A Linguistic Approach to Some Parameters of Layout: A Study of Enumerations

Ch. Luc, M. Mojahid, J. Virbel
IRIT
Université Paul Sabatier
118 route de Narbonne
31062 Toulouse cedex
France
e-mail : {luc,mojahid,virbel}@irit.fr

Cl. Garcia-Debanc
Laboratoire Jacques Lordat
Maison de la Recherche
5 allées Antonio Machado
31058 Toulouse cedex
France

M.-P. Péry-Woodley
ERSS
Maison de la Recherche
5 allées Antonio Machado
31058 Toulouse cedex
France
e-mail : pery@univ-tlse2.fr

Abstract

This paper reports on a multi-faceted study of enumerations involving linguists, psycholinguists and computer scientists. Our first point is that layout must be seen as a combination of lexico-syntactic and visual features, which we call "formatting", rather than restricted to visual features. The article lays down the theoretical bases for a model of text architecture making explicit the relations between discursive and visual formulations.
The corpus-based study of enumerations, which considers standard and non-standard forms, enables us to put the architecture model to the test. We identify major markers (typographical, layout, lexico-syntactic markers), and arrive at a fine-grained characterization and a classification of enumerations. As regards text organization, we show that in order to arrive at a complete and precise representation, we need to articulate and integrate the two models used: the architecture model and RST.

In this paper, we present a linguistic approach to some parameters of layout. We will however refer to "formatting", rather than layout, as the relations between syntactic, typographic and layout features are central to our outlook (see section 1). The term "formatting"[1] denotes all these features.
Our motivation is twofold. First, we consider formatting devices to have a specific semantics: it can be shown that different formatting choices represent different text structures, associated with different textual meanings; moreover, formatting differences have an impact on comprehension and recall (see last section). Our first motivation is to determine the best text formatting according to specific needs. Secondly, it has been shown in previous work (Pascual 1996) that text formatting cannot be done at the final step of a text generation process: certain formatting decisions have to be made very early on, and are essentially involved with the text production process. We therefore want to characterize the dependencies between the visual structure of texts and other text structures, as the retorical component.

This paper will mainly focus on the study of enumerations. We start with a presentation of our linguistic approach to text formatting and some aspects of a model of text architecture. We go on to explain the choice of enumerations by referring to specific properties of writing and we give a general definition of enumerations. In a third section we analyze three examples extracted from a corpus (Virbel 1999), which illustrate different structural forms of enumerations. In the next section, we propose a way to represent these enumerations using RST (Mann & Thompson 1987) and the model of text architecture. Finally, we give some results of a psycholinguistic experiment designed to test the impact of formatting on comprehension and recall of information.

A Linguistic Approach to Text Formatting

Virbel shows (Virbel 1989) that there is a relation of functional equivalence between formulations based on visual formatting and discursive formulations. Thus, to each formulation using formatting devices can be associated an equivalent formulation with running text. A specific textual metalanguage was constructed by Pascual (Pascual 1991); this metalanguage is based on an inventory of linguistic counterparts of formatting properties. In the light of Harris' view of the relation language/sublanguage/metalanguage (Harris 1968) (Harris 1982), a method was developed for matching discursive and reduced formulations, i.e. formulations exhibiting "traces" of the metalanguage (typographical, positional and syntactic markers). The formatting

[1] The original term is "mise en forme matérielle".

properties of a text (visual and syntactic aspects) signal particular entities that we call *text objects*: chapters, sections, paragraphs, but also definitions, enumerations, etc. The set of text objects and their relations in texts defines the architecture of text. Text architecture is therefore an abstract component of text, in the same way, for example as the rhetorical component. Moreover, the examination of the utterances of this specialized sublanguage shows that they are built around performative verbs. This performativity is interpreted as textual speech acts (in the sense of Searle) whose illocutionary force is directed to the text itself.
On the basis of the linguistic method to capture textual metalanguage, a model of text architecture has been developed in (Pascual 1991). This model is based on the notion of *metasentence*, where a metasentence is a formal sentence of the metalanguage for text architecture. The model provides 30 metasentences which cover a large range of textual phenomena occurring in scientific texts. The text architecture of a document is represented through a metadiscourse which is a list of instantiated metasentences.

Example

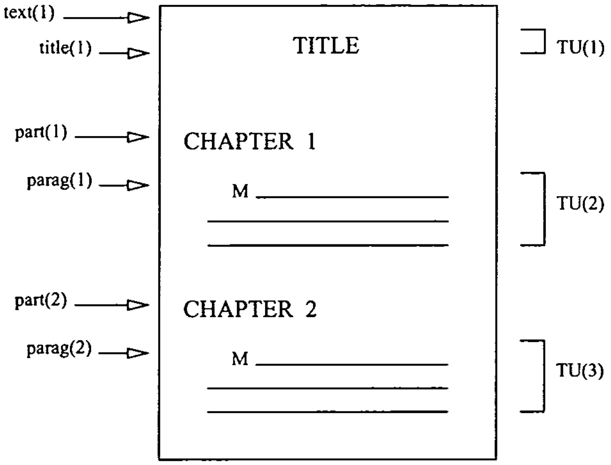


Figure 1: A text image

Let's take the text image[2] of figure 1. The corresponding metadiscourse is :
The author creates a text identified as *text(1)*.
The author gives a title to *text(1)* identified as *title(1)*.
The author attaches *TU(1)*[3] to *title(1)*.
The author composes *title(1)* from *TU(1)*.
The author organizes *text(1)* in two parts identified *part(1)* and *part(2)*.
The author assigns the level of *chapter* to *part(1)*, *part(2)*.
The author develops a paragraph identified as *parag(1)*.
The author attaches *TU(2)* to *parag(1)*.
The author composes *parag(1)* from *TU(2)*.
The author composes *part(1)* from *parag(1)*.
The author develops a paragraph identified as *parag(2)*.
The author attaches *TU(3)* to *parag(2)*.
The author composes *parag(2)* from *TU(3)*.
The author composes *text(1)* from *title(1)*, *part(1)*, *part(2)*.

