Signalling in written text: a corpus-based approach

Marie-Paule PÉRY-WOODLEY
ERSS/PRESCOT, Université de Toulouse Le Mirail
5 allées Antonio-Machado
31058 Toulouse cedex, France
pery@univ-tlse2.fr

Abstract

The concern of this paper is the signalling of segments and relations in written texts. It explores the role of visual formatting and its relation to lexical and other markers. Through a corpus-based study of a specific "text object" – definitions – in instructional texts, it brings together two models of text structure: RST and the model of text architecture. Unlike RST, this latter model gives a central place to signalling, establishing a theoretically-motivated relation of functional equivalence between markers based on typography or layout and lexico-syntactic markers. Definitions in the corpus are characterised on the basis of configurations of markers, and their occurrences charted in the global structure of the text. The distribution of definition patterns highlights the dynamic nature of text: markers of a specific text object vary systematically according to where it occurs in the structural hierarchy of the text. The study establishes a relation between text objects and RST segments, thus opening the range of discourse markers to include visual formatting, and providing RST segments with a textual status.

Introduction

Discourse relations are heterogeneous; text organisation seems to work on several distinct levels (Cf. Moore and Pollack 1992). This complexity has been the focus of much research recently, with a number of authors appealing to Halliday's tripartite distinction of linguistic metafunctions – ideational, interpersonal and textual – in order to articulate different perspectives on discourse organisation, or different levels of description (Maier and Hovy 1993, Bateman and Rondhuis 1997). These authors explored ways in which the metafunctions could provide an organising principle for the classification of discourse relations and markers (otherwise classified as semantic vs. pragmatic, subject-matter vs. presentational, etc.). The *textual* metafunction, described by Halliday and Hasan (1976) as "the text-forming component in the linguistic system", comprising "the resources that language has for creating text" (ibid: 26) has tended to receive the least developed treatment. The focus of this paper is the textual metafunction, and its aim is to contribute to an understanding of the "resources" that are exploited to create textual meaning, more specifically markers of relations and segment boundaries.

My approach belongs in corpus linguistics, and is therefore guided by an awareness of the diversity of language productions. A first factor of variation is domain: a number of studies are concerned with the linguistic characterisation of domain sublanguages (Grishman and Kittredge 1986; Sager, Friedman et al. 1987) A second factor is genre, which subsumes social function, discourse purpose, channel. This study focusses on written texts with a specific discourse function – instructional – within a particular domain: software manuals. The specificity of written texts and its relevance to an understanding of discourse organisation must be stressed: firstly, in most cases, writing implies that the writer[1] and the intended audience do not share the context of communication. This has two major consequences for the organisation of written text: a) a written text is generally a monologue, where topics are introduced, continued or dropped not through negociation between discourse participants but on the sole basis of the writer's representations and intentions; b) there is a requirement for explicitness in the signalling of the various levels of meaning. Secondly, a written text is a visual object, and its visual properties are directly involved – and exploited by readers – in the construction of meaning. The choice of instructional texts derives from a hypothesis linked to the explicitness requirement: the social function of these texts is such that their writers are likely to try and leave as little interpretative leeway as possible. They therefore constitute a good starting point for a study of organisational signals.

Discourse theorists are generally agreed on a recursive structuring involving text segments and discourse relations. Many questions remain open, however, over the signalling of relations and the nature and status of the segments. In RST, the authors stress the absence of specific signalling of rhetorical relations. As for the segments concerned, the minimal units are defined as "typically clauses", but Mann and Thompson specify that the relations in fact hold between the

[1] I use the word *writer* for convenience, even though the production of a text may involve several agents.

meanings and intentions represented by the clause (Mann and Thompson 1989; Mann, Matthiessen and Thompson 1992). In other words, there is an exploitable correspondence between the syntactic unit clause, identifiable on the basis of surface characteristics, and the unit of meaning which is the argument of a relation. But what of the larger segments formed out of these basic units? Do they have a status of their own? Can they be identified on the basis of surface signalling? Some positive answers are proposed here, in the light of a model which describes texts in terms of an architecture of objects, and on the basis of a study of a specific text object – definitions – in software manuals. The notion of marker is broadened to include typographical and layout features, which we will see can be functionally equivalent to lexical markers.

