RECITAL 2004, Fès, 21 avril 2004

Analyse macro-sémantique: vers une analyse rhétorique du discours

Antoine Widlöcher
GREYC - Université de Caen
awidloch@info.unicaen.fr

Résumé - Abstract

S'inscrivant dans les domaines du TAL, de la linguistique sur corpus et de l'informatique documentaire, l'étude présentée ici opère plus précisément dans la perspective d'une analyse macro-sémantique de la structuration discursive. Plus spécifiquement, nous proposons une analyse sémantique des structures rhétoriques du discours. Après avoir envisagé certaines voies ouvertes en la matière, nous définissons notre approche, et présentons les expérimentations conduites, dans le cadre du projet GeoSem, sur les structures énumératives dans le domaine géographique.

Within the frameworks of automatic NLP, corpus linguistics, and automatic document processing, this study aims more precisely at a macro-semantic examination of the discursive structuring. Specifically, a semantic study of the rhetorical structures of discourse is here suggested. After considering existing approaches, we present our own methodology, and relate experimentations, as part of the GeoSem project, on enumerative structures in geographical corpus.

Mots-clefs – Keywords

Macro-sémantique, analyse rhétorique, structure du discours, extraction d'information
Macro-semantics, rhetorical analysis, discourse structuring, information retrieval

1 Vers une analyse rhétorique

Dans les domaines du TAL, de la linguistique sur corpus et de l'informatique documentaire, cette étude vise plus précisément l'analyse *macro-sémantique* des *structures rhétoriques* du *discours*, c'est-à-dire des modes d'organisation du discours considéré à un niveau de granularité élevé. Sur l'aspect sémantique, disons simplement que, loin de nous en tenir à une caractérisation des formes de surface du discours, nous visons au contraire la détermination du sens qui s'y propage, de la *signification* engagée dans telle ou telle organisation *formelle* de surface. Nous prétendons du reste adopter un point de vue *macro*-sémantique. Si l'analyse sémantique au niveau lexical ou infra-lexical nous donne accès à des *atomes de signification*, et si l'analyse au niveau syntagmatique, propositionnel ou phrastique délivre des *molécules de sens* caractérisées

par une relative divisibilité et par une certaine *unité de sens*, pour notre part, nous observons le discours à un niveau de granularité supérieur, en étudiant l'organisation de ces unités de sens au sein de structures plus vastes que nous qualifions de *macro-molécules* de signification. Cependant, il ne s'agit pas simplement de proposer un niveau de granularité supplémentaire, mais de nous placer au point de vue du discours considéré dans sa globalité, pour percevoir les modes d'organisation assurant sa *cohérence* et son *unité*. Par ailleurs, notre démarche se veut *rhétorique*. Il ne s'agit pas seulement de repérer des *parties du discours*, ni de concevoir ce dernier comme leur simple juxtaposition, mais d'adopter un point de vue *logique* sur l'organisation discursive, afin de mettre en évidence sa dimension fondamentalement *relationnelle* et d'envisager la manière dont les sous-éléments interagissent.

À vrai dire, les applications possibles d'une telle approche sont nombreuses dans le domaine de l'informatique documentaire. S'il s'agit bien, *in fine*, de produire une *représentation sémantique du document* en considérant non pas seulement son contenu informationnel, mais également les principes organisationnels procédant à sa structuration, il va sans dire que le bénéfice en matière d'*indexation* sera tout à fait essentiel. Il deviendra possible d'indexer un document non plus seulement comme présentant, par exemple, des informations sur Basse-Normandie *et* Bretagne, mais comme *établissant une relation* entre ces informations. Incidemment les possibilités en matière de *recherche d'information* seront elles aussi renforcées. Nous pourrons envisager des requêtes de la forme : « les documents dans lesquels il est question de Basse-Normandie et de Bretagne et dans lesquels ces régions sont mises en relation de *contraste* » ou « les documents dans lesquels le fait B est *déduit* de A ». Par ailleurs, certains *motifs* rhétoriques s'avèrent efficaces pour l'*extraction d'information*. En procédant par exemple, par *concentration* du propos, ils permettent de capter une information *globale* que le reste du discours *diffuse*. L'extrait 1 fournira l'exemple d'une *amorce hyperonymique* qui joue très précisément ce rôle. On pourra envisager un système de résumé automatique reposant sur l'exploitation de tels principes.

