Semantic and Layout Properties of Text Punctuation

Elsa PASCUAL and Jacques VIRBEL
Programme de Recherche en Sciences Cognitives de Toulouse (PRESCOT)
Institut de Recherche en Informatique de Toulouse (IRIT/CNRS)
Université Paul Sabatier, 118 rte de Narbonne, 31 062 Toulouse Cedex, France
pascual@irit.fr, virbel@irit.fr

Abstract

Higher-level graphical and lexical punctuation (paragraphing, indentation, font changes, …) must be taken into consideration in comprehension processes and text generation. In this paper, we analyse a class of text punctuation marks which includes lexical units (chapter, introduction). We give a method for the analysis of the semantics of these units, in terms of metalanguage. In addition, their syntactic properties are considered using the contrastive properties of their layout.

1 Presentation

It is generally agreed that beyond intrasentence punctuation, there is a textual punctuation system linking a series of sentences to form particular textual entities (e.g. the paragraph), or which is placed outside the main text (such as foot-notes).

The reasons for this extension of punctuation to the text are probably various. Among them, the deciding factors seem to be the following:

- although including the same sentence material in a textual sentence could alter the interpretation according to the nature of the separators (i.e. [;] vs [:] or [,]), this is not necessarily the case, and the same series of this kind of lexical sentence (within the meaning of Nunberg, 1990) can be written as a series of textual sentences, with a one-to-one mapping between the two sets. This kind of phenomenon can happen at different textual levels. (Drillon, 1991, p. 442) has to our knowledge gone the furthest in this direction in that he writes "The newline is to the chapter what the semi-colon is to the sentence [...] while the paragraph is to the chapter what the full stop is to the sentence". Another example of this continuity between the two types of punctuation is given by the fact that the same material can be found in a bracketed position or between dashes within main text, or as a foot-note outside main text.
- it is recognized that these two types of punctuation are not entirely independent, but can in fact have various interpolations. Thus a multisentence or even multi-paragraph textual unit can make up a intrasentence segment: it is the case for example with the content of quotations or brackets, or even with units like "Definition", "Theorem", etc. in mathematical texts; this also happens in lists, where an item can have an multisentence internal structure whereas the whole list where this item appears makes up only a single textual sentence (Werlich, 1976; de Beaugrande, 1980).

Nevertheless, there does not seem to be a general agreement between researchers regarding the marks of ultrasentence punctuation. Also considered as such, according to authors: capital letters, indents, new lines, headings such as "chapter", "section"; etc., numbers, roman-italic contrasts and more generally body size or style of characters, differences in typographical and positional properties of titles, differences in justification and line length, etc.

In this paper, we will consider marks allowing certain types of text entities to be distinguished as textual punctuation, that is: headings (i.e. lexical marks) as well as typographical and/or positional marks applied to these headings and/or to distinguishable textual entities. A formal definition will be given below in Part II. These marks are generally seen as marks of text structure, without any link with the phenomena of intrasentence punctuation. Nevertheless (see also Dale, 1991; Hovy & Arens, 1991), examining them as punctuation marks can be worthwhile in that they play a double role of textual unit segmentation, and signposting the textual function of the segment, which, at a generic level is also the case of intrasentence punctuation.

A first difference is clearly that these marks are character strings making up lexical entities, and that as such they include a semantic reference which contributes in a determining way to the signposting of the textual function. In addition, they can contain, even within the same text, many typographical and positional variations, which can contribute to denoting the syntactic relations between the units concerned. These variations generally operate in a differential manner. Thus for example, relations of inclusion between textual units can only be shown by variations in titles linked to these units, and not by marks such as "chapter" or by numbers. In this case, identities and differences show the structure of units and not absolute values of the design properties.

Consequently, there is a relatively particular situation to be considered with regard to intrasentence punctuation. On the one hand, the possible syntactic relations between units can be, at least in part, determined by the semantics of the terms which they denote: thus there can be sections within chapters, (or the contrary, but not in the same text), but not chapters within chapters; in the same way, a chapter can contain, and begin by, an introduction, but an introduction can itself make up a chapter, and this in the same text (but not in the same chapter). On the other hand, this syntax can be detailed, other than by the order in which marks are written, by certain of their typographical and positional properties, but not necessarily all, and not always the same, according to the sequences of marks allowed.

