Computational Approaches to Discourse and Document Processing

Marie-Paule Péry-Woodley* – Donia Scott**

**CLLE-ERSS*
Maison de la Recherche
Université de Toulouse-Le Mirail
31058 Toulouse Cédex
France

pery@univ-tlse2.fr

*** Department of Computing*
The Open University
Walton Hall
Milton Keynes MK7 6AA
UK

D.Scott@open.ac.uk

ABSTRACT. *This introduction tracks the evolution of the definition and role of discourse issues in NLP from the knowledge-intensive "discourse understanding" methods of the 80's to the recent concern with "accessing contents" in vast document bases via data-intensive methods. As text/discourse linguistics moves toward corpus approaches, also in connection with the development of large text bases and of computational instruments, we explore potential new forms of convergence.*

RÉSUMÉ. *Cette introduction trace l'évolution de la définition et du rôle des questions discursives en TAL, depuis la « compréhension du discours » des années 80, avec ses méthodes fondées sur la connaissance, jusqu'à l'actuel « accès au contenu » dans de vastes bases de documents à travers des méthodes faisant appel à de très gros volumes de données. La linguistique du texte et du discours se tournant de plus en plus vers l'approche corpus, en lien également avec le développement de bases de textes et d'instruments informatiques, de nouvelles formes de convergences se font jour.*

KEY WORDS: *discourse processing, document processing, discourse structures.*

MOTS-CLÉS: *traitement du discours, traitement du document, structures discursives.*

TAL. Volume 47 – n°2/2006, pages 7 à 19

1. Preamble

This is the first issue of TAL entirely and explicitly devoted to discourse. A number of recent issues have had themes which touched on discourse questions: Dialogue (2002), Automatic Summarization (2004), Anaphora Resolution (2005). In TAL's "big sister", *Computational Linguistics*, discourse is discreetly but, it seems, increasingly present: CL had its own issue on summarisation, featuring discourse-based approaches, in 2002, and since the beginning of the 21st century a steady discourse-related "thread" can be identified around topics such as automatic text segmentation, discourse relations, information structure and centering. Current NLP applications concerned with these topics range from the classic ones – generation, summarisation – to less obvious candidates such as information extraction or question answering, where an awareness of the desirability of taking into account the structured nature of texts is gradually appearing. A recent survey presented discourse research in an NLP context as focusing on two fundamental questions:

> "First, what information is contained in extended sequences of utterances that goes beyond the meaning of the individual utterances? Second, how does the context in which an utterance is used affect the meaning of the individual utterances or parts of them?" (Moore and Wiemer-Hastings, 2003:439)[1].

We will explain in this introduction how, by associating "discourse" and "document", we wish to set a somewhat broader scene, influenced by our perception of the evolving convergence between text/discourse linguistics, document processing and NLP.

2. Discourse and NLP: a changing relationship

2.1. From sentence processing to *discourse understanding*

One view of discourse processing is as an extension of sentence processing. In Allen's *Natural Language Understanding* for example (Allen, 1987) – to refer back to a classic textbook – one sees a clear bottom-up incremental approach starting with syntactic processing, then on to semantic interpretation, i.e. representing sentence meaning, and finally taking in context and world knowledge. The emphasis is on deriving a logical form, with quantifiers and case roles the major issues, which means that the scope is generally that of the syntactic clause or semantic proposition rather than sentences in all their naturalistic complexity. Along the same lines, Moore and Wiemer-Hastings state in their recent survey: "The juxtaposition of

[1] A definition borrowed from B. Grosz in the *Discourse and Dialogue* section of the *Survey of the State of the Art in Human Language Technology* (Cole *et al.,* 1998).

individual *clauses* may imply more than the meaning of the *clauses* themselves" (Moore and Wiemer-Hastings, 2003: 439, our emphasis). The next level focuses on context, and in this view it could be said that discourse is defined as "propositions in context". In accordance with Allen's roots in Artificial Intelligence, context is first envisaged as extra-textual, and a major place is given to knowledge representation and pragmatics. The discourse questions which he then raises when envisaging propositions in their textual context still constitute the object of much current research: segmentation, i.e. the grouping of "sentences addressing the same topic"; focus tracking; discourse relations; tense tracking in narratives. Allen defines his objective as describing "the basic techniques that are used in building computer models of natural language production and comprehension" (Allen, 1987:1). These objectives appear to have diversified with the coming of age of Natural Language Processing.