The text architecture of this example is as follows: the text is constituted by a title and two chapters, each chapter is composed by a paragraph.
In the metadiscourse, the text objects are represented by a unique identifier (as *text(1)*, *part(1)*, ...). Two metasentences play a particular role in the metadiscourse: the metasentence "attach" and the metasentence "compose". For the first one, TU corresponds to a text unit, a segment of running text, which does not content another text object. The second metasentence is used to specify how text objects are put together inside another text object (the largest is the whole text itself). There exists rules that control the composition of text objects (see (Pascual 1996), (Pascual & Virbel 1996), (Luc 1998) for precisions).
Moreover, two formal properties of the model control the introduction of a metasentence in a metadiscourse: a relation of *Precedence* and a relation of *Obligation*. These relations are anti-reflexive, anti-symmetric, transitive and partial. For the graph of these relations, see (Pascual 1991).
A metadiscourse is therefore a set of instanciated metasentences that observe properties of coherence and cohesion (Luc 1998).

After this rapid account of our theoretical framework, we now focus our paper on enumerations. We will see below how the model of text architecture represents enumerations.

The Case of Enumerations

The impact of the invention of writing has tended to be envisaged mostly along two lines: writing has made it possible for speech to defy time (through specific notation systems); writing has given speech the property of ubiquity (through the duplication and the mobility of written records). But the conception of alphabetic writing as more or less regular correspondence between sounds and characters (writing as code or transcription) has limitations which have been pointed out by a number of authors:

- on the one hand, there are many important aspects of speech which are not represented in writing, or only in a very imperfect manner (rhythms, melodies, accents, ...). These shortcomings are clearly exposed in Nunberg's detailed study of punctuation (Nunberg 1990).
- on the other hand, a character or a string of characters cannot escape having layout and morphological properties (shape, size, incline, ...). These properties have no direct oral counterparts.

[2] The notation "M __ " represents a string of characters, i.e. a segment of running text.
[3] TU designates a Text Unit.

Obviously, moving from one-dimensional speech to a two-dimensional page and on to a three-dimensional book (i.e.: spatialization of language) opens up fresh expressive and cognitive possibilities (linked to morpho-spatial interpretations). This new potential, which is mostly outside the realm of oral communication, at least in complex forms (itineraries, inventories, recipes, ... (Ong 1988), (Goody 1977)), results from the construction of original concepts: page, double page, window, margin, title, paragraph, enumeration, margin notes, footnotes, ... These concepts are necessary for the control of the activities linked to literacy .

We became particularly interested in enumerations because they constitute such a remarkable case of exploitation of these possibilities. First, the written form allows the development of enumerations which can be as long and as embedded as necessary. Secondly, enumerations are an obvious case of correspondence between discursive forms based on adverbial expressions (firstly, secondly, ...; then, moreover, finally; ...), and syntactically reduced forms containing typographical and positional traces of these reductions (numbering, diacritics, horizontal spaces, ...). Finally, whatever the length and the complexity of items, writing offers unprecedented possibilities by allowing, unlike speech, the construction of enumerations out of syntactically or textually heterogeneous items.

These properties are even more remarkable when one considers that enumerations are extremely frequent and highly important in instructional texts. Their univocity is a necessary condition for the well-formedness of instructions, and the efficiency of their formulation determines their optimal use (e.g. in terms of a better recall or speed of retrieval).

A Definition of Enumerations

Enumerating has been defined as follows: "to enumerate is to attribute an equal level of importance to entities and to classify these entities according to various criteria" (Pascual 1991). This definition conforms to the widespread view of enumerations: items correspond to entities which are functionally equivalent and are realized through identical formatting (bullets, numbering, line breaks, ...).

When we started our fine-grained corpus-based study of enumerations (see next subsection), we came across many cases that contradicted this definition, which focuses on the parallelism between function and formatting in enumerations. These "exceptions" led us towards a new conception of the enumerative structure.

Indeed, we show that the speech act involved in enumerating is realized by a set of linguistic and visual properties which cannot be dissociated. These properties are the counterparts of the mental act of making an inventory of the entities which can be enumerated. Most importantly, we stress the relation between the possibility of co-enumerating items (presented as of equal importance) and the possibility of co-enumerating the entities in the inventory. This approach takes into account different types of enumerations which do or do not have the same syntactic or visual item structures.

Construction of a corpus

We started our study by collecting enumerations (Virbel 1999). Our objective being to construct a model encompassing non-standard forms, our corpus-collection followed somewhat unusual principles: we focussed on enumerations in contradiction with the first definition, i.e. enumerations with items which do not have the same function or formatting realization in texts. Moreover, we selected only enumerations displaying heterogeneous spacing, typographic or syntactic marks.

