Integrating Content and Style in Documents: A Case Study of Patient Information Leaflets

Article *in* Information Design Journal · November 2001

DOI: 10.1075/idj.9.2-3.05bou · Source: CiteSeer

3 authors, including:

Nadjet Bouayad-Agha
University Pompeu Fabra
38 PUBLICATIONS **510** CITATIONS

SEE PROFILE

Donia Scott
University of Sussex
129 PUBLICATIONS **2,006** CITATIONS

SEE PROFILE

University of Brighton

ITRI-00-16

Integrating content and style in documents: a case study of patient information leaflets

Nadjet Bouayad-Agha, Donia Scott and Richard Power

January, 2000

Supported by EPSRC Grant L77102

Information Technology Research Institute Technical Report Series

ITRI, Univ. of Brighton, Lewes Road, Brighton BN2 4GJ, UK
TEL: +44 1273 642900 EMAIL: firstname.lastname@itri.brighton.ac.uk
FAX: +44 1273 642908 NET: http://www.itri.brighton.ac.uk

Also published in the *Information Design Journal* 9/2 & 3 (2000) pp. 161–176

Integrating Content and Style in Documents:
A Case Study of Patient Information Leaflets*

Nadjet Bouayad-Agha, Donia Scott, Richard Power
Information Technology Research Institute
University of Brighton
Lewes Road
Brighton BN2 4GJ, UK
email: {first-name.last-name}@itri.bton.ac.uk

Abstract

We envisage a novel computer tool for producing technical documentation, in which the author specifies the desired content and style, but the exact wording and layout i s determined by the system (including versions in languages the author need not know); a prototype of such a system is being developed in the ICONOCLAST project. Amongs t other things, the system must adapt the wording of the generated document to its punctuation and layout. By studying a corpus of patient information leaflets, we have found many detailed examples of this interaction, some of which are described here. In particular, we focus on ways in which the use of special layout patterns (e.g., vertical lists, boxes) changes the options for wording, sometimes licensing departures from normal conventions of grammar and punctuation.

1 Introduction

Perhaps the most powerful idea in recent document processing is that of specifying the style of a document separately from its content. This idea is implemented to some extent in modern word processors, which allow you to vary the appearance of a document systematically by loading a new style sheet; it is implemented more fully in systems based on the Document Style Semantics and Specification Language (DSSSL), which allow detailed control over the presentation of a text that has been marked up in SGML (Standard Generalized Markup Language).

Why is this idea so powerful? Suppose that a pharmaceutical company has designed patient information leaflets for all its products (these are the slips of paper inside medicine packs which explain how to take the medicine and warn you of possible side-effects). Drafts of the leaflets are printed, and passed on to the market research department for testing. One of the things that could result from a test is that patients have difficulty navigating the leaflets because paragraphs are too close together: instead of placing a new paragraph on the next line, starting with a tab, it would be better to leave a double line with no tab. Moreover, it could be found that patients dislike the indented square bullets in unordered lists, finding them ugly and that they demand round bullets aligned with the text. Implementing these changes is obviously far easier if the style for realizing paragraphs or unordered lists has been specified just once, in a general rule. With no separation between content and style, the author will have to go through every paragraph or list in every leaflet, correcting each case individually.

In existing word processors, control over style is limited by the categories available for describing the content of the text. You can mark a span of text as a paragraph, or as a major heading, but you cannot mark it as a list of warnings, or ingredients, or side-effects. There is therefore no opportunity for defining a domain-specific style rule such as *Express warnings by an enumerated list in descending order of severity and within a centrally placed box with a black border*. This limitation will be overcome when systems based on DSSSL come into common use. Through a 'Document

*This work is supported by the Engineering and Physical Sciences Research Council (EPSRC) Grant L77102.