Moore and Wiemer-Hastings' 2003 survey gives a good indication of this evolution: the first part of their article, "*Computational theories of discourse structure and semantics*", is an extensive survey of major discourse models (DRT, Grosz and Sidner's Theory, RST, SDRT) followed by a lengthy section on "generating coherent discourse". The difference in treatment is remarkable between those initial sections and the one entitled "*Current discourse applications*" – covering summarisation and question answering –, which occupies just 4 pages of the 46-page review, explaining that "many of the current techniques break with the theoretical traditions described in previous sections. Instead they rely on shallow text processing techniques and statistical methods (…)" [Ibid:473]. What Moore and Wiemer-Hastings' article points to is the gap which has developed in the 1990's between computational models of discourse, characterised by knowledge-intensive methods and an underspecified "understanding" aim, and computational techniques making use of the vast increases in computing power and memory capacity, with much more specific objectives, such as text segmentation, discourse parsing, rhetorical parsing. In order to examine this gap, we find it useful to take a closer look at "discourse" vs. "document".

2.2. From *discourse understanding* to *accessing document contents*

The term "document" in relation to NLP – but not generally associated with "discourse" – has in recent years been given a marked impetus in the francophone context by the document engineering community *via* the CIDE cycle of conferences (Colloque International sur le Document Electronique)[2], the ongoing RTP-DOC research network[3] "Documents et contenu : création, indexation, navigation", the related "Semaines du Document Numérique". The idea for the 2006 "Symposium on Discourse and Document"[4], which announced and prepared this special issue, was

[2] Proceedings published by Europia : http://europia.org/edition/livres/doc/CatDoc.htm
[3] RTP CNRS 33 [RTP-DOC] : http://rtp-doc.enssib.fr/sommaire.php3
[4] Proceedings available on the symposium's website : http://discours2006.info.unicaen.fr/

rooted in a project experimenting with a number of approaches to discourse structure in a document browsing system aimed at geographers[5]. In most cases, the word "document" reflects a strongly situated and applied approach, as below in the paragraph opening the *Document Processing* section of the *Survey of the State of the Art in Human Language Technology*:

> "Work gets done through documents. When a negotiation draws to a close, a document is drawn up, an accord, a law, a contract, an agreement. When a new organization is established it is announced with a document. When research culminates, a document is created and published. And knowledge is transmitted through documents: research journals, text books and newspapers. Documents are information organized and presented for human understanding." (Cole *et al.,* 1998: 223).

From this angle, *discourse understanding* gets reworked as *accessing document contents*. An interesting reflection on this revised objective can be found in Nazarenko (2005), an examination of the implicit semantics underlying recent computational methods of access to textual contents. The evolution referred to earlier in relation to technical computing developments is complex: not only is "understanding" no longer approached via knowledge-intensive methods, it is no longer called "understanding" but "accessing textual content". And it seems to have exploded into a diverse range of document-processing applications, with increasingly fuzzy boundaries: information extraction – which can be seen as one of the forms taken by "understanding"[6] – blends into question answering, or query-biased summarisation, or browsing aids.

Examining the difference between "contents" and "meaning", Nazarenko (2005) points to the vagueness of the term "contents", and to its application-bound character: the aim may be to identify specific elements of information, or to get an overall view of what the document is about. The author stresses the difference between asking "What is this document about?" and "What does this document say?". She surveys the related applications mentioned above, concerned with the latter question, and sees a very limited notion of "contents", revolving largely around named entities. She uses the term "scattered semantics[7]" to describe the way documents are approached from different unrelated angles, and as separate "text islands" which never become interconnected. Approaches are eclectic and pragmatic, which means that a broader range of "markers" are being considered (such as typographical properties, document structure: sections, titles); they cannot however lead to an integrated semantic representation, for lack of a model which

---

[5] GEOSEM – Traitements sémantiques pour l'Information Géographique : textes, cartes, graphiques. Projet du programme interdisciplinaire du CNRS "Société de l'Information". http://infodoc.unicaen.fr/geosem/

[6] See in the aforementioned Survey of the State of the Art in HLT (Cole *et al.*, 1998) the title of section 7.3: "Text *Interpretation*: Extracting Information" (our emphasis).