1 The textual level

Halliday (Halliday and Hasan 1976; Halliday 1985) examines "the text-forming component in the linguistic system" at three levels of organisation:

– the clause: use of word order to signal theme, of phonological prominence to signal new information;
– the group or clause complex: use of syntax to signal interclausal relations, of punctuation to mark the sentence;
– the text: use of cohesion devices (reference, substitution, ellipsis, conjunction, lexical cohesion).

What I propose is an extension of this examination of "resources that language has for creating text" focussing on written text. Virbel and his group (Virbel 1985; Virbel 1989; Pascual 1991) have done extensive work on the visual aspects of text organisation, as one realisation of what will be called "formatting", though, as will be seen, it is formatting in a somewhat broader sense than the usual acception. The question which immediately arises is whether visual formatting can be seen as part of the resources of language. In answer to this question, Virbel (1985) convincingly shows the relation of functional equivalence (if one sets aside considerations of appropriateness to genre and stage in the text development) between formulations based on visual formatting and discursive formulations. The made-up examples in figure 1 will explain:

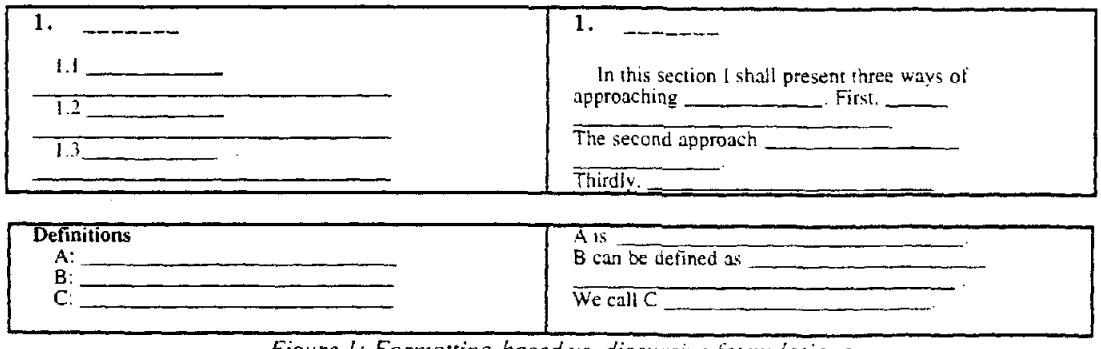

Figure 1: Formatting-based vs. discursive formulations

In the first example, the claim is that the same structuring is created – and meant to be recognised by the reader – in the text images on the left and on the right. Similarly, in the second example, three definitions are formulated, and meant to be recognised as definitions, in both cases. The formulations on the left are based mostly on layout, typography and enumerations, while those on the right, though not devoid of visual formatting, rely more on discursive means. These examples have been made fairly clear-cut for the purposes of the demonstration, but in-between formulations are obviously possible. The resources available for written text organisation thus appear as a continuum from wholly discursive to wholly visual. There seems to be no hard and fast conventions for layout and typographical enhancement, but rather a general principle of contrast.

On these observations the main tenets of the model of text architecture were formulated (Virbel 1985; Pascual 1991):

– these formulations are perceived as "equivalent" because they are interpreted as performing the same "text act", here organising and defining. Success of such text acts is that they be recognised, and that the text segments concerned (the arguments of the performative) be understood as sub-parts or as definitions. These are metalinguistic performatives whose performativity is directed at the text itself as text, and not at its ideational content or interpersonal purpose.
– the textual metalanguage, exemplified by the fully discursive formulations, is part of the language, and therefore open to description in terms of operator-argument relations (after Harris 1968; 1982). The operators are verbs such as *organise, entitle, illustrate, conclude, define*...; their arguments are text segments called text objects. A text object is therefore a

segment corresponding to a specific metalinguistic formulation and signalled by formatting. The notion of formatting[2] covers lexico-syntactic, typographical, layout, and punctuation markers.