Type Definition' (DTD) the author can create a set of categories for marking up the content of the text, including domain-specific categories like 'warning' or 'ingredient'. Layout rules relating to these categories can then be formulated in the DSSSL language. To obtain an output document, the marked-up text, along with the DTD and the style rules, will be processed by a DSSSL engine such as JADE (Clark, no date), the result being a document in a standard format such as RTF (Rich Text Format) which can be read into a word processor for final editing (if needed) and printing.

Although DSSSL allows more control than existing word processors, it still represents only a small step towards the objective of separating content from style. Returning to the scenario envisaged earlier, suppose that during market research, patients complain about the style in which the leaflets are written. For instance, they might object that the sentences are too long, or that technical words like 'larynx' are used instead of familiar ones like 'throat', or that warnings are listed in running prose instead of a series of bullet points. Since these criticisms have nothing to do with the information expressed by the leaflet, it seems appropriate to regard them as matters of style rather than content; nevertheless, they cannot be addressed by style rules in DSSSL, since they require rewording of the text.

In this article we introduce the ICONOCLAST project, now in progress at the Information Technology Research Institute, University of Brighton, which aims to achieve a full separation between content and style. This is done through the technology of Natural Language Generation (NLG), in which a computer program generates documents, perhaps in several languages, starting from a completely precise description of the desired information content. In the next section we will explain how this technology works, and how it might eventually be applied in document design.

2 Natural Language Generation

The central feature of NLG is the use of a *formal language*, rather than a natural language such as English or French, to represent the information content of a document. The prime example of a formal language is mathematics; another example is a computer programming language such as Pascal. Of course traditional mathematics (arithmetic and algebra) has limited coverage: it is useful only for expressing equalities and other relationships among quantities. The theoretical possibility of generalising the coverage of mathematical notation was suggested by the philosopher Leibniz (Russell, 1945), who envisaged that in the future, once this notation was fully developed, philosophical disputes might be resolved by routine calculation. His insight led to the development of modern formal logic during the 19th and early 20th centuries, especially by Frege and Russell (Russell, 1945). Frege invented a mathematical notation known as *predicate calculus*, which can be extended so that it has virtually the entire descriptive and logical coverage of natural language; this formal language is the basis of modern work on mechanical reasoning as well as language engineering.

The crucial properties of predicate calculus, and similar logical languages, are as follows.

- They represent only the bare informational aspect of meaning (i.e. facts and logical truths), ignoring other aspects of meaning such as atmosphere or emotional tone.
- Each formula has a single precise meaning, unlike a sentence in natural language which may have several possible interpretations.
- They are not biased towards any particular natural language: exactly the same formula can express the meaning of an English sentence and the meaning of the corresponding French sentence.
- They convey no stylistic information whatever. If two leaflets express the same content, then their meanings can be represented by exactly the same formulae, even if the first leaflet is couched in a convoluted technical style while the second is worded simply and informally.

In short, by using a logical language we can abstract the information content of a document from all its other features, including not only the formatting and layout, but also the linguistic style, the wording, and even the natural language in which it is written. Such a clear separation of content from all aspects of presentation brings many potential benefits. Interpreting the 'style sheet' more broadly as a presentation brief, we can go far beyond the control over formatting and layout allowed by word processors or by DSSSL. Our presentation brief might require that the meaning be expressed in concise but informal language, using sentences no longer than 15 words — and in all official languages of the European Union. If market research then showed

that patients in Italy and Germany prefer a more formal style, the Italian and German leaflets could be automatically regenerated after a slight modification to the style parameters. A task that previously took months would become as easy as switching the paragraph formatting from a tab to a double line.

How close are we to creating computer systems with this power? The answer to this question is complex, since some lines of research have advanced further than others. Let us first list the components that such a system would need.

1. A logical language rich enough to express the range of possible meanings in our target documents (e.g. the content of patient information leaflets).
2. Grammatical and lexical rules for all supported natural languages, allowing any relevant meaning to be expressed in a variety of ways.
3. Some way of modelling all the presentational features that document designers need to control, all the way from paragraph formatting to linguistic style to choice of natural language.
4. Some way in which the document designer can define the required meaning and presentation without having to learn knowledge engineering or mathematical logic.
5. Rules describing the interactions among style, layout and wording in the different supported languages. For instance, what are the stylistic implications of using 'larynx' instead of 'throat'? How should the wording of a series (i.e., list) be changed if it is presented through bullets rather than by running text?