[7] Our rendering of "sémantique éclatée".

would "make it possible to understand the role and contribution of each [level of analysis] in the overall analysis result, and to define on this basis an architecture for semantic analysis" (Ibid.: 225). A similar position is found in the recent posting for the *Workshop on Semantic Content Acquisition and Representation*[8]:

> "Text (and language in general) has ABOUTNESS; it has meaning, or semantic content. We as (computational) linguists are highly adept at dissecting text on a number of different levels: we can perform grammatical analysis of the words in the text, we can detect animacy and salience, we can do syntactic analysis and build parse trees of partial and whole sentences, and we can even identify and track topics throughout the text. However, we are comparatively inept when it comes to identifying the semantic content, or meaning, of the text."

The term ABOUTNESS – typographically emphasised in the opening sentence of the posting – is a case in point. *Aboutness* is one of the defining properties of *topic*, a notion which, though central to text/discourse linguistics, is very difficult to handle beyond the propositional level. Identifying and tracking the topics developed in documents is understandably a major concern for many applications (see Ferret, this issue; see also the TDT (*Topic Detection and Tracking*) evaluations held since 1998[9]). But defining discourse topic (theoretically and operationally) is an area where the gap between modelling and data intensive approaches is at its widest: van Dijk's influential view of discourse topic as a semantic macro-structure, or the construction of discourse topic as a Discourse Representation Structure in SDRT – at once based on and informing the construction of representations for successive propositions – are a very long way from the lexical repetition approach to automatic text segmentation into topically cohesive sections. The link between the two is the hypothesis that data-intensive methods pick up on lexical epiphenomena of topic organisation, but this is not an easy hypothesis to validate at this stage.

So at one end of the spectrum we have *discourse understanding*, strongly theory-based but without much impact on most current applied NLP work, at the other end, bitty *access to contents*, which computes or extracts disparate types of textual information it cannot integrate. The association between discourse and document in the title of this special issue reflects a desire to confront this tension. The two families of approaches delineated above also differ in their attitude to data: while a minimalist approach is traditional in discourse understanding research, which mostly works from short constructed examples, data-intensive approaches are prominent in many recent applications, which rely on vast repositories of "naturally occurring" text. These differences are explored in the next two sections.

---

[8] SCAR, to be held in May 2007: http://www.sics.se/~mange/scar2007/

[9] National Institute of Standards and Technology: http://www.nist.gov/speech/tests/tdt/

2.3. Discourse *and* document; discourse *as* document

The tension referred to above may be reformulated – in a somewhat rough and ready manner – in terms of local *vs*. global coherence or micro- *vs*. macro-structure. Psycholinguistic studies of discourse processing insist on the interaction between microstructures and macrostructure in the dynamic construction of a coherent interpretation. Macrostructures are constructed partly[10] on the basis of the processing of microstructures, and in turn inform this processing. This presents NLP with a major difficulty as macrostructures are not manifested linguistically. Among the indicators of macropropositions that can be identified are "titles of text, summaries, and topical sentences, often at the beginning or end of a paragraph" (Louwerse and Graesser, 2005). These are indeed some of the features listed by Nazarenko (2005) as new types of information exploited in applications such as summarisation and information extraction, but of little concern to most research in computational discourse semantics.

A different – and more linguistic – take on the notion of document finds its roots in functionalist approaches, with their emphasis on communication as the primary function of language, (and as what actually *shapes* the forms languages take), and on the role of external (cognitive and sociocultural) factors in explaining linguistic phenomena. Two major tenets of Halliday's systemic functional linguistics of relevance here are:

- the *text*[11], defined functionally as "a unit of language in use", is "the unit of the semantic process":

> "A text, as we are interpreting it, is a semantic unit, which is not composed of sentences but realized in sentences." (Halliday, 1977-2003: 45-46)[12]:

- choices of wording (at different granularity levels) are seen as resulting from choices within systems corresponding to three components (called metafunctions): interpersonal (interaction between speaker/writer and addressee); ideational ("construing" experience: participants, processes and circumstances); textual (presentation of ideational and interpersonal meaning as information in text unfolding in context: theme, cohesion).