Parallèlement à une étude générale de la rhétorique et de sa portée en TAL, une part importante de notre travail a consisté à évaluer la pertinence de notre approche dans une perspective plus limitée, dans le cadre du projet GeoSem (Enjalbert *et al.*, 2003; Bilhaut *et al.*, 2003a). Dédié au traitement sémantique du document géographique, ce projet a concentré une part importante de son activité à l'analyse de la « dimension » textuelle de l'information géographique, en obtenant en particulier de bons résultats en matière de détection d'expressions temporelles (Bilhaut *et al.*, 2003b) et spatiales (Mathet *et al.*, 2003), en opérant principalement à des niveaux de granularité d'ordre syntagmatique ou propositionnel. Nous nous sommes fixé pour tâche d'envisager un traitement sémantique des structures discursives de plus haut niveau, mais pouvant tirer bénéfice des résultats obtenus sur ces micro-structures, en limitant notre investigation à ce micro-domaine particulier, et à un motif rhétorique particulier : la structure énumérative.

2 Éléments bibliographiques

En amont des tentatives computationnelles d'exploitation du phénomène rhétorique, du côté des approches logico-philosophiques et linguistiques, nous pouvons tout d'abord nous arrêter sur (Aristote, b; Aristote, a) où ARISTOTE jette les bases d'une véritable théorie systématique de la *composition du discours*, en mettant *rhétorique* et *argumentation* au cœur de sa réflexion. À travers en particulier la distinction entre raisonnement *analytique* et raisonnement *dialectique* nous y trouvons des réflexions fondatrices sur les rapports entre rhétorique et logique, entre rhétorique et argumentation, ainsi que sur la dimension fondamentalement intentionnelle, finalisée

de ces dernières. Si la succession d'ARISTOTE a procédé à l'appauvrissement du champ d'application de la discipline en faisant tendre celle-ci vers une simple taxinomie des figures, théorie de l'ornementation et des effets, (Perelman, 1977), en renouant avec la tradition aristotélicienne, vise le retour à l'étude de la composition du discours et de l'argumentation. Sans discuter son choix de considérer la rhétorique comme la rencontre de l'intention d'un auteur (dimension finalisée) et d'un auditoire à convaincre (dimension socialisée), nous cherchons pour notre part à privilégier une approche basée sur la seule structure effective du discours.

En TAL, nous trouvons tout d'abord certaines théories généralistes visant la modélisation de la dimension rhétorique du discours, *en général*. (Mann & Thompson, 1987), à travers la proposition d'une *Rhetorical Structure Theory*, ouvrent la voie à une étude computationnelle des relations rhétoriques, par l'élaboration d'une typologie précise des relations basée sur un modèle en termes de *noyau* et de *satellite*, et sur la spécification des interactions qu'elles établissent entre les éléments en présence. Avec (Grosz & Sidner, 1986) se trouve mise en évidence la dimension intentionnelle de la structure rhétorique : l'inféodation du discours à un but du locuteur en détermine la structure. Les successeurs de ces deux théories fondatrices envisagent une conciliation des points de vue adoptés. La distinction entre structure *informationnelle* et structure *intentionnelle*, amorcée par la RST et radicalisée en ces termes par (Moore & Pollack, 1992), devient, en particulier, tout à fait centrale. Elle oppose les organisations du discours fondées sur la structure du réel (structures informationnelles) à celles qui reposent sur l'intention de l'auteur (structures intentionnelles). Cependant, nous considérons trop restrictive l'échelle à laquelle opèrent ces théories : s'intéressant principalement aux enchaînements d'ordre syllogistique entre syntagmes, propositions et phrases, et adoptant une lecture *ascendante* de la composition textuelle, elles renoncent à certaines structures que nous considérons comme pertinentes.