Very broadly, we would assess progress on these various fronts as follows.

Representing meaning: Work on representing the informational content of texts is well advanced.

Grammatical and lexical rules: Work on formalising the grammatical and lexical resources of natural languages is also well advanced. Large-scale resources have been developed for English and several other languages.

Modelling presentational features: The main issue here is to identify the features that document designers most need to control. This is one of the aims of the ICONOCLAST project. Research is in its early stages.

Editing meaning: Perhaps the main obstacle to the commercial application of NLG technology has been the lack of any straightforward way in which a document designer can specify the desired meaning. Note that this meaning cannot be defined through a natural language text, or even by a text with SGML mark-up. It must be defined in a precise logical language. Recently we have found a potential solution to this problem, called 'WYSIWYM editing' (WYSISYM standing for 'What You See Is What You Meant') (see Power and Scott, 1998; Power, Scott and Evans, 1998; Scott, Power and Evans, 1998).

This method has proved effective for simple semantic material, but we have not yet confirmed that it is effective for complex material. We plan to use WYSIWYM in ICONOCLAST both for editing the meaning and for editing the presentation brief.

Interactions among style, layout and wording: Research on this topic is just beginning. This is the main aim of our work for ICONOCLAST.

In the remainder of this paper, we present some early results from ICONOCLAST on the interactions among style, layout and wording. We think these findings may prove interesting from two points of view. On the one hand, they can be expressed as formal rules to be incorporated into a future document generation system of the kind envisaged above. On the other hand, they may provide more immediate help to document designers, by indicating some aspects of usage and good practice which have not previously been formalised.

3 Issues in the Automatic Generation of Text with Layout

Our starting point in ICONOCLAST has been to uncover some of the interactions that occur between content and style by performing a detailed analysis of a corpus of laid-out documents. Although we expect to produce algorithms that can be applied to a wide range of texts, in order to get the most from our analyses we have had to confine our study initially to one genre. We have chosen to look at patient information leaflets (which we will hereafter refer to as "PILs"), which have many of the properties we need for our analysis: they make heavy use of layout; all cover the same topics, for example what's in the drug, how to use it, what to do before you use it, what to do after you've used it, precautions such as warnings and side effects; each manufacturer has its own distinctive house-style; there are many examples of different leaflets with the same content (ie., the same drug marketed by more than one company), and they are freely available. The aim of ICONOCLAST is thus to automatically generate this kind of document.

We have constructed a comprehensive on-line corpus of over 500 PILs from the *Compendium of Patient Information Leaflets* produced by the Association of the British Pharmaceutical Industry (ABPI), complete with their layout. In the remainder of this paper, we describe, by reference to examples taken from the corpus, some of the features of PILs that the ICONOCLAST system will be generating. In what follows we draw examples from 6 representative leaflets, produced by different companies and covering 3 types of products: antibiotic creams, hormone-replacement patches and anti-inflammatory skin creams.

In this article we focus on three aspects of PILs, common to many other genres of *visually informative documents* (Bernhardt, 1986) where the interaction between layout and text present special challenges for the automatic generation of documents: the grammatical structure of the constituent sentences, the organisation of information, and the way in which information is referred to.

3.1 Grammaticality

It's fairly well accepted that the notion of "grammaticality" of sentences differs between spoken and written language. We expect the basic units of written texts to be sentences that conform to the standards of the language as found in grammar books (e.g., Quirk et al 1985), but do not hold to such a strict requirement for speech; conversely, what is well-formed in speech often does not pass muster in text. Our analysis of the corpus of PILs leads us to conclude that laid-out texts share many of the properties of speech in this respect. Simply put, laid-out texts abound with sentences which would be considered to be grammatically ill-formed in standard, running texts but which are "extra-grammatical" in that they conform to a grammar acceptable to the genre.