Halliday stresses the way these components work *together* to form cohesive text ("texture"):

> "Texture is not something that is achieved by superimposing an appropriate text-form on a pre-existing ideational content. The textual

[10] Readers' goals and world knowledge – including textual knowledge such as knowledge of genre and layout conventions – also play a major role.
[11] Halliday's terminology includes neither *discourse* nor *document*, but *text,* and *texture* for the properties of cohesion-coherence which characterise it.
[12] Page numbers are as in the 2003 re-edition.

> component is a component of meaning along with the ideational and interpersonal components. Hence a linguistic description is not a progressive specification of a set of structures one after the other, ideational, then interpersonal, then textual. The system does not first generate a representation of reality, then encode it as a speech act, and finally recode it as a text, as the metaphors of philosophical linguistics seem to imply. It embodies all these types of meaning in simultaneous networks of options, from each of which derive structures that are mapped on to one another in the course of their lexicogrammatical realization." (Halliday, 1977-2003:45).

This functional view informs our interest in discourse *as* document and not just as a linguistic object beyond the sentence. It entails a multi-dimensional view of discourse level phenomena: lexical cohesion for instance (apprehended e.g. via lexical chains or through text-tiling type methods) needs to be considered in its relations with other structuring principles. It has inspired much research on computational approaches to discourse: around Rhetorical Structure Theory (Mann and Thompson, 1988), whether focussing on generation or discourse parsing (Marcu, 2001; 2005); and in direct connection with the Hallidayan notion of *cohesion*, in the fertile field of automatic segmentation based on lexical cohesion[13]. With more specific relevance to our argument, it also constitutes a major source of inspiration (along with Nunberg's *Linguistics of Punctuation*, 1990) for researchers attempting to theorise document structure in order to integrate a broad conception of layout into the interpretation and generation of written texts (Bateman *et al*., 2001; Delin *et al.*, 2002; Power *et al*, 2003).

Envisaging layout in terms of an *abstract document structure* (Power *et al.*, 2003), and as an integral part of discourse construction in relation with other structuring principles, such as rhetorical structure, can be seen as an extension of the investigation of the textual component. A related though independent approach was developed around the notion of *Text Architecture* (Virbel, 1985; Luc and Virbel, 2001), insisting on the relations (including equivalence) between discursive and visual realisations of particular *text objects*, and on the semantic import of layout choices within the abstract architecture of a text. Such research offers prospects for a theoretically-motivated integration of the scattered approach to "contents" described in Nazarenko (2005).

2.4. Documents as discourse: the "corpus linguistics effect"

This widening gap between computational theories of discourse on the one hand and discourse processing techniques on the other, pointed out in their different ways by Moore and Wiemer-Hastings (2003) and Nazarenko (2005), is obviously linked to the growing availability of vast volumes of textual data and to the massive increase in processing power. These developments have led to the spread of data-

[13] The initiators of these segmentation techniques (Morris and Hirst, 1991; Hearst, 1997 *inter alia*) systematically refer to *Cohesion in English* (Halliday and Hasan, 1976).

intensive techniques in NLP, but have also meant that empirical linguistics has acquired tools to match its aims. Large-scale corpus-based studies of discourse structure are however still rather infrequent, as their requirements both in terms of corpus collection and preparation, and in terms of computational tools, are heavier than for other fields of linguistics (see Péry-Woodley, 2005 for an analysis). Yet they may hold a major key for a rapprochement between the two poles. Empirical approaches are necessary for the validation of findings from theoretical or small-scale descriptive approaches, as the tendency in text/discourse linguistics has been to use data illustratively, with very few instances, and limited concern for replicability. Systematic corpus-based studies are also an essential step in a focused collective reflection on categories and methods, which in turn should lead to better tools for discourse annotation (interactive tools for automatic or semi-automatic annotation, cf. Orasan, 2005; Webber and Byron, 2004), and form the basis for cumulative research building on previous results.