This model of text organisation centres on the identification and characterisation of segments at the textual level. To what extent do text objects identifiable through formatting correspond to segments at other levels of description? If a correspondence can be established, the notion of marker, mostly geared toward lexical markers, could be radically broadened. But this requires that the relations between the textual and the other levels of text organisation be better understood. In order to broach these questions, our approach (Pascual and Péry-Woodley 1997a; 1997b; 1997c) has been to examine a specific text object in a subset of a particular genre. Some elements from the study of definitions in a software manual are presented in section 2.

2 Definitions in a software manual

2.1 Methodological preliminaries

2.1.1 The corpus

Our corpus consists of three software manuals. For the initial exploration of the text object definition, we selected a limited sub-corpus extracted from the manual of a text analysis and categorisation system called SATO[3]. The manual is organised in 7 chapters, numbered 1 to 7, and a number of peripheral objects such as acknowledgments and index. Our sub-corpus is chapter 6 (78 pages, 49 000 words), which is devoted to the description of the commands of one of the two main modules making up the system. The analysis below focusses on section 6.1, dedicated to a specific type of commands called "analyseurs".

2.1.2 The method

We produced a representation of the text in terms of the higher levels of architecture (parts, titles, paragraphs, examples, etc.). This representation was obtained on the basis of a top-down analysis by a first coder, the starting point being the visual formatting features – traces on the text's surface of the textual metalanguage – which make these text objects identifiable. Jointly, a bottom-up RST analysis was performed by a second coder. Definitions were then identified intuitively by the two coders. There was general agreement, though there remains some uncertain cases which will not be dealt here. Definitions in the corpus are signalled by configurations of lexico-syntactic, typographical and layout markers. Our final model of the grammar of definitions in this corpus, presented fully in Pascual and Péry-Woodley (1997b), is the result of several cycles of approximation-refinements. It presents a number of basic patterns which are one level of abstraction removed from the surface forms: they allow the grouping together of surface forms in terms of an analysis in Harrisian elementary phrases and transformations.

2.2 Representing the higher levels of text structure

2.2.1 A synthetic representation

The partial representation in figure 2 is a hybrid one: it shows the convergence between an analysis in terms of text objects and an analysis in terms of rhetorical relations. The schemas are therefore labelled both in terms of clausal units and relations, and in terms of text objects (see key below figure 2). "Part" is used as a generic term subsuming chapter, section, sub-section, etc. When it coincides with numbered parts in the manual, the original numbers are used (part 6.1.1); non-numbered parts are attributed a number (parts 26 to 31). For reasons of readability and space, figure 2 focusses on part 6.1.1[4].

The structure represented displays great regularity: it is a series of nested elaborations, which correspond to nested definitions. As mentioned before, part 6.1 of our manual describes/defines a set of commands called "analyseurs". At the first level (not shown), there is a preamble (pre 1) which is the nucleus of eight elaborations (parts 6.1.1 to 6.1.8) dealing with each "analyseur" in turn. Pre 1 is itself an elaboration schema. Part 6.1.1 is structured in the same way as part 6.1, with a preamble (pre 2) and an elaboration. Again the preamble is an elaboration schema. The body of part 6.1.1 is again an elaboration schema with a preamble (pre 3) as its nucleus, and three elaborations, of which the last two, an explanation and an example, will be analysed no further. The analysis of the remaining elaboration (parts 26 to 31) reveals a more complex structure, where related spans are not strictly adjacent[5]: text-span 7-8 and clause 9 are the nuclei of elaboration relations involving parts 26 and 27 (elaborating 7-8) and parts 28 to 31 (elaborating 9).

---

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

[3] SATO (Système d'Analyse de Textes par Ordinateur) is a system developed by F. Daoust at the Centre d'ATO of the University of Quebec at Montreal. It is the software used to search for occurrences of definitions in our corpus.

[4] The reader is asked to ignore at this stage the indications of definition types (BP, RP1-5), which will be dealt with in sections 2.3 and 2.4 below.

[5] I realise this is not conform to the tenets of RST. This anomaly seems linked to the list structure typical of the genre, which will be discussed later.