This poses special problems for the automatic generation of PILs by the ICONOCLAST system. Typically, natural language generation systems are developed to guarantee that the text it produces conforms to the accepted grammar of the language. With the aim of generating PILs, we now have the additional requirement of ensuring that deviations from the standard (i.e., normative) grammar occur only in the right places and in the right way. We describe in this section some of the types of "extra-grammaticality" in terms of deviations from the standard conception of an orthographic sentence (i.e., minimally a grammatical clause; beginning with a capital letter and ending with a full-stop or question mark) that we are having to cover in ICONOCLAST. The most common locations of extra-grammaticalities in PILs are headings, warnings, conditionals and lists.

Headings

Perhaps the most noticeable feature of headings is that they are strongly marked by layout. Typical devices used to set headings apart from the preceding and following text include the use of white space, borders, a different font or face, and increased type-size — either individually or in combination.

Another notable feature is that headings only very rarely consist of a standard grammatical sentence. For example:

What this medicine is used for
(Voltarol, Geigy)

• **How to use your cream**
(Betnovate, Glaxo)

Even in the rare cases where they do consist of a standard grammatical sentence, the accompanying punctuation is typically non-standard. While almost all will begin with a capital letter, *none* will end with a full stop. However, interrogative forms tend *always* to end with a question mark. For example:

What should I do if I forget to use the Cream?
(FuciBET, Janssen-Cilag)

What is Diclomax SR used for?
(Diclomax, Parke-Davis)

The strong use of layout features in headings appears to licence such deviations from the normative grammar of the language.

Warnings

Warnings in PILs (not surprisingly) are also often strongly marked by layout features. Most typical devices are the use of bold face and upper case, often together. Unlike headings, however, they are almost always a normatively grammatical sentence – and this is perhaps strongly related to the fact that they are also almost always in the imperative form. Despite this they do not always end with a full stop. It is important point to note, however, that deviations from standard sentence punctuation seem to occur only in those cases where the warning is marked by layout. Compare, for example:

IT IS IMPORTANT TO READ THIS CAREFULLY BEFORE STARTING TREATMENT
(Betnovate, Glaxo)

with

If you accidentally take too many Voltarol 75mg SR tablets, tell your doctor at once or contact your nearest hospital casualty department.
(Voltarol,Geigy)

Conditionals

Conditionals in the PILs corpus follow the usual "If X is true then do/do not do Y." or "Unless X is true then do/do not do Y." format. They depart from the normative grammar, however, in that although they invariably consist of a standard grammatical sentence marked by the standard punctuation forms, they often contain within them other sentences. Once again, one finds that such deviations are invariably accompanied with layout features. Consider the three examples below, all taken from the same leaflet:

If you are not sure then follow the advice on the back of this leaflet.
(Betnovate, Glaxo)

Here we have what we would normally expect to find in a standard text.

In the second example

Unless told by your doctor:

- YOU SHOULD NOT use more than this.
- YOU SHOULD NOT use on large areas of the body for a long time (such as nearly every day for many weeks or months).

(Betnovate, Glaxo)

the consequents are each presented as a separate orthographic sentence, even though they logically form part of the larger sentence beginning with *unless* and ending with *months*. The logical structure of the sentence proper is, however, strongly marked by layout: the conditional adverb of the antecedent clause in bold and the presentation of the consequent as a vertical list marked with indentation and dashes. Without the use of these layout devices, this sentence would have had to be presented as something like: *"Unless told by your doctor, you should not use more than this or on large areas of the body for a long time (such as nearly every day for many weeks or months)"*.