Our corpus is constituted of 75 enumerations extracted from various texts: scientific articles, books, newspaper articles, on-line texts, instructional texts, ... Most of the enumerations selected are in French but we have some examples in English. It would appear from this initial collection exercise that non-classical forms of enumerations are more common in French than in English; they do occur in English however, and must therefore be taken into account.

A first observation is that, in our corpus, all enumerations (i.e. sets of items) are preceded by that we call an *introducer* of the enumeration. It is characterized by a combination of markers which may be lexical ("the following points"), typographical (":"), positional (line break) or syntactic (the introducer and the first item form a text clause). The introducer serves to announce the enumeration without being a part of it. We could, for example, consider that numbered pieces of informations (bigger than a text sentence) are considered as an enumeration if they are preceded by an initial organizer.

A striking feature is the diversity of enumerative forms in texts. Moreover, they can be present in texts at various levels: sentence level[4], paragraph level or part level. Most of the time, there is no direct correspondence between the structure of the enumeration (i.e. the structure of the items) and the structure of the text (sentences, paragraphs and parts).

We isolated two main categories of enumerations:

- enumerations in which at least one item (and/or introducer) is smaller than a text sentence. In this case, items could correspond to a clause but we observe many cases in our corpus where the items are smaller than a clause. In a simple enumeration, all combinations can be encountered: for example, the introducer and the first item form a clause, and the other

[4] We use the term sentence in the sense of text sentence as defined in (Nunberg 1990).

items are separate sentences or clauses. These observations raise a number of problems: syntactic and semantic relations between the items, and between the introducer and the items; the problems of RST segmentation and analysis of these enumerations.

- enumerations in which items are bigger than a text sentence. In these cases, the problem comes from the interaction between the enumerative structure and other text structures as paragraphs, parts. Indeed, most of the times these structures do not correspond: a model which aims to represent a complete text structure must not be hierarchical.

Some Examples of Enumerations

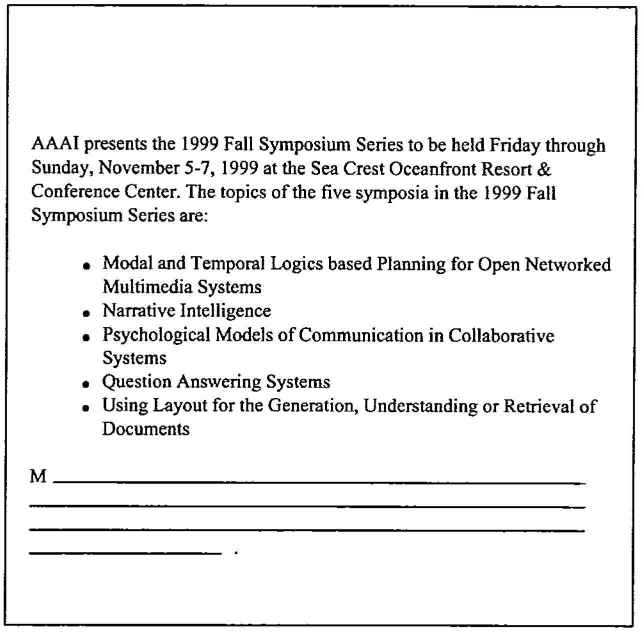
In this section, we give a detailed account of three examples of enumerations. We make use of text images, which have been developed to allow a representation of the architectural structure of texts (see section 1). Text images only present the entire strings when it is relevant to our purpose.
The first one is a classical version of enumeration. This enumeration is formed with a introducer and five regular items. This type of enumeration is probably the most common in texts (especially in English texts).
The second example was found on a web page. It exhibits identical formatting for items that are not syntactically equivalent.
The last example is extracted form a draft version of the article "Unresolved Issues in Paragraph Planning" by Eduard Hovy. It is characterized by two levels of structure: a segmentation in paragraphs and an enumeration.

Example of a classical enumeration

This enumeration (figure 2) correspond to the most widely shared view of the enumerations. It presents a characteristic structure: an introducer and five items. The introducer is characterized by lexico-syntactic ("five") and typographical (":") features and a vertical space with the first item. The features for the items are: the use of bullets, they are on a new line, an indentation.
The introducer is an incomplete sentence: it corresponds to factorization of informations; the items are equivalent from a syntactic and a semantic point of view.

The Web example

At first sight, this example (see FIG. 3) appears regular. It starts with an introducer signalled by lexical and typographical markers ("following four claims" and a colon). Moreover the items are presented with identical formatting: numbering, line break and indentation. But, if we look at the syntactic structure of the items, we notice that they are not equivalent. Indeed, the last item is a subordinate clause to the third item ("where"). Moreover, the first item is a sort of introduction to the second and third items: it introduces the notions of "contents of thought" and "contents of concepts". Contrary to the first example, the items of this enumeration cannot be interchanged.
These kinds of constructions occur frequently in our corpus: equivalent formatting and presentation are used for entities that are not functionally or syntactically equivalent. In order to describe these differences in the internal structure of enumerations, we appeal to the basic linguistic notions of paradigmatic versus syntagmatic relations. Through these relations we are able to specify the parallelism or non-parallelism between items without a semantic connotation. It can be compared with the Nucleus/Satellite distinction in the RST. In our example, we have a syntagmatic relation between item 3 and item 4, a paradigmatic relation between item 2 and item 3.



