Journée d'étude de l'ATALA
La Rochelle, 22 juin 2004
dans le cadre de la Semaine du Document Numérique

APPEL A COMMUNICATION

Modéliser et décrire l'organisation discursive à l'heure du document numérique

CALL FOR PAPERS

Modelling and describing discourse organisation in the age of the digital document

Journée organisée par/workshop organised by
Marie-Paule Péry-Woodley,
ERSS/Université de Toulouse-Le Mirail (pery@univ-tlse2.fr)

[English text below]

La Semaine du Document Numérique a pour objectif de réunir des communautés de recherche traitant du document numérique à partir de divers points de vue : média, modes de médiation (techniques et sociaux), relation avec l'activité humaine. La journée d'étude de l'ATALA se propose d'aborder ces questions sous l'angle linguistique en mettant en avant le fait que les documents numériques sont des discours, présentant une organisation interne qui demande à être comprise et qui peut être utilisée dans des systèmes informatiques. Cette journée d'étude vise à mettre en présence trois thématiques concernées par le développement du document numérique : l'étude de l'organisation discursive, les linguistiques de corpus, les applications informatiques visant l'exploitation de documents.

Pour les linguistiques du texte et du discours, l'essor du document numérique entraîne de nouvelles possibilités et de nouvelles interrogations, parmi lesquelles :

- l'application au discours des méthodes d'analyse de corpus : comment constituer des observables pertinents dans ce domaine ?
- le développement de nouveaux modes d'accès aux documents, mettant l'accent sur l'exploitation de la structuration interne des documents et de sa signalisation ;
- l'impact des nouveaux types de documents sur les notions fondamentales du domaine : cohésion, cohérence, signalisation métadiscursive.

Cette journée d'étude sur l'organisation discursive à l'écrit a pour but de rassembler des travaux issus de trois champs qu'il semble important de rapprocher à la lumière de ces nouvelles perspectives :

1. Organisation discursive

Pour appréhender une suite d'énoncés comme discours, il faut en saisir l'organisation (percevoir les segments, leur hiérarchie, leurs mises en relation). Une longue et fertile tradition aborde l'étude de cette organisation à travers la notion de relations de discours, liens sémantico-pragmatiques entre segments (propositions ou groupes de propositions) (cf. Péry-Woodley (coord.) 2001). D'autres modes d'organisation sont envisagés, notamment par le biais de la notion de thème ou topique, ou plus récemment à travers la notion d'encadrement du discours (Charolles 1997).
Les travaux dans ce domaine se situent sur un continuum allant de la modélisation "conceptuelle" à des approches empiriques (segmentation automatique, cf. Hearst 1997 ; analyses surfaciques – manuelles ou automatiques - cf. Teufel et Moens 1999). Le défi est de tenir les deux extrémités du continuum pour faire le lien entre les réalisations dans les textes et les processus qui sous-tendent fondamentalement l'organisation discursive à différents niveaux de grain (organisation locale vs. organisation globale). La relation entre modélisation et études empiriques est à présent souvent problématique, ces dernières courant le risque de perdre de vue la structure en se focalisant sur les marqueurs (connecteurs par exemple), alors que les

modèles conceptuels sont difficiles à tester empiriquement. Les approches sur corpus – facilitées par la prolifération des documents numériques – sont en train, là aussi, de changer la donne (cf. Conrad 2002).

2. Études en corpus des corrélats linguistiques de l'organisation discursive

Comme l'ont noté plusieurs auteurs (Biber et al 1998 inter alia), bien que les travaux sur l'organisation discursive soient presque toujours fondés sur l'analyse de textes attestés, le corpus y joue plus souvent le rôle de réservoir d'exemples que d'objet d'analyse à proprement parler. Il faut souligner les difficultés de la mise en œuvre d'une "approche corpus" dans le domaine de l'organisation discursive : tant pour la constitution du corpus (les techniques fondées sur l'échantillonnage l'excluent…), qu'en ce qui concerne le rôle de la quantification, et surtout la définition d'observables pertinents, à même de permettre la mise en relation entre marqueurs de surface (qui peuvent n'être que des épiphénomènes) et principes d'organisation sous-jacents (multiples et interconnectés).
Par ailleurs, on note dans les approches empiriques du discours un clivage entre approches linguistiques (couverture faible et fiabilité forte) et approches numériques (couverture forte et fiabilité faible). L'articulation entre ces approches ouvre des perspectives d'avenir, à la fois pour la compréhension des principes d'organisation discursive et pour les applications.