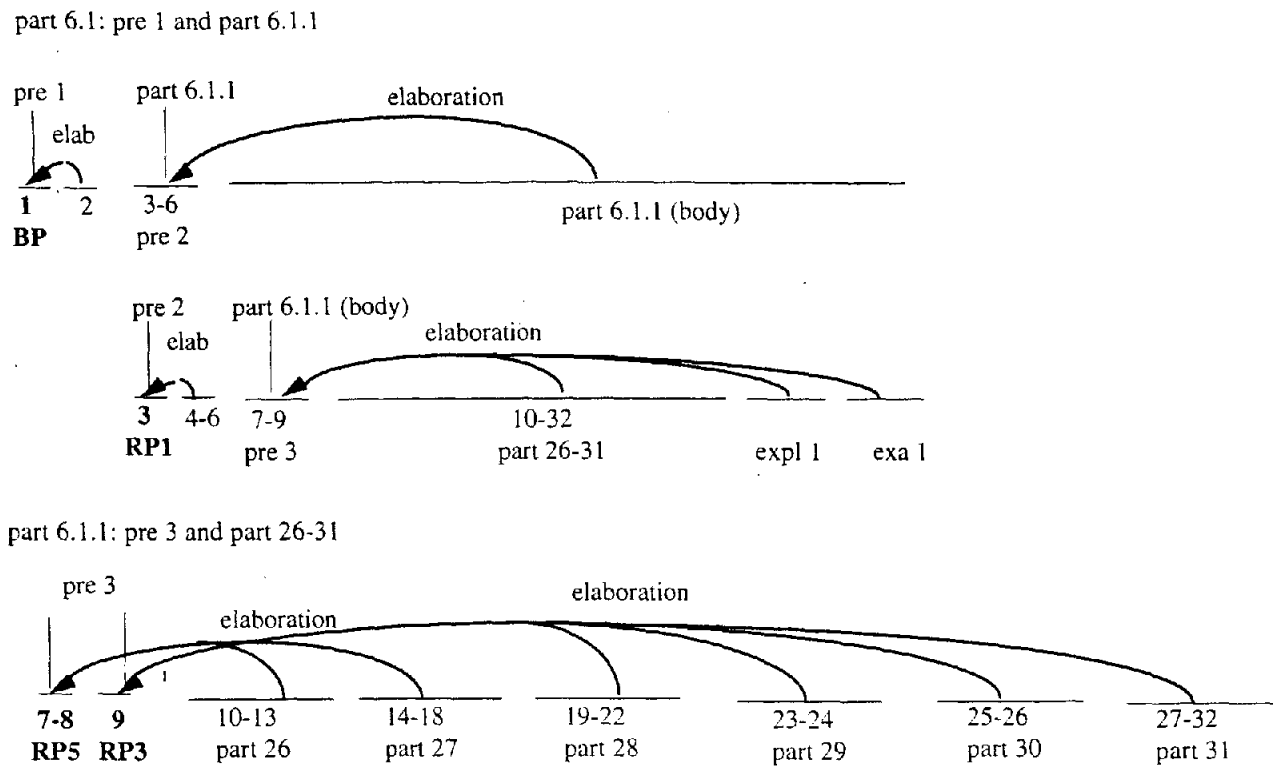

Figure 2. RST/architecture representation

Key to figure 2: 1-32 = clausal units
pre = preamble
expl = explanation
exa = example

2.2.2 Architectural segments vs. rhetorical segments

In this analysis, text objects, identified on the basis of formatting features, are all RST text-spans. This implies that formatting features can also be markers of rhetorical segments. The authors of RST, whilst stating that the analysis can be approached top-down as well as bottom-up, do not give any indication as to the identification of high-level segments. Yet analysts performing a top-down RST analysis are bound to use formatting to delimit high-level text-spans, as part of the interpretation process. The model of text architecture is an attempt at making explicit this aspect of text-meaning production. The congruence between architectural and rhetorical segments displayed in the reference text may not be generalisable. It is probably desirable, however, at least in certain genres, and could be developed into a principle in generation and composition instruction.
In this analysis, RST text-spans acquire a status at the textual level. This may be an organisational status, such as parts at different levels of the hierarchy, or a functional status, such as definitions. There appears to be a strong correspondence between some text objects and particular relation schemas: the definition patterns detailed in the next section are the nuclei of definitional text-spans which are all elaboration schemas[6]. Finally, definitions can be made up of definitions, just as elaboration schemas can be made up of elaboration schemas.