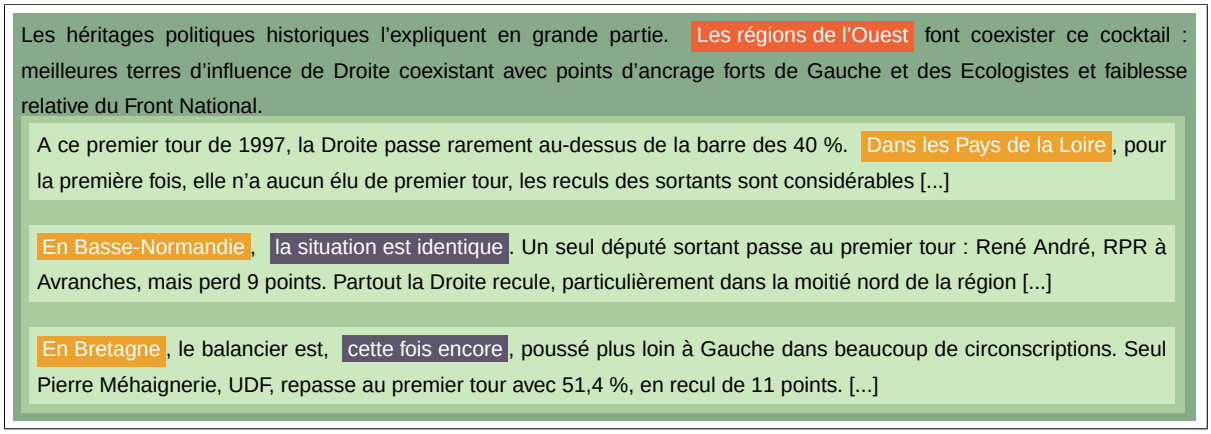
Enfin, toujours dans le champ du TAL, il s'avère également fructueux de nous intéresser à des travaux visant l'étude d'organisations discursives spécifiques telles que les structures énumératives. Si les recherches présentées dans (Garcia-Debanc *et al.*, 2002) nous permettent sur ce point d'isoler certains concepts essentiels, notre propre optique s'en écarte cependant assez largement : ces travaux s'appuient sur l'organisation typo-dispositionnelle du document, quand nous envisageons, au contraire, de nous baser intégralement sur son analyse discursive.

3 Modèle de l'organisation du discours

L'extrait 1 de notre corpus[1] permettra d'entrevoir les différents éléments dont nous visons l'analyse. Trois *blocs* textuels consécutifs présentent chacun une région géographique (« Pays de la Loire », ...), du point de vue d'un *phénomène* (Bilhaut *et al.*, 2003a) commun (ici, des résultats d'élections). Ils constituent trois *items* d'une *énumération*. Entre ces éléments coénumérés, nous observons la spécification d'une *relation* : la « situation est identique » en Basse-Normandie et dans les Pays de la Loire, et cette similarité vaut encore entre Bretagne et Basse-Normandie. L'environnement immédiat de l'énumération révèle qu'elle est englobée dans une structure plus vaste que nous qualifions de *structure énumérative*, introduite par une *amorce hyperonymique* dans laquelle est indiquée la *classe* (« régions de l'Ouest ») dont les items sont les *instances* et qui présente une *loi générale* ensuite *validée* pour chacun des items.

Ainsi, nous avons tout d'abord vu émerger différents *blocs* textuels caractérisés par un certain degré de *cohésion sémantique*, ou disons, d'unité de sens. Les items, l'énumuration ou même

[1] Pascal BULÉON. *Quarante années d'évolution politique de l'Ouest de la France : 1960-2000*.



Extrait 1: Exemple de structure énumérative

la structure énumérative considérée dans sa globalité constituent de telles unités que nous nommons *régions rhétoriques* et qui représentent le premier objet de l'analyse rhétorique. Nous avons par ailleurs observé l'articulation du discours autour de certaines structures procédant à l'*organisation* de l'information. Nous nommons *schémas rhétoriques* de tels principes organisationnels. À ce niveau de réflexion, nous retrouvons l'énumération et la structure énumérative présentes dans l'extrait proposé, mais pouvons également citer à titre d'exemple les structures démonstratives ou les structures méréologiques. Enfin, et cet aspect constitue le troisième élément fondamental de la structuration discursive, nous avons vu apparaître des interactions que nous qualifions de *relations rhétoriques* entre les différents éléments de sens entrant dans la composition des *schémas rhétoriques*. Ainsi, à l'intérieur de la structure énumérative, nous avons observé la présence d'une relation de *similarité* entre les items co-énumérés et d'une relation de type *thèse / instanciation* entre amorce et énumération.

4 Expérimentations sur le discours géographique

Le travail mené dans le cadre du projet GeoSem répond à plusieurs délimitations de notre domaine d'investigation. Tout d'abord, nous nous limitons à l'étude d'un *schéma* rhétorique particulier : la *structure énumérative*. Par ailleurs, le domaine est *défini* (domaine géographique), et nous nous autorisons l'accès à certaines connaissances préalables à son égard : profitant du fait que ce *monde réel* soit fortement structuré, nous nous appuyons par exemple sur une préconnaissance de son organisation (le département de la Manche est en Basse-Normandie...). Nous pourrons par ailleurs tirer profit des *connaissances* acquises aux niveaux de granularité inférieurs, fournies par exemple par les analyseurs spatiaux et temporels déjà évoqués. D'autre part, le corpus envisagé est fondamentalement *descriptif* et nous pouvons nous attendre à ce qu'il existe, entre ce monde *réel* (représentation conventionnelle) et le monde *décrit* (représentation discursive), un certain parallélisme. De plus, ce corpus se caractérise par un style assez *expositif* et presque *pédagogique* d'où il résulte une grande *lisibilité* de la structure discursive.