The following two parts deal with these two aspects.

2 Semantics of text punctuation

As we have outlined above, textual punctuation allows the following items to be perceived:

- particular entities, that we will call *textual objects*: chapters, sections, paragraphs, listings, lists and introductions, but also definitions, commentaries, theorems, and demonstrations, etc.
- the relations between these textual objets: inclusion; semantic link (theorem / demonstration, any textual object / commentary); logical link (introduction / textual object introduced); etc.

The set of textual objects and their relations in a text defines the *architecture* of the text. Textual architecture, a central notion in this part of the paper, is therefore an abstract component of text, in the same way, for example, as rhetorical component (Mann & Thomson, 1987; Maier & Hovy, 1990).

2.1 Characterizing text architecture from text punctuation marks

A close examination shows that there is a correspondence between text punctuation marks and linguistic phrases: each textual object can be expressed by a family of formulations, from running text to richer formatted text.
For example, a definition can be expressed by one of these three formulations:

I define A as B.
Definition: A is B
A is B

In fact, the combination of all possible typographic, positional and syntactic variations for a textual object (like a definition) leads to a number of possible forms of the order of one million. In this example, it can be seen that: in the first formulation, the linguistic means are predominant; in the last formulation, the definition is only partially expressed in words, but this is counteracted by the predominance of non lexical punctuation marks. This means that each formulation of a textual object with formatting devices can be associated with an equivalent formulation in running text. Then, instead of studying millions of forms of the same textual object, the study is restricted to only running text, expressing explicitly the textual object corresponding to the formulation. In this case linguistic methods of production and analysis can be used.

2.2 Textual function of punctuation marks

Producing an inventory of the linguistic counterparts of the formatting properties, and building up classifications of them, based on linguistic properties, leads to the development of specialized sublanguage (according to Harris, 1968) for text architecture. The examination of the utterances of this specialized sublanguage shows that they are built on performative verbs and propositional attitudes such as *to divide into, to enumerate, to define, to introduce, to entitle, to give an example, to make an observation*, etc. The performativity of the elements of this lexicon is interpreted according to J. Austin and J. Searle's research about speech acts (Austin, 1962; Searle, 1975). Indeed, these utterances are speech acts, where the illocutionary force is directed toward the text itself, indicating how the text segment must be interpreted as a specific textual object with regard to this force. Consequently, these linguistic utterances define a metatextual dimension of the text (Harris, 1968). Thus, this sublanguage is a metalanguage to express the text architecture. Some examples of (natural) sentences of such a metalanguage are: "I conclude by saying that", "this paper is divided into nine

sections", "after having defined this notion, I can formulate the following theorem", etc.
The hypothesis of the existence of textual metalanguage within text leads to an abstract interpretation of punctuation marks: they are seen as traces of specific speech acts. Following Z. Harris's theory about the relation language/sublanguage/metalanguage (Harris, 1968), we obtain a method for matching formulations in running text and reduced formulations with traces (here, the traces are typographic, positional and syntactic).

Having given this linguistic basis we can formally define what we consider to be *textual punctuation, it is any manifestation of architectural metalanguage in a text.* That is: lexical marks coming directly from metalanguage as well as any traces of reduction of metalanguage in the form of typographical and positional syntactic marks (nominalization for example).

2.3 Formalizing semantic expression of text punctuation

2.3.1 Production of the metalanguage

Now that the semantics of punctuation marks has been defined (see section above), the next question is the production of the metalanguage. The method we used to produce the metalanguage has been proposed by (Virbel, 1989). It is a linguistic method, based on formal lexicology (Gross, 1975). This method allows us to completely capture the metalanguage. We briefly present it here, by means of an example.