2.3 Characterising definitions

2.3.1 The patterns

Definitions in our text are signalled through a combination of discursive and visual formatting features. These are sufficiently recurrent and regular to allow the formulation of a basic pattern (BP in Table 1), where every distributional slot is filled, and of five reduced patterns (RP1 to RP5), where one or more element is missing. There is a gradation in the number of reductions: RP1 and RP2 involve one reduction; RP3 and RP4 involve 2 reductions; RP5 involves 3 reductions.

[6] Other such correspondences between particular expressions of the textual metalanguage (metasentences) and RST relations have been suggested in Pascual and Péry-Woodley (1997a).

Given the objectives of this paper, Table 1 only shows patterns actually occurring in the corpus. If the aim was to generate all possible formulations, whether in order to capture all potential forms for automatic recognition, or for text generation, it would obviously be easy to complete the table.

The patterns always coincide with the beginning of a paragraph; the word being defined is always typographically marked (capitals, bold, inverted commas). These layout and typographical features are an integral part of the patterns.

	Nc1	*Nn*	*Vi*	*Nc2*	*Vc*	*VP*
BP no reduction	§ La commande *The command*	**Distance** ***Distance***	est *is*	un analyseur lexico-statistique. *a lexico-statistical analyser.*	Elle permet de *It is used to*	comparer statistiquement les lexiques de deux sous-textes quelconques d'un corpus *statistically compare the lexica of any two sub-texts of a corpus*
RP1 reduction *Vi Nc2* **RP2** reduction *Nc1*	§ L'analyseur *The analyser*	**COMPARAISON** ***COMPARISON*** § Le filtre *The filter*	est *is*	un patron de fouille *a search pattern*	permet de *is used to* qui permet de *which is used to*	marquer ... *mark ...* définir ... *define...*
RP3 reduction *Nc1, Vi Nc2* **RP4** reduction *Vi Nc2 Vc*	§ L'analyseur *The analyser*	§ **EXPORTER** ***EXPORT*** **SEGMENTATION** ***SEGMENTATION***			permet d' *is used to*	enregistrer ... *record ...* découpe ... *segments..*
RP5 reduction *Nc1, Vi Nc2, Vc*		§**APPLIQUER** ***APPLY***				lance ... *starts...*

Table 1: Definition patterns

Key (the classes are distributional classes which have been functionally labelled, apart from the final verb phrase):

Nc: classifier noun
Nn : domain-specific name
Vi : "is-verb" {être, désigner}
Vc : "can-verb" {permettre, servir à, avoir pour effet, être utilisé pour, ...}
§ : indicates the start of a paragraph.

2.3.2 Interpreting the variation

A definition typically consists of two functional elements: the class, expressed by a hypernym, and the specificity, expressed by a modifier attached to the hypernym. Table 1 shows that the corpus displays little variation as regards the specificity (*Vc* VP or just VP), but the class can be expressed twice (*Nc1* and *Nc2*), once (*Nc2*) or not at all (in RP5). Before moving on to the next section, concerned with the distribution of these different patterns within the hierarchy of the text, I shall report some recent observations on "class-less" definitions: all occurrences of RP5 are found in list structures where the class is indeed expressed, but in the header of the list and not in every definitional item. Ongoing analysis of other software manuals confirms the regularities underlying the variations in the use of lists in definitions. The three examples in Figure 3 show how the class relation may be formulated with differing levels of reliance on visual clues. In the rightmost formulation, the interpretation of "Display" as a type of command relies solely on layout clues:

Three commands may be applied ...:	**Commands**:	**Commands**:
Display is a command which	– Display: this command ...	– Display: *<function>*
Export is a command allowing	– Export: this command ...	– Export: *<function>*
Print is a command which	– Print: this command ...	– Print: *<function>*