In the third example

IF you find your condition gets worse during treatment you may be allergic to the cream or have a skin infection which needs other treatment.
STOP USING THE CREAM AND TELL YOUR DOCTOR AS SOON AS POSSIBLE
(Betnovate, Glaxo)

both antecedent and main consequent are full grammatical sentences. The first is presented in the standard way, but the second does not have the expected final full stop. However, we see once again that the complex of antecedent + consequent is bracketed by the layout: the use of capital letters for the conditional adverb of the antecedent ("if") and throughout for the consequent. A related phenomenon is shown below:

...**If you get any of the following:**

Stomach pain, indigestion, heartburn or feeling sick for the first time.
Any sign of bleeding in the stomach or intestine, for example, passing black stools.
...
An unexpected change in the amount of urine produced and/or its appearance.

STOP taking the tablets and tell your doctor.
(Voltarol,Geigy)

When a number of conditionals occur together, other devices are required. For example:

- Sunbathing – always make sure your patch is covered by clothing.
- Using a sunbed – either cover up your patch as above or take it off and put it back on after your shower when your skin is completely cool and dry.
- Swimming – you can wear your patch beneath your swimming costume during swimming.

(Estraderm TTS, Ciba)

Each problem-solution relation is presented within a bulleted point in a single orthographic sentence. This sentence is (normatively) non-grammatical: it expresses the problem in a VP-ing form separated with a dash from its solution which is expressed in a full clause. Also, the use of the dash seems compatible with the abrupt syntactic change occurring between the two propositions it links. It is interesting to note that when this kind of content is presented in a more conventional form, the grammatical properties are similarly more conventional. For example, compare what we have just seen with the following example taken from another company's leaflet for an almost identical product, and with therefore much the same conditions of usage :

Can I wash, bathe or shower as normal?
YES, but do not scrub too hard in case you loosen the edges of the patch.
Can I go swimming with the patch on?
YES, the patch will not be affected.
Can I sunbathe with the patch on?
YES, but keep the patch covered to avoid direct sunlight.
(Evorel, Janssen-Cilag)

Once again we see that layout is used to distinguish the antecedent from the consequent. In this case, however, the conditionals are presented as a dialogue between the reader and the writer, and the normative rules for written dialogue apply.

Lists

The layout device of vertical lists presents a number of exceptional features. We saw before that items in a list can be orthographic sentences even when they form part of a larger sentence unit. We have also encountered many cases where units as small as a single noun attract standard sentence punctuation. For example:

Are you taking any of the following:

- Anticoagulants (blood thinning tablets like warfarin)?
- Lithium or digoxim?
- Methotrexate?
- ...
- Any other medicines which your doctor does not know about?

(Voltarol, Geigy)

Also prevalent are cases where no standard punctuation marks whatever are used inside a list. For example:

You should find that

two fingertips of cream will treat both hands or one foot
three fingertips of cream will treat one arm
six fingertips of cream will treat one leg
fourteen fingertips of cream will treat the front and back of one trunk

Do not worry if you find you need a little more or a little less ...

(Betnovate, Glaxo)

In this example, the items in the list would perhaps be more suitable to an even stronger graphical device like a table. In many respects this list combines the graphical features of a table with the linguistic features of text, in the process losing other features of each form.

Laid-out lists present particular problems for ICONOCLAST since (a) they seem to break so many of the rules of the normative grammar but yet appear to behave in a fairly well ordered (but not yet well understood) manner and (b) they are extremely prevalent in the PILs domain. Non-laid out (i.e., horizontal) lists, on the other hand, appear not to deviate from the standard.

3.2 Organising Content

A major activity of writing is the process of organising the content of the document — deciding how best to distribute the required informational content between clauses, sentences, paragraphs, sections and the like. Once done, the writer has to make sure that he or she presents the information in a way which makes its organisation clear to the reader. Punctuation is one the many devices available to the writer for doing this. Layout, as we will show, is yet another.