This method consists in searching the terms of the natural language observing certain properties. In the present example, they are the following:

- these terms refer to textual object *types*.
- they indicate a relation between textual objects,
- they fit in a set of schemata[1] where:
  - Ni are noun phrases representing either the author or a textual object.
  - V is a performative verb (or verbal expression).
  - prepi are prepositions (depending on the verb).
  - loc is a prepositional or adverbial phrase.
  - V-nom is the nominalization of V.
  - V-sup and V'-sup are the verbs associated to the nominalization.
  - brackets indicate an optional element.
  - + indicates a choice.

The schemata are the following:

(S1) N0 V (prep1) N1 loc N2
Marie introduit texte1 par texte2
Mary introduces text1 by text2
Marie traduit texte3 par texte4
Mary translates text3 into text4

(S2) N0 V-sup (prep2) V-nom (prep1 + de) N1 loc N2
Marie effectue la traduction de texte3 sous la forme de texte4
Mary carries out the translation of text3 in the form of text4

(S3) N2 V (prep1) N1
Texte2 introduit texte1
Text2 introduces text1

(S4) N2 V'-sup (prep2) V-nom (prep1 + de) N1
Texte2 constitue l'introduction de texte1
Text2 constitutes the introduction of text1

450 French terms verify all these properties at the same time. Amongst those terms are: introduction (*introduction*), traduction (*translation*), résumé (*summary*), réponse (*answer*), démonstration (*proof*), commentaire (*comment*), annotation (*annotation*), etc. A complete inventory of textual objects is developed by

- doing this work for all other classes of objects,
- for each set of terms, classifying them and following logic, semantic and formatting criteria.

[1] As the study has been done for the French language, here we give the syntactic schemata for this language. A translation of the examples associated to them is given.

For a more detailed presentation of the linguistic background of this work, see (Virbel, 1989; Landelle, 1987; Pascual & Virbel, 1993; Pascual, 1991; Pascual, 1996).

2.3.2 Formalizing the metalanguage : a model of representation of text architecture

On the basis of the linguistic method proposed above, we developed a model to represent text architecture. This model is based on the notion of *metasentence*, where a metasentence is a formal sentence of the metalanguage for text architecture.
Restricting our study to the context of scientific and technical documents, we extracted thirty classes of metasentences for the French language. These classes correspond to hundreds of French verbs (or verbal expressions).

Here we give some examples[2] of metasentence classes, in which:

- Mi are entities which may be instantiated by nouns taken from the corresponding set.
- idi represent textual object identifiers.
- idi.Mj means that the identifier idi must be of type Mj, i.e. must belong to the set associated with Mj.
- n is an integer.

1. *The author organizes id0.Mo into n parts identified by id1, ..., idn.*

   Mo belongs to the set of textual objects which can be divided into parts.

2. *The author assigns the level of Mcn to id1.part, ..., idn.part.*

   Mcn ∈ {book, volume, part, chapter, section, etc.}

3. *The author numbers id1.Mn, ..., idn.Mn.*

   Mn belongs to the set of textual objects which can be numbered.

4. *The author uses the Ms system to number id1.Mn, ..., idn.Mn.*

   Ms ∈ {Roman caps, Roman lower case, Arabic numerals, alphabet caps, etc.}

5. *The author introduces id0.Min by an introduction identified by id1.*

   Min belongs to the set of textual objects to which an introduction can be added.

6. *The author gives id0.Mt a title identified by id1.*

   Mt belongs to the set of textual objects to which a title can be added.

The first class of metasentences describes the organization of a textual object into *parts* such as a text, a part, a comment, etc. (The word *parts* has a technical meaning in this model). The second class represents books, volumes, parts (in the usual meaning), chapters, etc., which are *parts* (in the technical meaning) to which the author assigns a level. The third class represents the numbering of objects that can be numbered (which are described in the associated set). The fourth class indicates the system chosen for numbering. The two last classes represent the possibility, for a textual object, to be introduced or entitled. For the complete inventory of the classes of metasentences, see (Pascual, 1991).

