edu:80/orgs/tei/.

Waller, Robert, 1988. *The Typographic Contribution to Language: Towards a Model of Typographic Genres and Their Underlying Structures*. Ph.D. thesis, Department of Typography & Graphic Communication, University of Reading.

WWP, 1999. The women writers project. http://www.stg.brown.edu/projects/wwp/. Brown University.