In order to train a computer to generate documents, we need to be able to provide it with formal grammars governing the operation of the available devices. As we mentioned earlier, a number of formal grammars already exist. However, although the normative rules of English punctuation are provided in a number of style manuals —perhaps the best known of these is *The Chicago Manual of Style* (CMS, 1993)— and although punctuation is something that all writers understand and use (some more successfully than others), a formal grammar of its behaviour has only recently emerged (Nunberg, 1990), and we will be drawing heavily on it in ICONOCLAST. No such grammar yet exists for layout, and a large part of our work in developing the system will be to uncover (from our study of the PILs corpus) and formalise some of the rules that operate in the production of laid-out text. We describe in this section some of the phenomena we are addressing in ICONOCLAST pertaining to the interactions between the organisation of content and its presentation in laid-out texts.

Factoring Information

One of the powerful guiding principles for text organisation is factoring; that is, organising the text around the common elements of its content. Consider the following pair of excerpts taken from two leaflets from different companies but pertaining to the same type of product (hormone replacement patches). The propositional content of these examples are almost identical – the values of 3 domains: sizes, quantities contained and quantities delivered. Their expression, however, is widely variant.

Evorel patches contain a natural oestrogen called oestradiol. The patches come in 4 different sizes: Evorel 25, Evorel 50, Evorel 75 and Evorel 100. The patches contain 1.6 mg, 3.2 mg, 4.8 mg, 6.4 mg of oestradiol and deliver 25, 50, 75 and 100mcg of oestradiol respectively per 24 hours.

(Evorel, Janssen-Cilag)

In this leaflet, the information is presented in *horizontal lists in running text*. Each set of values from each domain is presented in a separate clause of its own: sentence 2 is about sizes, sentence 3 is about quantity contained and sentence 4 is about quantity absorbed. The order of distribution of the information is signalled with the adverb *respectively* which, for example, signals to the reader that it is Evorel 25 which contains 1.6mg of the drug, of which 25mcg is released every 24 hours.

A rather different picture emerges in the second example:

> **Estraderm TTS patches contain a substance called oestradiol. They come in three sizes:-**
>
> - Estraderm TTS 25 containing 2mg of oestradiol. Your body will absorb 25 micrograms of oestradiol each day whilst you are wearing Estraderm TTS 25 patch.
> - Estraderm TTS 50 containing 4mg of oestradiol. Your body will absorb about 50 micrograms of oestradiol each day whilst you are wearing an Estraderm TTS 50 patch.
> - Estraderm TTS 100 containing 8mg of oestradiol. Your body will absorb about 100 micrograms of oestradiol each day whilst you are wearing an Estraderm TTS 100 patch.
>
> *(Estraderm TTS, Ciba)*

This time the information is organised in a *vertical list in laid-out text*. Each item of the list is organised around the domain of size; this element has been factored out and used as the *key domain* (Douglas and Hurst, 1995) and is announced in the lead-in sentence "They come in three sizes". The effect of the selected distribution of the content is like reading out the following table from left to right, top to bottom:

Size of patch	*Quantity contained*	*Quantity delivered*
Estraderm TTS 25	2mg	25mcg
Estraderm TTS 50	4mg	50mcg
Estraderm TTS 100	8mg	100mcg

The tabular organisation of the content in this example is emphasised by the use of linguistic parallelism. Such strong parallelism in text so close together would be stylistically infelicitous were it not for its appearance within a list. Additionally, the fact that the information is presented in a vertical list licences, as we saw in an earlier example, the use of extra-grammatical sentences: each item begins with a verbless sentence. Finally, because the information is laid out in a vertical list, the nominal items can be elaborated upon in the second sentence; this would not be permitted in continuous prose, where such elaboration could only be inserted within parentheses. Similar observations are made by Twyman (1985).

Joining and Separating Information

The following example (encountered earlier in our discussion of grammaticality, but repeated here for convenience) illustrates some of the ways in which layout can be used as both a joining and separating device:

> . . . **If you get any of the following:**
>
> Stomach pain, indigestion, heartburn or feeling sick for the first time.
> Any sign of bleeding in the stomach or intestine, for example, passing black stools.
> . . .
> An unexpected change in the amount of urine produced and/or its appearance.
>
> **STOP taking the tablets and tell your doctor.**
>
> *(Voltarol,Geigy)*

Note that the lead-in and lead-out elements of the list are emboldened, while the items of the list are not. This non-conventional use of style alternation allows the reader to join the discontinuous segments together in a single sentence, much as though the list of side-effects were

a parenthetical. This can be performed in a more straightforward manner by the reader than if alternation had not been used. The parenthetical nature of the list items with respect to the lead-in and lead-out is further emphasised by the indentation of the items relative to the lead-in and -out.