Figure 3: Lists and the expression of class

2.4 Mapping definitions onto the overall structure

The RST/architecture representation in figure 2 above indicates the position of different definition patterns in the structure. The nucleus of the preamble (elaboration schema) to part 6.1 is a basic pattern (BP). At the next level down, the nucleus of the preamble to part 6.1.1 is a reduced pattern of type RP1 (one reduction). Down one more level, the preamble to the body of part 6.1.1 comprises two reduced patterns of type RP5 and RP3 respectively, i.e. patterns having lost two or three elements compared with the basic pattern. There is therefore an apparent correlation between definition type and text structure. We went on to investigate this correlation for the whole of part 6.1. The results are presented in figure 4 in terms of occurrences of definition patterns in the numbered text parts. They show that the distribution suggested in figure 2 is a constant over the 8 sub-parts. The definitions, or rather definition nuclei – as the elaborations must be seen as part of the definitions – which initiate each sub-part (6.1 to 6.4) are all representatives of the basic pattern. One step below in the hierarchy, the definition nuclei which initiate parts 6.1.1 to 6.1.8 are mostly reduced patterns of type RP1 (6 out of 8), with one instance of basic pattern and one of reduced pattern RP4. In the parts which make up parts 6.1.1 to 6.1.8, the patterns showing multiple reductions dominate[7]:

Part 6			
	Part 6.1 : BP		
		Part 6.1.1 : RP1	
			Part 6.1.1(body): RP5
			RP3
		Part 6.1.2 : RP1	
			Part 32: RP3
			RP5
			RP5
			RP5
		Part 6.1.3 : BP	
		Part 6.1.4 : RP1	
		Part 6.1.5 : RP1	
		Part 6.1.6 : RP1	
		Part 6.1.7 : RP4	
		Part 6.1.8 : RP1	
	Part 6.2 : BP		
	Part 6.3 : BP		
	Part 6.4 : BP		

Figure 4. Distribution of definition patterns in part 6

This study attempts to relate a fine-grained analysis of a specific text object and the organisation of a large segment of text. The regularities in the distribution of definition patterns are of interest with respect to the dynamic aspect of text construction. Definitions in the corpus are seen as text objects which correspond to elaboration schemas whose nuclei are characterised by regular formatting patterns (lexico-syntactic, typographical and layout). Within these patterns, the classifier Nc states the class (what type of command it is) while the modifier (Vc VP) expresses the specificity. What the distribution of these patterns within the text as a whole shows is that the expression of the class can disappear at the lower hierarchical levels, when the classificatory elements have already appeared at structurally higher levels, leaving definitions entirely focussed on the functional aspects (what the command does). With each new part there is therefore an evolution from definitions which situate the command within the universe of the system to definitions which focus solely on what can be done with the command.

Conclusion

The above representations come out of a study starting from premises somewhat apart from most work on discourse organisation. The first is that there is a specific textual level of organisation which is signalled through what has been called "formatting". This textual level is seen as participating in the construction of textual meaning, in an interaction with other levels which has yet to be fully understood. The second is that formatting may be to some extent constrained by genre and domain, and that it therefore makes sense to identify generalisable traits within a genre/domain before going on to look for constants across genres/domains. The third is that it may be enlightening to focus on a specific text object, but view its behaviour within the text as a whole. This leads us to encompass a much larger text than is usually the case in detailed studies of discourse organisation, while adopting a fine-grained analysis for the text object under study.

Formatting as presented here provides a novel and theoretically-motivated way of envisaging the textual metafunction. It opens up the notion of discourse marker for written text, situating typographical and layout clues in a relation of functional equivalence with "classical" linguistic clues. Where there is congruence between RST and architectural segments, formatting markers are clues to discourse structure. The regular lexico-syntactic, layout and typographical patterns which we have called definition patterns have a dual status: they signal definitional text objects as well as being nuclei of a particular type of elaboration schema.

Whereas RST analysis is presented as essentially based on an interpretative process, fundamentally

[7] The detail of this level has only been given for parts 6.1.1 and 6.1.2 for readability's sake. The distribution is however constant throughout.

independent from any specific surface markers. the analysis of architecture centres on the signalling of textual objects through formatting. This paper has brought to light some convergence between the results of the two analyses in texts subject to high requirements of explicit signalling. This is a step towards understanding the linguistic resources brought into play for the signalling of discourse relations. Future work on these issues could take a number of distinct but potentially converging viewpoints: starting from special formatting devices. such as parentheses or footnotes; starting from specific text objects. to extend the study of definitions to other corpora or to examine other functional text objects such as examples or conclusions; taking particular relations as the starting point. to investigate relations which are reputed to have no marker – e.g. elaboration – in the light of the broader conception of signalling developed here.