The journal *Computational Linguistics* devoted a special issue to "Empirical Studies in Discourse" ten years ago, with a useful introduction (Walker and Moore, 1997) listing a number of ways in which empirical methods can "help researchers discover general features and generate hypotheses", among which:

> "(1) Tagging of discourse phenomena in corpora; (2) Induction of algorithms or discourse models from tagged data; (3) Comparison of algorithm output to human performance; (4) Human scoring of an algorithm's output; (5) Task efficacy evaluation based on the domain; (6) Ablation studies where algorithm features are systematically turned off." (Walker and Moore, 1997: 1)

Discourse annotation projects have progressed apace in these last ten years, mostly for English but also for French, among which Discourse Treebanks (Penn Discourse Treebank, Miltsakaki *et al.*, 2004; Prasad *et al.*, this issue; RST Discourse Treebank, Carlson *et al.*, 2002) and anaphoric annotation projects (in Britain at UCREL[14], in France with the ANANAS project[15], or the "Modern French Corpus including Anaphors Tagging" (Tutin, 2002)). These projects mostly involve annotation by hand, though they make use of pre-processing techniques, and of course aim to provide tagged data usable for a variety of NLP applications. A different empirical approach to discourse aims to test linguistic hypotheses – e.g. on the semantics of discourse markers – on large corpora, calling upon NLP techniques to do so (Bestgen, 2006; Piérard and Bestgen, this issue). Finally, extensive corpus-studies form the necessary basis to the constitution of lists of markers to be used in the automatic identification of specific discourse structures (see Jackiewicz, also Couto and Minel, this issue).

The "document" element of our title fits at once easily and uneasily with current empirical approaches. Easily because the term implies attested texts, produced in

[14] http://www.comp.lancs.ac.uk/computing/research/ucrel/annotation.html
[15] http://www.atilf.fr/ananas/

real situations, and that is the basic tenet of corpus linguistics. Uneasily because document-level approaches a) exclude the sampling techniques used in many corpora; b) are particularly interested in documents with complex document structure; c) require an awareness of and an approach to genre. As part of the cumulative process we call for, there are prospects for “second-generation” research which will use “first-generation” annotated corpora to track the document-level patterns formed by discourse phenomena, or their interactions with document structure.

3. Overview of the issue

Taken together, the articles in this Special Issue evidence the closer integration of the notions of “discourse” and “document” that is now emerging in the field. Their impetus comes not just from the need to fill theoretical gaps but from an interesting and telling range of technology-led sources that force us to address new forms of texts and to support new text-centred applications. Computational linguistics has long sought "killer applications", and for many the closest we are to realising this in the current context is through information extraction, question-answering, and automatic summarisation systems. However, discourse has to-date not been a major feature of the work in these areas. This situation is beginning to change. We see several examples of this in this Special Issue.

Taking Question-Answering as their objective, *Verberne* and her colleagues, explore the extent to which properties of a discourse can be used to guide the identification in documents of answers to one of the least studied, and least successful, types of questions put to such systems: "Why". Although their work is still in its early stages, the results of their explorations have been extremely promising. *Prasad* and her colleagues describe an annotation scheme they have developed for the attribution of the arguments of discourse relations, and indeed for the relations themselves. By annotating the Penn Discourse Treebank with information about the source and degree of factuality of the propositions, facts and eventualities contained in a discourse, they aim to get systems to learn to recognise these properties in text. *Jackiewicz* is also concerned with the multiple layers of discourse within a text, and with ways of taking this complexity into account in NLP and Semantic Web applications: her corpus-based study of reported speech (direct quotations) in newspaper articles leads to a semantic typology which distinguishes four major types of quotation practices marked by increasing writer involvement. Methodologies differ however: in *Prasad et al.*, systematic (manual) corpus annotation is both a way of refining a model and constructing a resource for NLP; *Jackiewicz’s* method relies on extracting relevant contexts from the corpus on the basis of pre-existing lists of reporting verbs used as seeds. These text extracts are then analysed with the objective of producing a typology and lists of surface markers for the automatic identification of the different types of structures.

The imperative for intelligent search has inspired renewed efforts on document segmentation, and we present two examples of this. *Pierard and Bestgen* present a large-scale corpus-linguistics methodology to evaluate the segmentation function of different temporal adverbials, calling upon two indices of thematic continuity/discontinuity which are amenable to being studied in large volumes of text: paragraph breaks and an index of lexical cohesion resulting from latent semantic analysis. In the context of text segmentation techniques based on lexical reiteration, *Ferret* explores and evaluates two possible ways of improving the detection of similarities between discourse units: an endogenous method based on non-supervised techniques of topic discovery; and an exogenous method relying on a network of lexical co-occurrences built from a large corpus.