The use of a vertical list to present the various side-effects provides additional opportunities for grouping the conceptually related symptoms together: stomach related problems, then blood problems, then skin problems, etc.

It is perhaps interesting to note that the layout feature *newline* can be used as an additional separating device to the punctuation feature *comma*. Its equivalent in this respect in continuous prose is the use of the *semi-colon* arising from the application of the *comma promotion rule* (Nunberg, 1990), whereby separating commas are *promoted* to semi-colons when the items they separate also contain commas. An example of comma promotion can be seen in the following:

> Do not use Evorel patches if you have any of the following: severe liver, kidney or heart disease; blood clots (thrombosis); inflammation of the veins; ...
> *(Evorel, Janssen-Cilag)*

However, it seems that the newline provides an even stronger perception of separation than the semi-colon, and thus makes the information easier to locate. Indeed, laid-out texts clearly provides more opportunities for signalling the separation of information units, since it allows for two levels of promotion whereas running text allows for only one (imagine a list within a list within a list).

3.3 Scope of a Referring Expression

Writers have at their disposal a number of ways to refer to information contained elsewhere in the document, and perform this task with ease. Computer algorithms for doing this already exist (e.g., Dale and Reiter, 1995), but none take account of the effect of layout on the behaviour of referring expressions, and there are indeed no available studies on how these two aspects of documents interact. But interact they clearly do. For example, common sense tells us that a referring expression such as *this* appearing in a context where it immediately follows some boxed-in text, will more likely be interpreted as referring to the information expressed by the full content of the boxed paragraph than to a single element within it. Many, rather more subtle, effects such as these appear to operate. Part of our work in ICONOCLAST is to uncover and formalise some of them.

The following example from our corpus illustrates the relation of scope of a referring expression to layout. Such examples are rather common in the PILs corpus.

> As with any other medicines, in a few women, Evorel patches may cause some unwanted effects. Most are usually minor and will disappear within a couple of months, for example:
> – Headaches
> – Local skin irritation
> – Nausea
> – Tender breasts
> – Irregular vaginal bleeding
>
> Rarely, dizziness, bloating, fluid retention, weight gain and leg cramps may occur.
>
> If such side effects are prolonged, or if you react to the patch itself with skin rashes or irritation, remove the patch and ask your doctor for advice. If you notice any other symptoms not listed above whilst using the patch, please tell your doctor.
> *(Evorel, Janssen-Cilag)*

The referring expression *such side effects* is introduced in a new paragraph, and its scope is both the horizontal series in the immediately preceding paragraph and the vertical series in the paragraph before that. If the sentence containing the referring expression had been lined up after the horizontal series, then its scope would only be that of the horizontal series. On the other hand, *any other symptoms not listed above* refers meta-textually to a region of the document and seems to cross those paragraph boundaries: had it been attached to the horizontal list, it seems that it would refer back not only to the horizontal list but to the vertical list as well.

We can see the effect quite clearly by making a few small changes to a part of the Diclomax leaflet describing the drug's side effects. Compare:

> Other reactions more rarely seen are:
>
> - Peptic ulcers and gastrointestinal bleeding
> - Bloody diarrhoea
> - Skin troubles such as eruptions, itching or bruising
>
> If these side effects occur you should tell your doctor.

with:

> Other reactions more rarely seen are:
>
> - Peptic ulcers and gastrointestinal bleeding
> - Bloody diarrhoea
> - Skin troubles such as eruptions, itching or bruising
>
> If these side effects occur you should tell your doctor

In the first case, the expression *these side effects* is interpreted as a reference to *all* the reactions listed. But in the second, it seems to refer only to the reactions given in the last bullet item in the list.