Acknowledgements

This paper derives from work carried out over several years with Elsa Pascual. who died accidentally last summer. It is written in her memory.

I would like to thank A. Borillo and J. Rebeyrolle (ERSS. Toulouse). C. Garcia-Debanc. C. Luc and J. Virbel (PRESCOT. Toulouse). and D. Scott (ITRI. Brighton) for their comments and suggestions on earlier versions of this paper.

References

Bateman. J. and K.J. Rondhuis (1997). *Coherence Relations. Towards a General Specification.* Discourse Processes 24/1. pp. 3-50.

Grishman. R. and R. Kittredge (1986) *Analyzing language in restricted domains. Sublanguage description and processing.* Hillsdale. N.J.. Erlbaum.

Halliday. M. A. K. and R. Hasan (1976). *Cohesion in English.* London. Longman.

Halliday. M. A. K. (1985). *An Introduction to Functional Grammar.* London, Edward Arnold.

Harris. Z.S. (1968). *Mathematical Structures of Language.* New York, Wiley & Sons.

Harris. Z.S. (1982). *A grammar of English on mathematical principles.* New York. Wiley-Interscience.

Luc. C. (1998) *Relations constructives entre phrases élémentaires et relations rhétoriques dans le cadre de la génération automatique de textes à consignes.* Paper submitted to RECITAL'98. Le Mans. France. 8-9 September 98.

Maier. E. and E. Hovy (1993). *Organizing discourse structure relations using metafunctions.* In "New Concepts in Natural Language Generation". H. Horacek & M. Zock, ed.. Pinter, London.

Mann. W.C. and S.A. Thompson (1989). *Rhetorical structure theory: A theory of text organization.* In "The Structure of Discourse". L. Polanyi. ed., Ablex. Norwood, N.J.

Mann. W.C., Matthiessen, M.I.M. and S.A. Thompson (1992). *Rhetorical structure theory and text analysis.* In "Discourse Description: Diverse Linguistic Analyses of a Fund-raising Text", W.C. Mann & S.A. Thompson. eds.. John Benjamins, Amsterdam/Philadelphia. pp. 39-78.

Moore. J.D. and M.E. Pollack (1992). *A problem for RST: the need for multi-level discourse analysis.* Computational Linguistics, 18/4. pp. 537-544.

Pascual. E. (1991). *Représentation de l'architecture textuelle et génération de texte.* Doctoral thesis, Université Paul Sabatier, Toulouse.

Pascual. E. and M.-P. Péry-Woodley (1997a). *Définition et action dans les textes procéduraux.* In "Le texte procédural : langage. action et cognition", E. Pascual. J-L. Nespoulous & J. Virbel, eds., PRESCOT, Toulouse. pp. 223-248.

Pascual. E. and M.-P. Péry-Woodley (1997b). *Modélisation des définitions dans les textes à consignes.* In "Cognition. Discours procédural. Action". J. Virbel. J-M. Cellier & J-L. Nespoulous. ed.. PRESCOT. Toulouse. pp. 37-55.

Pascual. E. and M.-P. Péry-Woodley (1997c). *Modèles de texte pour la définition.* "Proceedings of 1ères Journées Scientifiques et Techniques du Réseau francophone de l'Ingénierie de la Langue de l'AUPELF-UREF", AUPELF-UREF, Avignon, pp. 137-146.

Péry-Woodley. M.-P. and Rebeyrolle. J. (1998) *Domain and genre in sublanguage text: definitional microtexts in three corpora.* "Proceedings First International Conference on Language Resources & Evaluation". Granada. May 1998. ELRA. Paris. pp. 987-992.

Sager. N.. C. Friedman. et al.. (Ed.) (1987). *Medical language processing. Computer management of narrative data.* Reading. MA., Addison-Wesley.

Virbel. J. (1985). *Langage et métalangage dans le texte du point de vue de l'édition en informatique textuelle.* Cahiers de Grammaire. 10. pp. 5-72.

Virbel. J. (1989). *The contribution of linguistic knowledge to the interpretation of text structures.* In "Structured Documents", J. André. V. Quint & R. K. Furuta. ed.. CUP, Cambridge. pp. 161-181.