Undoubtedly, the emergence of the WWW has been an important catalyst for research in computational linguistics. However, the imperative of dealing with the immense quantity of text that this technology produces has largely overshadowed the need to address an equally obvious fact: the web has led to a new textual genre of hypertextual documents. The non-linearity of hypertext, and its encapsulation as a digital, screen-based form, presents new challenges for readers and writers alike. *Mancini, Scott and Buckingham-Shum* discuss the problem of signalling discourse structure in hypertextual documents, and propose a solution that maps linguistic markers of discourse relations (ie., subordinating and coordinating conjunctions, and conjunctive adverbs) onto dynamic, visual properties. Their contribution provides a significant departure from our usual conceptualisations of discourse markers as strictly linguistic objects.

Framing these works on either side of the theory/application divide are contributions by *Danlos* and *Max*, both concerned with issues of well-formedness and coherence. The prevailing discourse theories are scrutinized by *Danlos*, who examines their potential for strong generative capacity: allowing all and only well-formed structures that correspond to felicitous discourses. *Max* explores aspects of the class of natural language tools that make use of *symbolic authoring*[16] to allow subject-matter experts to edit the underlying meaning of a document, and describes a particular instantiation of such a tool. Finally, *Couto and Minel*'s contribution approaches the theory/application divide in an original way, taking document browsing as a starting point to propose a declarative language designed to represent the knowledge a reader uses in the course of browsing, i.e. an application-specific way of approaching aspects of discourse organisation. Along the way, the authors raise questions about the data structures needed to represent texts and about the nature and respective status of different text objects.

[16] "Symbolic authoring" was introduced by Scott and Evans (1998) to describe a technique for generating language-neutral symbolic representations of the content of a document.

References

Allen, J., *Natural Language Understanding*, Menlo Park, CA, Benjamin Cummings, 1987.

Bateman, J., Kamps, T., Kleinz, J., Reichenberger, K., “Towards Constructive Text, Diagram, and Layout Generation for Information Presentation”, *Computational Linguistics*, vol. 27 no. 3, 2001, p. 409-449.

Bestgen, Y., "Improving Text Segmentation Using Latent Semantic Analysis: A Reanalysis of Choi, Wiemer-Hastings, and Moore", *Computational Linguistics*, vol. 32 no. 1, 2006, p. 5-12.

Carlson, L., Marcu, D., Okurowski, M. E., "Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory", in J. van Kuppevelt, R. Smith (Eds.), *Current Directions in Discourse and Dialogue*, Dordrecht, Kluwer Academic Publishers, 2002, p. 85-109.

Cole, R. A., Mariani, J., Uszkoreit, H., Varile, G. B. (Eds.). *Survey of the State of the Art in Human Language Technology*, Pisa, Giardini, 1998 (Electronic version available on: http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html).

Delin, J., Bateman, J., Allen, P., “A Model of Genre in Document Layout”, *Information Design Journal*, vol. 11 no. 1, 2002, p. 54-66.

Halliday, M. A. K., “Text as semantic choice in social contexts”, in J. Webster (Ed.), *The Collected Works of M.A.K. Halliday (Volume 2): Linguistic Studies of Text and Discourse,* London, Continuum, 2003, p. 23-81 (reprinted from van Dijk, T., Petöfi, J.S. (Eds.), *Grammars and Descriptions*, Berlin, Walter de Gruyter, 1977, p.176-226).

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

Hearst, M., "TextTiling: segmenting text into multi-paragraph subtopic passages", *Computational Linguistics*, vol. 23 no. 1, 1997, p. 33-64.

Louwerse, M. M., and Graesser, A. C., "Macrostructure", In K. Brown (Ed.), *Encyclopedia of Language and Linguistics*, Elsevier, 2005, Vol. 7.

Luc, C., and Virbel, J., "Le modèle d'architecture textuelle : fondements et expérimentation", *Verbum*, vol. 23 no. 1, 2001, p. 103-123.

Mann, W. C., Thompson, S. A., “Rhetorical structure theory: Toward a functional theory of text organization”, *Text*, vol. 8 no. 3, 1988, p. 243-281.