3. Applications informatiques visant l'exploitation de documents numériques

Les applications qui ont pour unité pertinente le document sont peu concernées par la question de l'organisation discursive. Celles, en revanche, qui visent la navigation intra-documentaire, la synthèse sélective, la visualisation multi-échelle, doivent pénétrer dans les documents et ont donc intérêt à ne pas les envisager comme de simples « sacs de mots », mais à prendre en compte la structuration en « blocs » thématiques et/ou rhétoriques, ainsi que l'architecture du texte (cf. Luc et Virbel 2001). Ces visées font émerger de nouvelles questions, telles que l'articulation de niveaux d'organisation dans les documents longs (où l'aide à la navigation acquiert une pertinence particulière).

Cet appel concerne à la fois des travaux qui se réclament des interactions entre ces domaines, et des travaux qui se situent dans l'un d'eux mais dont les auteurs sont intéressés par le dialogue proposé. Les études descriptives faisant la part belle à la réflexion méthodologique sont particulièrement sollicitées. Quelques thématiques pertinentes (liste non limitative) :

- Repérage d'objets ou de zones de texte correspondant à des actes textuels ou discursifs (conclusions, explications, évaluations, …)
- Marqueurs d'organisation discursive (des marqueurs aux relations, approche inductive) : connexion, indexation (cadres), métadiscours textuel
- Caractérisation linguistiques de fonctions discursives (des fonctions aux marqueurs : approche déductive)
- Segmentation (automatique ou manuelle) : "topic shifts", indices (lexico-syntaxiques, typo-dispositionnels) de bornes de segments
- Articulation entre organisation locale et globale
- Impact du genre discursif sur l'organisation et sa signalisation linguistique
- Analyse et exploitation de l'architecture des documents
- Approches topologiques
- Annotation discursive

SOUMISSION (MODALITES)

Un résumé de deux à quatre pages doit être envoyé avant le 30 janvier 2004
par courrier électronique en format word, pdf ou ps à Marie-Paule Péry-Woodley (<pery@univ-tlse2.fr>).

Les notifications d'acceptation seront données pour le 15 mars 2004.

CALL FOR PAPERS

Modelling and describing discourse organisation in the age of the digital document

The Digital Document Week aims to gather research communities dealing with digital documents from a variety of angles: media, technical and social modes of mediation, relation with human activity. This ATALA workshop wishes to broach these questions from a linguistic point of view, focussing on digital documents as discourse, characterised by an internal organisation which needs to be understood and may be exploited in computer-based systems. The workshop aims to bring together three research areas concerned with the development of digital documents: the study of discourse organisation, corpus linguistics, computer-based applications for the exploitation of digital documents.

For text and discourse linguistics, the proliferation of digital documents leads to new opportunities and new research questions, such as:
- the application of corpus analysis methods to discourse: what kind of data can be regarded as relevant at this level of linguistic investigation?
- the development of novel ways of accessing documents, which leads to a new emphasis on text structure and the potential exploitation of surface markers;
- the impact of new document types on basic concepts in the field: cohesion, coherence, metadiscursive signalling.

This workshop on written discourse organisation aims to bring together research from three domains which must seek points of convergence in the light of these new prospects:

1. Discourse organisation

In order to apprehend a sequence of utterances as discourse, it is necessary to understand its organisation (to identify its segments and perceive their hierarchy and their relations). An old and fertile tradition approaches discourse organisation via the notion of discourse relations: semantico-pragmatic links between segments (propositions or sets of propositions) (cf. Péry-Woodley (ed) 2001). Other modes of organisation may be envisaged, via the notion of theme or topic for instance, or more recently through the discourse framing hypothesis (Charolles 1997). Research in this field can be placed in a continuum from pure "conceptual" modelling to empirical methods (automatic segmenting, cf. Hearst 1997; shallow analyses – human or automatic - cf. Teufel et Moens 1999). The challenge is to hold both ends of the continuum in order to draw connections between the way "things are put" in texts and the processes underlying discourse organisation at different levels of granularity (local vs. global organisation). The relationship between modelling approaches and empirical research has often seemed problematic, with empirical studies running the risk of losing track of structure as they focus on surface markers, while conceptual models tend to be difficult to test empirically. Corpus-based approaches – greatly facilitated by progression into the digital age – are in the process of bringing considerable changes in the discourse field, as they have done elsewhere in linguistics (Conrad 2002).