4 Conclusion

Our aim to automate the production of visually informative documents makes the explicitation and formalisation of the relation between layout and wording during the composition process a fundamental issue. These relations are used with great frequency by expert writers and information designers, but the rules governing their behaviour are (a) highly complex and (b) not generally available for study since they form part of the writer/information-designer's implicit knowledge. The first phase of our work in developing ICONOCLAST is therefore to provide a systematic description of the interaction between language and layout/punctuation devices, as well as explanatory principles relating this behaviour to linguistic and perceptual properties. In this article, we present some of our preliminary findings on some of the topics that must be addressed in the development of the system.

We have presented a selection of environments in the PIL genre where normative grammar rules for text are consistently flouted. In all cases that we have encountered, the text in question also includes the use of layout features, for example, contrasting font and face, indentation, bulleted lists. In other words, the *extragrammatical* texts are also strongly graphical. Our observations on the corpus strongly suggest that the use of layout features licences deviations from standard written English. We have also shown that layout provides ways of structuring information that are radically different from running prose.

We also observed some other remarkable non-conventional uses of layout — such as indentation and face alternation for bracketing in the context of a vertical list with lead-in and lead-out text — which appear to behave in much the same way as parentheses do in running text. Lastly, we have pointed out that layout plays a role in the interpretation of the scope of a referring expression in a similar way that a pair of parentheses prevents the material that is within it to be referred to outside its scope (Nunberg, 1990). The parallelism and complementarity between punctuation and layout, which we have been able to touch on only briefly here, is yet another of the many issues being addressed in ICONOCLAST.

Our approach to uncovering the implicit rules of good information design is heavily data-driven rather than prescriptive, relying as we do on the behaviour of the individual elements of layout and linguistic features of documents in our corpus. Through

experimentation with the system, we will be able to refine the rules as the system develops. We expect that the relationships we uncover will show the way into the *good* composition of professional documents.

References

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

Bernhardt, Stephen (1985), "Text Structure and Graphic Design: the Visible Design". In *Systemic Perspectives on Discourse*, pp 18–38, Eds. J.D. Benson & W. Greaves, Vol 2.

Clark, James (no date) "Jade — James' DSSSL Engine" available on the World Wide Web at "http://www.jclark.com/jade".

CMS (1993), *The Chicago Manual of Style*, The University of Chicago Press, 14th edition.

Dale, Robert and Reiter, Ehud (1995). "Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions". In *Cognitive Science*, 19:233-263.

Douglas, Shona and Hurst, Matthew (1995), "Using Natural Language Processing for Identifying and Interpreting Tables in Plain Text", in *Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, Nevada.

Nunberg, Geoffrey (1990), *The Linguistics of Punctuation*, CSLI Lecture Notes, number 18, CSLI Publications, Stanford, CA.

Power, Richard and Scott, Donia (1998) "Multilingual Authoring Using Feedback Text", *Proceedings of the Joint COLING-ACL Conference (COLING-ACL'98)*, Montréal, Canada.

Power, Richard, Scott, Donia and Evans, Roger (1998) "WYSIWYM: Knowledge Editing with Natural Language Feedback", *Proceedings of the 13th Biennial European Conference on Artificial Intelligence (ECAI-98)*, Brighton, UK.

Quirk, Randolph, Greenbaum, Sidney, Leech, Geoffrey and Svartvik, Jan (1985) *A Comprehensive Grammar of The English Language*, Longman.

Russell, Bertrand (1945) *A History of Western Philosophy*, New York: Simon and Schuster.

Scott, Donia, Power, Richard and Evans, Roger (1998) "Generation as a Solution to its Own Problem", *Proceedings of the 9th International Workshop on Natural Language Generation (INLG-98)*, Niagara-on-the-Lake, Canada.

Twyman, Michael (1982) "The Graphic Presentation of Language", *Information Design Journal*, pp 2–22.