Marcu, D., *The Theory and Practice of Discourse Parsing and Summarization,* Cambridge, MA, MIT Press, 2001.

Marcu, D., “Discourse Parsing, Automatic”, in K. Brown (Ed.), *Encyclopedia of Language and Linguistics*, Elsevier, 2005, Vol.3.

Miltsakaki, E., Prasad, R., Joshi, A., Webber, B., "Annotating Discourse Connectives and their Arguments", *HLT/NAACL Workshop on Frontiers in Corpus Annotation*, Boston, MA. 2004.

Moore, J. D., Wiemer-Hastings, P., “Discourse in Computational Linguistics and Artificial Intelligence”, In A. C. Graesser, M. A. Gernsbacher, S. R. Goldman (Eds.), *Handbook of Discourse Processes*, London, Lawrence Erlbaum, 2003, p. 439-485.

Morris, J., Hirst, G., "Lexical cohesion computed by thesaural relations as an indicator of the structure of text", *Computational Linguistics*, vol. 17 no. 1, 1991, p. 21-48.

Nazarenko, A., “Sur quelle sémantique reposent les méthodes automatiques d'accès au contenu textuel ? ”, in A. Condamines (Ed.), *Sémantique et Corpus*, Paris, Hermès, 2005, p. 211-244.

Nunberg, G., *The Linguistics of Punctuation,* Menlo Park, CSLI, 1990.

Orasan, C., “Automatic annotation of Corpora for Text Summarisation: A Comparative Study”, In *Proceedings of 6th International Conference, CICLing2005*, Mexico City, 2005, Berlin, Springer-Verlag, p. 670-681

Péry-Woodley, M.-P., "Discours, corpus, traitements automatiques", in A. Condamines (Ed.), *Sémantique et Corpus*, Paris, Hermès, 2005, p. 177-210.

Power, R., Scott, D., Bouayad-Agha, N., “Document Structure”, *Computational Linguistics*, vol. 29 no. 2, 2003, p. 211-260.

Scott, D. and Evans, R. “Multilingual Document Management Without Translation: Using natural language generation in the Multilingual Information Society”. *Elsnews*, vol. 7 no. 1, February 1998.

Tutin, A., "A Corpus-based Study of Pronominal Anaphors in French", *DAARC 2002*, Lisbonne, Portugal, 2002.

Virbel, J., “Langage et métalangage dans le texte du point de vue de l'édition en informatique textuelle”, *Cahiers de Grammaire*, vol. 10, 1985, p. 5-72.

Walker, M., Moore, J. M., "Empirical studies in discourse", *Computational Linguistics*, vol. 23 no. 1, 1997, p. 1-12.

Webber, B., Byron, D. K., "Discourse Annotation", *ACL 2004 Workshop on Discourse Annotation*, Barcelona, Spain, 2004.

Comité de lecture spécifique :

N. Asher (CNRS, U. Toulouse 3, France)
J. Bateman (U. Bremen, Germany)
Y. Bestgen (U. C. Louvain, Belgium)
N. Bouayad-Agha (U. Pompeu Fabra, Barcelona, Spain)
M. Charolles (U. Paris 3, France)
N. Colineau (CSIRO, Australia)
D. Cristea (U. Iasi, Romania)
L. Danlos (U. Paris 7, France)
L. Degand (U. C. Louvain, Belgium)
P. Enjalbert (U. Caen, France)
S. Ferrari (U. Caen, France)
B. Grau (U. Paris-Sud, France)
C. Hallett (Open University, U.K.)
A. Hartley (U. Leeds, U.K.)
N. Hernandez (U. Nantes, France)
J. Karlgren (Swedish Institute of Computer Science, Sweden)
L. Kosseim (U. Concordia, Quebec, Canada)
G. Lapalme (U. Montréal, Quebec, Canada)
H. Le Thanh (Hanoi University of Technology, Vietnam)
N. Lucas (CNRS, U. Caen, France)
C. Mancini (Open University, U.K.)
A. Max (U. Paris-Sud, France)
J.-L. Minel (CNRS, U. Paris 10, France)
C. Paris (CSIRO, Australia)
R. Power (Open University, U.K.)
H. Saggion (U. Sheffield, U.K.)
S. Teufel (U. Cambridge, U.K.)
K. van Deemter (U. Aberdeen, U.K.)