AAAI presents the 1999 Fall Symposium Series to be held Friday through Sunday, November 5-7, 1999 at the Sea Crest Oceanfront Resort & Conference Center. The topics of the five symposia in the 1999 Fall Symposium Series are:

- Modal and Temporal Logics based Planning for Open Networked Multimedia Systems
- Narrative Intelligence
- Psychological Models of Communication in Collaborative Systems
- Question Answering Systems
- Using Layout for the Generation, Understanding or Retrieval of Documents

M ____
____
____
______________ .

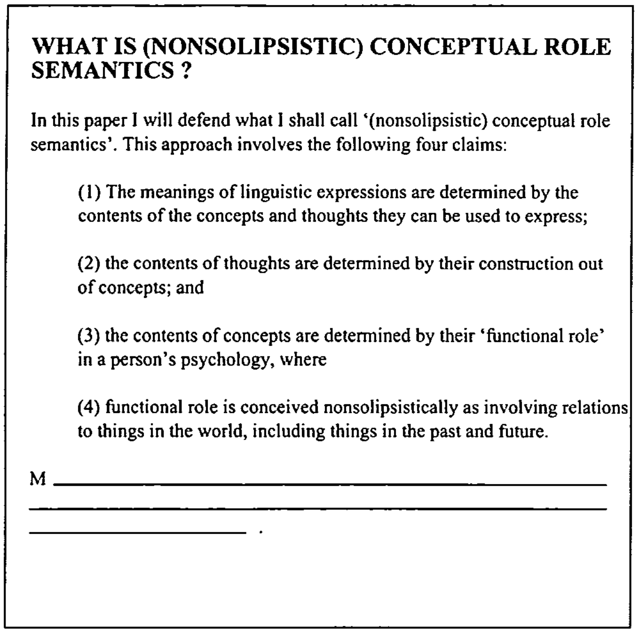
Figure 2: From the AAAI Fall symposium Call for Registration

The Hovy example

We present this example as a text image (figure 4, at the end of the paper). The boxes used correspond to the physical pages. The image respects the number and the position of paragraphs in the pages. The title of the part appears in bold face, and only significant parts of running text are reproduced. The image also reveals the two-level structure of the part: in the left margin, we indicate the different paragraphs (numbered from 1 to 9) and in the right margin the items of the enumeration.

This example presents three interesting cases:

- uncertain membership of an item in an enumeration;
- interdependencies between two enumerations; and



WHAT IS (NONSOLIPSISTIC) CONCEPTUAL ROLE SEMANTICS ?

In this paper I will defend what I shall call '(nonsolipsistic) conceptual role semantics'. This approach involves the following four claims:

(1) The meanings of linguistic expressions are determined by the contents of the concepts and thoughts they can be used to express;

(2) the contents of thoughts are determined by their construction out of concepts; and

(3) the contents of concepts are determined by their 'functional role' in a person's psychology, where

(4) functional role is conceived nonsolipsistically as involving relations to things in the world, including things in the past and future.

M

Figure 3: Extracted from a web page

- double structure of the part: paragraphs and enumeration.

The first problem comes from the last item. Indeed, the first three items of the enumeration are introduced by classical lexico-syntactic markers ("One issue ..."; "A second issue ..."; "A third issue ..."). Then, the next paragraph starts with "Furthermore", which is a textual organizer and may introduce an item. But it is impossible to ascertain that it is the fourth item of the enumeration as the marker belongs to a different class and no clue is given in the introducer of the enumeration as to the number of items to expect. Yet, here, the ambiguity is raised when one reads the last paragraph, which is not the case in several examples of our corpus. When an ambiguity persists, we call this phenomenon the uncertain membership of an item to an enumeration.

The second point is the relation between two enumerations. In the example, the last paragraph of the part is an enumeration. This enumeration is composed of four items and is a summary of the previous one, as is made clear by its introducer. These two enumerations are linked: each item of one enumeration corresponds to an item of the other enumeration. So, the ambiguity introduced with "Furthermore" is raised by the last paragraph: the part which starts with "Furthermore" is indeed an item of the first enumeration. This case of dependencies between enumerations is more frequent in the opposite form: the introducer of an enumeration is itself an enumeration whose items introduce the items of the main enumeration. In this case, there is an isomorphism between the structure of the two enumerations. A case in point is the current segment of this article: the current text is the second point of an enumeration whose introducer is an enumeration. There are other examples of dependencies between enumerations: for example, an item is constituted by an enumeration.

Finally, we consider the two levels of structure highlighted in the text image: there is a structure based on a paragraph segmentation and another based on the enumeration. We observe that the introducer of the enumeration and the first item: the first paragraph is composed of the introducer of the enumeration (i.e. the first sentence) and the beginning of the first item. The remainder of the first item corresponds to the next two paragraphs. The structure is very specific and typically non hierarchical: a hierarchical model of representation of the text structure (as for example, RST) is not able to represent this structure. We will see below how the model of text architecture represents this kind of structure.