To represent a given text, the corresponding classes are instantiated.
For example, the following[3] text is represented as follows :

[2] As this study has been done for the French language, we give here the terms which seem to apply to English as well.
[3] The notation: M___________ is used for any character strings.

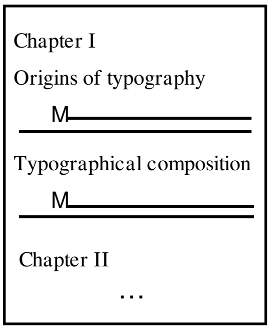

Figure 1

The author creates a text identified as t1
The author organizes t1 into two parts identified by p1 and p2
The author assigns the level of chapter to p1 and p2
The author numbers p1 and p2
The author uses the Arabic numerals system to number p1 and p2
The author organizes p1 into two parts identified by p3 and p4
The author gives p3 a title identified by ti1
The author gives p4 a title identified by ti2
etc.

This representation of the architecture of a text is called *metadiscourse*. All metadiscourse fulfils a certain number of properties ensuring that it is well formed.

3 Syntax of text punctuation

Text linguistic rules, and in particular the relation of inclusion between textual objects impose a certain syntax on textual punctuation marks, in terms of:

- order of lexical marks according to their content,
- typo-positional properties of lexical marks and /or textual segments.

Thus combinations of lexical marks, in a given order, lead necessarily to certain interpretations and rule out others (Cf. section 1. below). When there are several possible interpretations, morpho-positional rules are applied in order to convey the good interpretation or to desambiguate (Cf. section 2. below).

In this paper, we have chosen three lexical marks to illustrate this point in detail.

3.1 The linguistic component: order of lexical marks

A systematic linguistic study allows the set of accepted interpretations to be obtained from sequences of marks. These interpretations are described in terms of the model of representation of the textual architecture, given above.
We are going to present some results from our study in an informal way, summarized in two tables. The first deals with the analysis of sequences of two marks taken from the set {<*chapter* n°>, <title>, <*introduction*>}, and the second with sequences of three marks from the same set. The informal notation used to describe the corresponding relations of inclusion is the following: brackets are used for levels of hierarchy for the textual objects marked; the symbol + is used for interpretations which are accepted and - for those which are not.

sequences of marks	interpretation in hierarchical terms	
<*chapter* n°> <title>	(<*chapter* n°> <title>)	+
	(<*chapter* n°>(<title>))	+
<title> <*chapter* n°>	(<title> <*chapter* n°>)	–
	(<title>(<*chapter* n°>))	+
<title> <*introduction*>	(<title> <*introduction*>)	–
	(<title>(<*introduction*>))	+
<*introduction*> <titre>	(<*introduction*> <title>)	+
	(<*introduction*>(<title>))	+
<*chapter* n°> <*introduction*>	(<*chapter* n°> <*introduction*>)	+
	(<*chapter* n°>(<*introduction*>))	+
<*introduction*> <*chapter* n°>	(<*introduction*> <*chapter* n°>)	–
	(<*introduction*>(<*chapter* n°>))	+

Table 1

Thus, the sequence of marks <*chapter* n°> <title> can be interpreted as a chapter with a title (first line of the table), or as in figure 1, where the title has a lower hierarchical level to that of the chapter (second line).