2. Corpus-based studies of linguistic correlates of discourse organisation

As noted by several authors (Biber et al 1998 inter alia), though research on discourse organisation tends to make regular use of authentic data, the corpus is often seen as a source of examples rather than the object of the analysis as such. The implementation of a fully-fledged "corpus approach" in the field of discourse

organisation carries with it many difficulties: corpus construction (common sampling-based techniques make it impossible…), the role of quantitative analysis, and most of all definition of relevant data making it possible to draw the connection between surface markers (which may be just epiphenomena) and the multiple principles underlying complex hierarchic organisation.
A gap can also be observed between linguistic approaches (low coverage and high reliability) and numerical approaches (high coverage and low reliability). Articulating these approaches may open new prospects, leading to fresh insights into discourse organisation principles as well as more operational methods for applications.

3. Computer-based systems for the exploitation of digital documents

Applications for which the relevant unit is the whole document are little concerned by questions of discourse organisation, but those concerned with intra-document browsing, selective synthesis or multi-level visualisation must work their way inside the documents and therefore cannot consider them as simple "bags of words": they have to take into account the organisation into thematic or rhetorical chunks and text architecture (cf. Luc & Virbel 2001). These objectives bring about new research questions, in particular around the articulation of different organisational levels in long documents (where browsing aids acquire particular relevance).

This call for papers concerns researchers who are already working on these interactions, as well as those whose work is in one of the domains referred to but who are interested in a dialogue with other discourse approaches. Descriptive studies which pay specific attention to methodology will be particularly welcome.

Some relevant themes (non-exhaustive list):

- identification of objects or text zones corresponding to text or discourse acts (conclusions, explanations, evaluations, …)
- discourse organisation markers (from markers to relations: inductive approach): connection, indexing (discourse frames), textual metadiscourse
- linguistic characterisation of discourse functions (from functions to markers: deductive approach)
- segmentation (automatic or manual): "topic shifts", clues to segment boundaries (lexico-syntactic, typographical, dispositional)
- articulation between local and global organisation
- impact of discourse genre on discourse organisation and its linguistic markers
- analysis and exploitation of document architecture
- topological approaches
- discourse annotation

SUBMISSION (MODALITIES)

A summary (2-4 pages, Word, pdf or ps) to be e-mailed before January 30th 2004 to Marie-Paule Péry-Woodley (<pery@univ-tlse2.fr>).

Notification of acceptance will be given by March 15th 2004.

Références/References

Biber, D., Conrad, S., & Reppen, R. (1998). Corpus linguistics: Investigating language structure and use. Cambridge: Cambridge University Press.

Conrad, S. (2002). Corpus linguistics approaches for discourse analysis. Annual Review of Applied Linguistics, 22, 75-95.

Charolles, M. (1997). L'encadrement du discours : Univers, champs, domaines et espaces (Cahier de Recherche Linguistique 6): Université de Nancy2.

Hearst, M. (1997). TextTiling: segmenting text into multi-paragraph subtopic passages. Computational Linguistics, 23(1), 33-64.

Luc, C., & Virbel, J. (2001). Le modèle d'architecture textuelle : fondements et expérimentation. Verbum, 23(1), 103-123.

Péry-Woodley, M.-P. (ed.) (2001). Cohérence et relations de discours à l'écrit. Présentation. Verbum, 23(1).

Teufel S. & Moens, M. (1999). Discourse-level argumentation in scientific articles: human and automatic annotation. In: Towards Standards and Tools for Discourse Tagging. ACL 1999 Workshop.