These observations raise questions about the function of paragraphs in written texts. Heurley (Heurley 1997) suggests a distinction between paragraphs, defined as visual text units, and informations blocks, which are structurally or semantically organized text units. His definitions are based on the results of a psycholinguistic experiment which tested the role of paragraphs in the reading process. This definition fits with our observations of interactions between paragraphs and enumerations.

To conclude this section, we would like to stress the importance of considering cases which question the classical view of enumerations and stretch the existing models of text organization. A study of enumerations must not be limited to classical enumerations, but must also look at special cases: these cases, even if they are not the most frequent, are present in texts, and are significant of what writers actually do when they enumerate.

Some results from the corpus analysis

The corpus analysis took as its starting point a number of informal observations:

- items within an enumeration may belong to different classes of syntactic or textual constituents;
- irregular factorization (e.g. of prepositions) and coordination are frequent;
- organizers within an enumeration may be heterogeneous;
- the final boundary of enumerations is often unclear;
- complex dependencies may exist between embedded or adjacent enumerations.

In order to refine these observations, we proceeded to a systematic examination of 45 enumerations from the corpus. This led to the elaboration of an analytical framework and representation language allowing the identification of major structural forms and marker configurations. Enumerations have been characterized in terms of:

- their introducer,
- the relation between the items,
- organizers and other item markers.

Introducers Introducers belong to one of two types: incomplete sentence or leading sentence. An incomplete sentence introducer is a syntactically unfinished sentence, whose missing constituent/s is/are provided by the items of the enumeration. A leading sentence introducer is a syntactically complete sentence which announces the items to be enumerated. These two types of introducers are characterized by specific configurations of lexico-syntactic, typographical and layout markers.

Relation between items In a classical enumeration, the items are in a paradigmatic relation (syntactically or textually equivalent constituents). In a significant number of our enumerations however, the items are in a syntagmatic relation. In most of the cases characterized as syntagmatic, each of the items is a constituent, and the set of items forms a sentence. There are also examples of a syntagmatic relation at discourse level, whereby for example the items of the enumeration constitute the points of an argument. Finally, there are a number of hybrid enumerations where some items of a paradigmatic enumeration are in a syntagmatic relation. This is the case with the Web example.

Organizers and other item markers We have sought to bring out regularities in the form of the items by looking at the co-occurrence of the following markers:

- typographical markers (-, italics, bold, capital initial letter, ...);
- different forms of numbering;
- lexico-syntactic markers.

Whereas the first two types apply to all enumerations, the third concerns only "classical" paradigmatic enumerations, which take two forms:

1. items display parallel structure, defined in syntactic, typographical, and layout terms, as in the first example.
2. each item begins with an organizer such as *first*, *in the first place*, *secondly*, ..., *finally* or, as in the Hovy example, a classifier with a numeral (*One issue...*, *A second issue...*, *A third issue...*). These organizers can often be heterogeneous: in the Hovy example, the last item is introduced by an organizer from another "family", Furthermore.

Some tentative observations can be derived from this analysis regarding the interaction between lexico-syntactic and visual markers. As predicted by the architecture model, the use of visual devices is lower when items are structured by lexical organizers; on the other hand they are very prevalent in hybrid enumerations, and, more unexpectedly, with parallel structure.

The Representation of Enumerations

Beyond the characterization of enumerations, another objective of this study is to propose connections between two models of text structure: RST (Rhetorical Structure Theory) (Mann & Thompson 1987) which has become the reference theory for text generation and the model of text architecture. The detailed analysis of examples of enumerations found in texts leads us to the view that neither of these theories taken separately are able to represent enumerations correctly (Luc, Mojahid, & Virbel 1999). The RST allows for the construction of treelike structures for representing texts: in the case of enumeration, we have shown that these kinds of structure are not appropriate. On the other hand, the model of text architecture is not able to represent the rhetorical structure of texts. We propose compositions and combinations of these two models for representing enumerations.

An earlier study focusing on definitions (Péry-Woodley 1998) has drawn connections between these two models. It suggested that the markers which signal text objects (i.e. typographic, positional and/or lexico-syntactic) may serve as a basis for RST segmentation. This approach resulted in a dual representation of text (RST and text architecture), which took the form of an RST-tree where Text Spans are also text objects.

The problem of segmentation of texts

A major problem for a common representation comes however from the segmentation of text. Whereas an architecture segmentation is mainly based on visual features, the RST segmentation, even if no strict formal definition of basic units is given by Mann and Thompson, is mainly based on syntactic features: "...for interesting results, the units should have independent functional integrity..."; "...the units are roughly clauses..." (Mann & Thompson 1987)

As was proposed for definitions, a better solution could be to determine common Text Units for both theories. In practice, most of the time, the Text Units correspond to text clause.

In the case of enumerations, some items or introducers are smaller than a text clause: the segmentation should then be based on the visual features. An example is the classical enumeration presented in the previous section: the introducer is syntactically incomplete, it should be considered as a RST text unit.

On the other hand, in the Hovy example, the RST segmentation must not be based on visual features. Indeed, the visual structure of the text imposes a segmentation in paragraphs but the semantic structure, and so the RST analysis, corresponds to the enumeration.

The Web example

Here, a hierarchical structure seems appropriated to represent this example from two different perspectives (RST and architecture). The structure is presented in figure 5.