L'analyse rhétorique procède ici en deux temps : la détection du *schéma* rhétorique de type structure énumérative d'une part, et l'analyse des *relations* rhétoriques entre items d'autre part. En sortie, nous visons à la fois la segmentation du texte en conséquence, et la production d'une structure de traits fournissant l'interprétation sémantique associée (Fig. 1). Pour la détection

des structures énumératives et de l'organisation hyperonymique, l'analyseur s'appuie sur la structuration *informationnelle* du discours. Le domaine géographique est *fortement structuré* et peut être au moins partiellement connu du système : l'analyseur dispose, en entrée, d'une pré-connaissance *hiérarchique* du monde géographique (Caen est une ville du Calvados, qui est un département de Basse-Normandie...). D'autre part, le discours analysé, fondamentalement descriptif, présente un certain degré de transparence entre le monde qu'il décrit et la description qu'il en donne. La détection de structures énumératives peut donc s'appuyer sur cette propriété informationnelle et procèder par détection d'*isomorphismes* entre discours et domaine.

L'analyse des *relations* rhétoriques existant entre les items consiste principalement à interroger les relations de contraste et de similarité, telles qu'étudiées dans (Widlöcher *et al.*, 2004), sous le nom de CURS (*Contrast/Uniformity Relational Structures*). L'analyseur s'appuie tout d'abord sur une série d'indices fréquents et fiables de type *cue-phrases* qui manifeste la présence de relations de cette nature. L'exemple présenté plus haut en donne un aperçu à travers l'usage qu'il fait des expressions comme « la situation est identique ». Nous parlons alors de *mise en relation rhétorique explicite*. D'autres indices plus ténus et *implicites* peuvent être utilisés, de manière incrémentale : la confirmation du sens attribuable à tel indice par tel autre permettra d'avancer. Il est tout d'abord possible de *déduire* la relation par comparaison des représentations sémantiques obtenues pour chacun des éléments en présence. Notre exemple (voir Fig. 1), révèle ainsi que chacun des items est caractérisé par la présence forte d'informations relatives à la droite (« la Droite », « RPR »...) et qu'il existe donc une certaine similarité entre les trois régions sur ce plan. Nous parlons alors de *mise en relation rhétorique par comparaison des représentations isolées*. En ce sens, un usage efficace peut être fait des marques de quantification *effective* (15.000, 16%...) ou *connotée* (recul, augmentation...). Dans l'extrait, nous apprenons ainsi que la similarité porte plus précisément sur le recul de la droite dans chacune des régions.

L'implémentation d'un premier analyseur fondé sur ces principes à été réalisée sous la forme d'un module pour la plate-forme LinguaStream[2]. Il offre de bons résultats quand la structuration discursive est relativement marquée. L'extrait de corpus proposé plus haut en fournit un exemple, et correspond à une sortie effective de notre analyseur (Fig. 1).

5 Perspectives

Il conviendra tout d'abord, à présent, de valider nos hypothèses et l'analyseur proposé, sur un corpus plus large et moins « favorable ». Plus fondamentalement, nous envisageons de dépasser les restrictions méthodologiques et d'aborder d'autres *schémas rhétoriques* (progression thématique, démonstration...) et d'autres *relations rhétoriques* (celles de la RST, loi générale/variations contextuelles...). D'autre part, si notre corpus nous a conduit à privilégier la structuration *informationnelle* du discours, ancrée dans la structure du domaine, nous envisagerons à présent sa dimension *intentionnelle*, en abordant des corpus plus propices à cette étude (discours politiques...). Enfin, dans la continuité de cette prise en compte de l'intentionnalité, nous suivrons dorénavant une orientation plus *logique* en abordant la dimension *argumentative* de l'organisation discursive, et en envisageant les *assouplissements* de la logique pouvant conduire à une *logique naturelle* capable d'en rendre compte. Sur ce point, on pourra se demander si la *rhétorique des figures*, si l'approche plus stylistique de la rhétorique, ne répond pas précisément au besoin d'une *logique souple* propre à décrire l'élaboration du discours.

[2] `http://www.linguastream.org`.