sequences of marks	interpretation in hierarchical terms	
<*chap* n°> <title> <*intro*>	(<*chap* n°> <title> <*intro*>)	–
	(<*chap* n°>(<title>(<*intro*>)))	+
	(<*chap* n°>(<title> <*intro*>))	–
	(<*chap* n°> <title>(<*intro*>))	+
<title> <*chap* n°> <*intro*>	(<title> <*chap* n°> <*intro*>)	–
	(<title>(<*chap* n°>(<*intro*>)))	+
	(<title>(<*chap* n°> <*intro*>))	+
	(<title> <*chap* n°> (<*intro*>))	–
<title> <*intro*> <*chap* n°>	(<title> <*intro*> <*chap* n°>)	–
	(<title>(<*intro*>(<*chap* n°>)))	+
	(<title>(<*intro*> <*chap* n°>))	–
	(<title> <*intro*> (<*chap* n°>))	–
<*intro*> <title> <*chap* n°>	(<*intro*> <title> <*chap* n°>)	–
	(<*intro*>(<title>(<*chap* n°>)))	+
	(<*intro*>(<title> <*chap* n°>))	–
	(<*intro*> <title>(<*chap* n°>))	+
<*chap* n°> <*intro*> <title>	(<*chap* n°> <*intro*> <title>)	+
	(<*chap* n°>(<*intro*>(<title>)))	+
	(<*chap* n°>(<*intro*> <title>))	+
	(<*chap* n°> <*intro*>(<title>))	+
<*intro*> <*chap* n°> <title>	(<*intro*> <*chap* n°> <title>)	–
	(<*intro*>(<*chap* n°>(<*chap* n°>)))	+
	(<*intro*>(<*chap* n°> <title>))	+
	(<*intro*> <*chap* n°>(<title>))	–

Table 2

3.2 The morpho-positional component

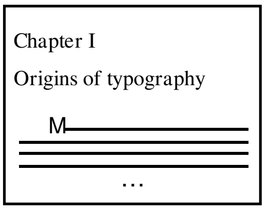

Figure 2

In the opposite text, a simple observation of the lexical marks cannot determine whether the title “Origins of typography” is that of Chapter I, or that of the first unit of Chapter I, without a term of division (for example: section).

The morpho-positional component is based on three principles:

a) An inventory of the morphological properties of character strings making up the lexical marks: typographical, in terms of body size, roman/italic contrast, bold, underlining, upper/lower case, typeface; positioning of these blocks : horizontal position, line structure (concatenation of marks on the same line) and vertical positions (inter-block values).

b) A criteria for contrast in communicational intention mark (Virbel, 1986, p. 38-40), founded on the perception of punctuation marks, in the sense of identity systems and differences in the values of properties (grouped in (a)). By this it is meant that (within the limits defined by perceptive capacities and physical design constraints) mainly relative values of certain properties will express structural values.

c) A functional selection of the types of property which are associated according to the nature if the ambiguity: this aspect, to our knowledge dealt with little or not at all in specialised work on page make-up and the arts of printing (however Cf. Twyman, 1981) for the role of vertical spaces with regard to titles), is nevertheless strongly testified, and it allows recourse to a strategy for a solution of the type “guided by the data”. Thus the ambiguity illustrated by figure 2 putting the marks <*chapter* n°> and <title> into play may be resolved by only taking into account the relative values of vertical spaces between these two marks, and between the second and the block of text, or their line structure, independantly of any other consideration (for example the body of these marks, or even their horizontal dispositions); the situation being very different if the two marks being considered were for example <*chapter* n°> and <*introduction*>, where the differences in horizontal position play a determining role.

Combining these three principles leads to the definition of a kind of rule where the conditional part is relative to the marks concerned by an ambiguity of interpretation (and implicitly the set of interpretations which can be associated to it) and the interpretative part gives, ideally according to a decreasing order of discriminative efficiency, the properties being usefully present in this case and the relative values of these properties.
The rule relating to the ambiguity of figure 2 thus has the following content :

if <*chap* n°> <:> <title>
then (<*chap* n°> <title>)
else if line-structure [<*chap* n°>,<title>] = one line
then (<*chap* n°> <title>)
else if horizontal position [<*chap* n°>] ≠horizontal position [<title>]
then (<*chap* n°>(<title>))
else if vertical-space [<*chap* n°>,<title>] > = vertical-space [<title>,< textual unit>]
then (<*chap* n°>(<title>))
else (<*chapter* n°><title>)

4 Conclusion

In this paper, we have presented the syntax and semantics of a limited set of text marks which can be seen as macro-punctuation marks, but with specific properties with regard to sentence punctuation: the lexicalisation of marks and the syntactic role of their layout.

We would like to go deeper into this approach, on two main themes :
- taking into account other classes of marks,
- studying the interdependence between this type of text punctuation and sentence punctuation.