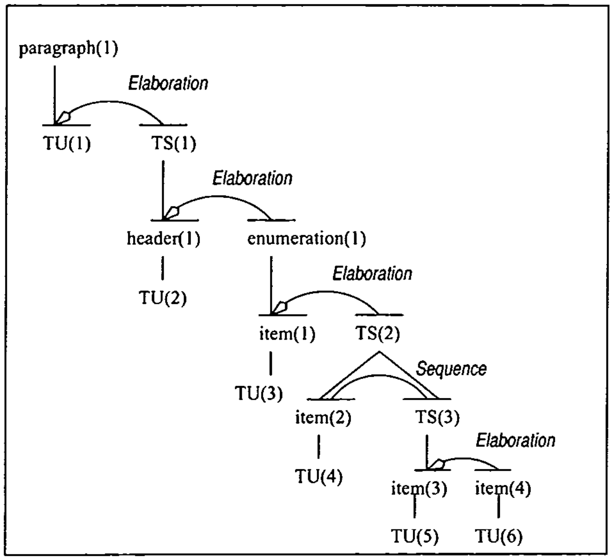


Figure 5: RST analysis of the Web example

The Text Units are common to the two models: they corresponds to the text sentences or the text clauses of the example. In accordance with the properties of the model of text architecture, we attach a TU to each text object. The text objects are indicated by their identifiers; the introducer is call a "header"[5]. This identifier is used, in the model of text architecture to designate a text object which introduces another text object but which is not included in this object. Some Text Spans do not correspond to text objects: they are marked out as TS.
This representation is mainly based on the RST principles: it is a treelike structure with Nucleus and Satellite and where we add text objects. This structure respects the four construction rules of the RST (Mann & Thompson 1987, p. 7-8): completeness, correctness, uniqueness and adjacency.

The Hovy example

We start with a partial presentation of the metadiscourse associated with the example. Indeed, the model of text architecture was constructed in a text generation perspective and is non-hierarchical. The properties and the constraints that govern the construction of a metadiscourse are flexibles enough to capture a text architecture as defined in the Hovy example.

We only present below an extract of the metadiscourse[6].

...
Ed[7] distinguishes an enumeration identified as *enum(1)*.
Ed organizes *enum(1)* in 4 items identified *item(1)*, *item(2)*, *item(3)*, *item(4)*.
Ed heads *enum(1)* by a header identified as *header(1)*.
Ed attaches *TU(1)* to *header(1)*.
Ed composes *header(1)* from *TU(1)*.
Ed develops a paragraph identified as *parag(1)*.
Ed attaches *TU(2)* to *parag(1)*.
Ed composes *parag(1)* from *header(1)*, *TU(2)*.
Ed develops a paragraph identified as *parag(2)*.
Ed develops a paragraph identified as *parag(3)*.
...
Ed composes *item(1)* from *TU(2)*, *parag(2)*, *parag(3)*.
....
Ed closes *enum(1)* by a coda identified as *coda(1)*.
Ed distinguishes an enumeration identified as *enum(2)*.
Ed heads *enum(2)* by a header identified as *header(2)*.
Ed attaches *TU(10)* to *header(2)*.
Ed composes *header(2)* from *TU(2)*.
...
Ed develops a paragraph identified as *parag(9)*.
Ed composes *parag(9)* from *header(2)*, *item(5)*, *item(6)*, *item(7)*, *item (8)*.
Ed composes *coda(1)* from *parag(9)*.
...
Ed composes *part(1)* from *parag(1)*, *parag(2)*, *parag(3)*, *parag(4)*, *parag(5)*, *parag(6)*, *parag(7)*, *parag(8)*, *parag(9)*.
...

The metasentence "Ed heads *enum(1)* by a header identified as *header(1)*" defines a link between the enumeration and its introducer, in this case a header. As was mentioned in the previous section, a header introduces a text object without being part of it. This representation fits with our definition for the introducer of an enumeration.
The problem of the first item is here solved through the double membership of TU(2) to item(1) and parag(1): parag(1) is composed by header(1) and TU(2), and item(1) is as well composed by TU(2).
We interpret the second enumeration in the last paragraph as a kind of conclusion to the first enumeration. We therefore use the metasentence "Ed closes *enum(1)* by a coda identified as *coda(1)*". A coda, as for the header, is a text object which concludes another text object without being included in it. We then compose the coda of a paragraph and this paragraph of an enumeration with four items and the associated header. The dependencies between the two enumerations are then represented in our metadiscourse by the metasentence "Ed closes ...".

[5] This term is a translation of the French term "chapeau". It is taken from the journalistic terminology.

[6] The complete metadiscourse of this example comprises 55 metasentences.

[7] Ed stands for the author.

Finally, the last presented metasentence indicate that the part is composed by the nine paragraphs: it brings to the fore the visual structure of the part.