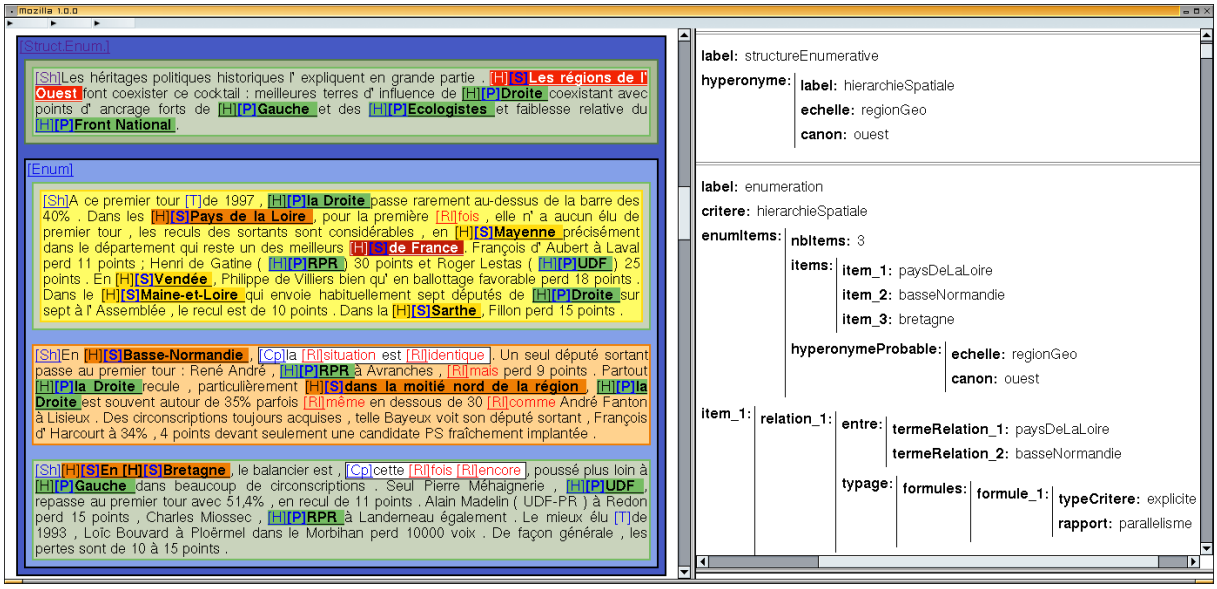


FIG. 1 – Sortie de l'analyseur

Références

ARISTOTE (a). *Organon.* Paris : Vrin.

ARISTOTE (b). *Rhétorique.* Tel. Gallimard.

BILHAUT F., CHARNOIS T., ENJALBERT P. & MATHET Y. (2003a). Passage extraction in geographical documents. In *Proceedings of New Trends in Intelligent Information Processing and Web Mining*, Zakopane, Poland.

BILHAUT F., HO-DAC M., BORILLO A., CHARNOIS T., ENJALBERT P., LE DRAOULEC A., MATHET Y., MIGUET H., PERY-WOODLEY M.-P. & SARDA L. (2003b). Indexation discursive pour la navigation intradocumentaire : cadres temporels et spatiaux dans l'information géographique. In *Actes de Traitement Automatique du Langage Naturel (TALN)*, Batz-sur-Mer, France.

ENJALBERT P. *et al.* (2003). *Projet GéoSem. Traitements sémantiques pour l'Information Géographique : textes, cartes, graphiques.* Rapport interne.

GARCIA-DEBANC C. *et al.* (2002). *Structures spatio-linguistiques du texte : traitements formels et cognitifs.* Rapport interne.

GROSZ B. J. & SIDNER C. L. (1986). Attention, intention, and the structure of discourse. *Computational Linguistics*, **12**(3), 175–204.

MANN W. C. & THOMPSON S. A. (1987). *Rhetorical Structure Theory : A theory of Text Organization.* Rapport interne ISI-RS-87-190, ISI : Information Sciences Institute, Marina del Rey, CA.

MATHET Y., CHARNOIS T., ENJALBERT P. & BILHAUT F. (2003). Geographic reference analysis for geographic document querying. In *Proceedings of Workshop on the Analysis of Geographic References, Human Language Technology Conference (NAACL-HLT)*, Edmonton, Alberta, Canada.

MOORE J. D. & POLLACK M. E. (1992). A problem for RST : The need for multi-level discourse analysis. *Computational Linguistics*, **18**(4), 537–544.

PERELMAN C. (1977). *L'Empire Rhétorique. Rhétorique et Argumentation.* Paris : Librairie Philosophique J.VRIN. Seconde édition 2002.

WIDLÖCHER A., FAUROT E. & BILHAUT F. (2004). Multimodal Indexation of Contrastive Structures in Geographical Documents. In *Proceedings of RIAO 2004*, Avignon, France. À paraître.