One of the objectives of this kind of work could be to better define knowledge on ultrasentence punctuation, which to our knowledge has not yet led to a stable consensus of opinion among specialists.

5 References

Austin J. L. 1962. *How to do things with words*. Oxford University Press. London.

Catach N. 1994. *La ponctuation.* Presses Universitaires de France.

Dale R. 1991. Exploring the Role of Punctuation in the Signalling of Discourse Structure. *Workshop on Text Representation and Domain Modelling*. T.U. Berlin. pp 110-120.

de Beaugrande R. 1980. *Text, Discourse, and Process*. Ablex.

Doppagne A. 1978. *La bonne ponctuation*. Duculot.

Drillon J. 1991. *Traité de la ponctuation française*. Gallimard.

Gross M. 1975. *Méthodes en syntaxe*. Hermann.

Harris Z.S. 1968. *Mathematical Structures of Language*. Wiley & sons.

Hovy E. H.& Arens Y. 1991. Automatic Generation of Formatted Text. *Proceedings of the 8th Conference of the American Association for Artificial Intelligence*. Anaheim, CA. pp 92-96.

Jones B. E. M. 1994. Exploring the Role of Punctuation in Parsing Natural Text. *COLING 94*. pp 421-425.

Landelle M. 1987. *Analyse syntaxique de l'expression de la segmentation dans le lexique français*. Rapport de DEA. Toulouse, France.

Maier E. & Hovy E. H. 1991. A metafunctionally motivated taxonomy for discourse structure relations. In *3rd European Workshop on natural language generation*. Judenstein, Austria. pp 38-45.

Mann W. C. & Thompson S. A. 1987. *Rhetorical Structure Theory : Toward of Functional Theory of Text Organization*. Research Report RR-87-190. USC/Information Sciences Institute.

Nunberg G. 1990. *The Linguistics of punctuation*. CSLI.

Pascual E. & Virbel J. 1992. Connaissances linguistiques et morpho-dispositionnelles pour le contrôle de la décomposition structurelle de documents. *Colloque National sur l'Ecrit et le Document (CNED).* Nancy, France. In : Bigre. 80. pp 217-224.

Pascual E. & Virbel J. 1993. On the Unsuspected Importance of Material Shaping in the Processes of Text Production and Understanding. *ICCS 93: International Colloquium on Cognitive Science.* Donastia-San Sebastian, Spain.

Pascual E. 1996. Integrating Text Formatting and Text Generation. *Trends in Natural Language Generation: an Artificial Intelligence Perspective*. G. Adorni & M. Zock Eds. Springer. Heidelberg. pp 205-221.

Pascual E. 1991. *Représentation de l'architecture textuelle et génération de texte.* Thèse de l'Université Paul Sabatier. Toulouse, France.

Searle J. R. 1975. A Taxonomy of Illocutory Acts. *Language, Mind and Knowledge*. K. Gunderson Ed. University of Minnesota Press. pp 344-369.

Twyman M. 1981. Typography without words. *Visible Language*. 15, 1. pp 5-12.

Virbel J. 1986. Langage et méta-langage dans le texte du point de vue de l'édition en informatique textuelle. *Cahiers de Grammaire*. 10. pp 1-72.

Virbel J. 1989. The contribution of linguistic knowledge to the Interpretation of Text Structures. *Structured Documents*. J. André, V. Quint & R. Furuta Eds. Cambridge University Press. pp 161-181.

Waller R. 1980. Graphic aspects of complex texts: Typography as macro-punctuation. In P. Kolers, M. Wrolstad & H. Bouma Eds: *Processing of Visible Language*. Plenum Press. pp 241-253.

Waller R. 1982. Text as Diagram : Using typography to Improve Access and Understanding. In D. Jonassen Ed. *The Technology of Text*. Educational Technology Publications. pp 137-166.

Werlich E. 1976. *A text Grammar of English*. Quelle & Meyer.

White M. 1995. Presenting Punctuation. *Fifth European Workshop on Natural Language Generation*. Leiden, The Nederlands. pp 107-125.