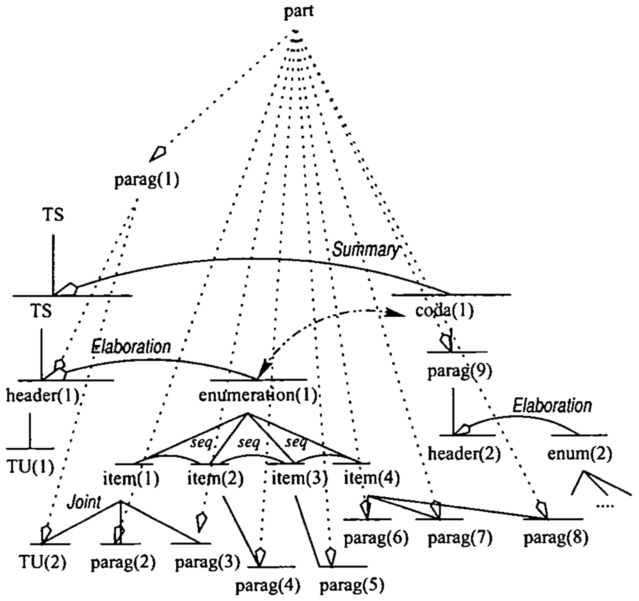


Figure 6: A representation of the Hovy example

Figure 6 is a representation of the RST and text architecture structure of the Hovy example. We only reproduce the relevant segments of text in the figure. The leaves of the representation are mostly paragraphs but a complete representation would have to take into account the RST Text Units.

The RST analysis is based on the enumeration segmentation: the relations link the text objects which are signalized by this segmentation. As we want to obtain as precise and complete representation as possible, we also show the link, as given by the metadiscourse, between the first enumeration and the coda.

The visual structure of the text, i.e. the segmentation in paragraphs, provides a parallel structure: we indicate this structure by dotted lines.

The whole structure is clearly non hierarchical. Further work will study properties and characteristics associated with these kinds of text structures in a text generation perspective. We are convinced that a complete and appropriate representation for the generation of formatted texts must be similar to this structure.

The Influence of Formatting on Text Comprehension

A psycholinguistic experiment was conducted in order to examine the influence of formatting on the comprehension and recall of a specific type of instructional texts: playing instructions for games (Grandaty, Degeilh, & Garcia-Debanc 1997). The model of text architecture provides the experiment with a framework for controlling different layout parameters. The experiment also aims to find out whether children are able to identify textual objects (such as titles, enumerations). Enumerations are a frequent device in playing instructions, where they concern three types of entities: objects, actions and situations.

The experiment takes as its starting point the relation expressed earlier between discursive and visual formulations. Several versions of the same text were constructed: from mostly discursive versions (enumerations using textual organizers and punctuation) to versions relying heavily on visual formatting devices (bullets and line breaks). These texts were experimented on different populations: children (9 and 12 year-olds) and adults. The data collected confirm the hypothesis that layout influences comprehension and recall for all groups, but the results are complex: different versions of a text give rise to varying performances according to the population concerned, and according to the cognitive process involved. For example, the version with visuo-spatial formatting leads to better selective retrieval (for all groups), whereas the version with discursive formulation is associated with better comprehension and recall (for adults and 12 year-old children). Our explanatory hypothesis is that discursive formulations require higher cognitive processing than visuo-spatial formulations, leading to better comprehension and recall.

This experiment convincingly shows the importance of adapting text formatting (in text generation, for example) according to user and communication goal.

Conclusion

We believe that layout must be seen as a combination of lexico-syntactic and visual features instead of only visual features, which is why we choose the term "formatting" to denote these features. This paper has laid down the theoretical bases for a model of text architecture making explicit the relations between discursive and visual formulations.

This corpus-based study of enumerations has enabled us to put the architecture model to the test. We have identified the major markers (typographical, layout, lexico-syntactic markers), and produced a fine-grained characterization of enumerations and a classification. We have shown that in order to arrive at a complete and precise representation, we need to articulate and integrate the two models used: the architecture model and RST.

This initial study opens up a number of research perspectives:

- pursuing the analysis of markers and relations which characterize enumerations;
- analysing further the status of enumerations within text, in particular links with other text objects such as titles or definitions;

- dealing with the tricky question of the final boundary of an enumeration.

Two main principles will continue to direct our approach to layout phenomena:

- focus on specific text objects, and
- an approach which combines different points of view within research on the genesis of text, coming from linguistics, psycholinguistics and computer science.

References

Goody, J. 1977. *The Domestication of the Savage Mind.* Cambridge University Press.

Grandaty, M.; Degeilh, S.; and Garcia-Debanc, C. 1997. Rôle de la mise en forme matérielle dans la constitution du sens des textes à consignes. In Pascual, E.; Nespoulous, J.-L.; and Virbel, J., eds., *Le texte procédural : langage, action et cognition*, 55–74. Mons, Gers: Prescot. Toulouse.

Harris, Z. 1968. *Mathematical Structures of Language.* John Wiley and Sons.

Harris, Z. 1982. *A Grammar of English on Mathematical Principles.* John Wiley & Sons.

Heurley, L. 1997. Processing units in written texts: Paragraphs or information blocks ? In Costerman, J., and Fayol, M., eds., *Processing Interclausal Relationships*, Studies in Production and Comprehension of Text. Lawrence Erlbaum Associates. 179–200.

Luc, C.; Mojahid, M.; and Virbel, J. 1999. Connaissances structurelles et modèles nécessaires à la génération de textes formatés. In *Actes du Colloque sur la Génération Automatique de Texte (GAT'99)*, 157–170.

Luc, C. 1998. Contraintes sur l'architecture textuelle. *Document numérique* 2(2):203–219.

Mann, W. C., and Thompson, S. A. 1987. Rhetorical structure theory : A theory of text organization. Technical report, ISI-RS-87-190, Information Sciences Institute, Marina Del Rey, Ca.

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

Ong, W. J. 1988. *Orality and Literacy. The Technologizing of the Word.* Routledge.

Pascual, E., and Virbel, J. 1996. Semantic and layout properties of text punctuation. In *International Workshop on Punctuation in Computational Linguistics, 34th Annual meeting of the Association for Computational Linguistics.* Univ. of California, Santa Cruz, USA.

Pascual, E. 1991. *Représentation de l'Architecture Textuelle et Génération de Texte.* Thèse de doctorat, Université Paul Sabatier.

Pascual, E. 1996. Integrating text formatting and text generation. In Adorni, G., and Zock, M., eds., *Trends in Natural Language Generation: an artificial intelligence perspective*, number 1036 in Lecture Notes in Artificial Intelligence. Berlin, New York: Springer-Verlag. 205–221. (Selected Papers from the 4th. European Workshop on Natural Language Generation, Pisa, Italy, 28-30 April 1993).

Péry-Woodley, M.-P. 1998. Textual signalling in written text: a corpus based approach. In Stede, M.; Wanner, L.; and Hovy, E., eds., *Proceedings of the Workshop "Discourse Relations and Discourse Markers"*, 79–85. Association for Computational Linguistics.

Virbel, J. 1989. The contribution of linguistic knowledge to the interpretation of text structure. In André, J.; Quint, V.; and Furuta, R., eds., *Structured Documents*, 161–181. Cambridge University Press.

Virbel, J. 1999. Structures textuelles - planches. fascicule 1 : Énumérations. Technical report, IRIT. Version 1.

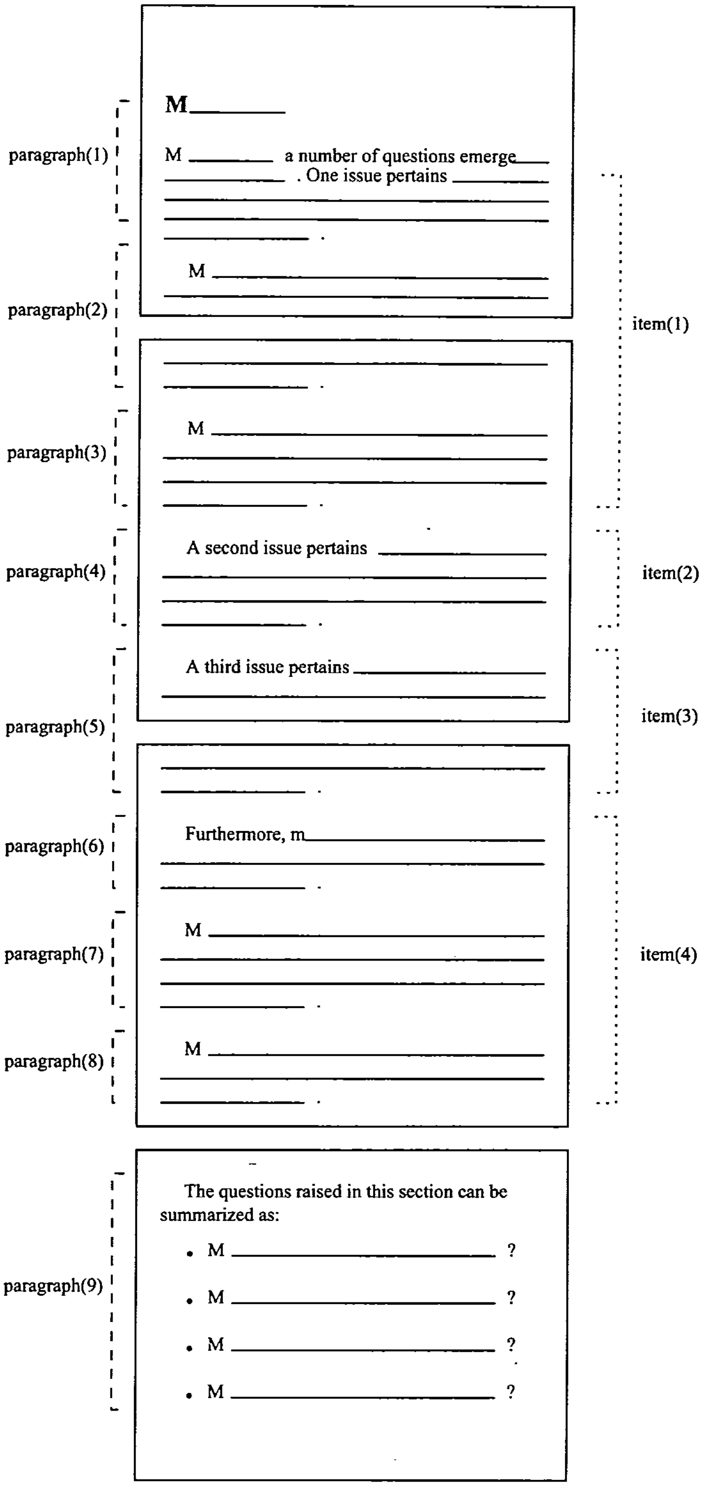


Figure 4: From a Draft Version of "Unresolved issues in Paragraph Planning"