Discours, corpus, traitements automatiques

Marie-Paule Péry-Woodley

Chapitre 5

Discours, corpus, traitements automatiques[1]

5.1. Introduction

Ce chapitre propose une réflexion méthodologique sur l'application des principes des linguistiques de corpus, et la mise en œuvre de leurs méthodes, à l'étude de l'organisation discursive. Parce que cette réflexion est issue de nos propres questionnements et pratiques, nous commencerons par exposer brièvement une problématique bien spécifique sur laquelle nous travaillons depuis quelque temps[2] : l'étude de l'organisation temporelle, à la lumière en particulier des hypothèses de Charolles sur l'encadrement du discours (Charolles, 1997). La réflexion méthodologique s'amorcera ensuite avec ce constat fait par plusieurs auteurs : dans la plupart des cas, les études sur le discours, même si elles se fondent sur des textes attestés, ne ressortissent pas pleinement aux études de corpus. Face à ce constat, dont nous devons reconnaître qu'il s'applique assez largement à nos propres travaux, nous chercherons à analyser les difficultés spécifiques rencontrées par les études sur le discours, qui les conduisent en général à si mal satisfaire aux exigences des méthodes d'analyse de corpus. Les interactions entre discours, traitement automatique des langues et analyses de corpus seront ensuite examinées à travers des applications comme le résumé automatique et l'aide à la navigation. Les questions posées par ces applications recoupent en de nombreux points celles qui motivent les études linguistiques du discours. La dimension applicative les oriente en revanche vers des techniques numériques qui ne sont pas celles typiquement utilisées dans ces études. Dans quelle mesure serait-il approprié de reprendre au compte d'études du discours en corpus certaines de ces techniques ? En amont des analyses discursives, la méthode de constitution de corpus peut aussi bénéficier de l'appel à des techniques de profilage automatique des textes. Autre point : un aspect particulièrement positif des linguistiques de corpus est l'accent mis sur la constitution de ressources collectives, c'est-à-dire mises à la disposition de la communauté des linguistes, constamment enrichies par les travaux des uns et des autres, et permettant ainsi de confronter les approches, de partager et de cumuler les acquis. Pour les travaux sur le discours, la contribution à une telle entreprise semble passer avant tout par l'annotation discursive de corpus, qui constitue également une condition de la quantification.

5.2. Un point de vue sur l'organisation discursive : l'encadrement temporel

5.2.1. Problématique linguistique

Nous poserons d'entrée notre motivation générale en reprenant la phrase placée par le linguiste informaticien D. Marcu en exergue de ses pages sur le discours : « Texts are not just simple sequences

---

Chapitre rédigé par Marie-Paule PERY-WOODLEY

[1]Merci à Anne Condamines, à Benoît Habert et à un relecteur anonyme pour leurs remarques et suggestions sur une première version de ce chapitre.

[2] Le Draoulec et Péry-Woodley, 2001 ; 2003 ; 2004 ; Ho-Dac *et al.*, 2001.

of sentences but rather complex artifacts that exhibit a sophisticated high-level, discourse/rhetorical organization/structure »[3]. Notre travail sur l'encadrement temporel s'inscrit dans l'étude de la cohérence discursive à l'écrit. Il concerne une question centrale pour l'organisation discursive, celle de la segmentation. Toute structuration passe en effet par une segmentation, segmenter impliquant à la fois diviser et regrouper en fonction d'un critère organisationnel. Que l'on envisage l'organisation discursive en termes de structure d'information, de structuration thématique, ou de relations de cohérence, la notion de segmentation est présente : recherches de critères de regroupement d'unités (en segments), identification de marques de rupture ou de discontinuité (entre segments), étude des relations (entre segments) qui les hiérarchisent et forment des segments de niveaux supérieurs. L'identification de segments à même de présenter une homogénéité sémantique et/ou de constituer des unités fonctionnelles est également centrale pour les recherches en TAL touchant au discours.

Dans un ouvrage consacré à la sémantique, il convient de situer notre travail dans le champ de la sémantique discursive. Précisons d'abord qu'il ne s'agit pas d'une démarche herméneutique. Nos recherches ne se focalisent pas non plus sur la mise en œuvre d'un calcul du sens visant à en donner une représentation formelle. Nous cherchons plutôt à mettre au jour des principes qui entrent en jeu dans la structuration d'un texte (point de vue du locuteur/scripteur) et dans la construction d'une interprétation cohérente (point de vue de l'interlocuteur/lecteur), à en étudier les corrélats linguistiques, et à examiner les articulations entre divers modes de structuration. Ces visées seront précisées dans le cadre des applications du TAL évoquées en 5.4.

Les textes étant des objets structurés complexes, plusieurs principes et modes de structuration sont à l'œuvre simultanément. L'hypothèse de l'*encadrement du discours* (Charolles, 1997) s'attache à rendre compte d'un mode de structuration particulier, dont nous tentons de faire apparaître l'interaction avec d'autres modes en ce qui concerne la dimension temporelle. Au cœur de cette hypothèse : les *expressions introductrices de cadres de discours* ; pour nous : les adverbiaux temporels en position préverbale. Selon Charolles, les expressions introductrices de cadres de discours signalent que « plusieurs propositions apparaissant dans le fil d'un texte entretiennent un même rapport avec un certain critère, et sont, de ce fait, regroupables à l'intérieur d'unités que nous appellerons cadres » (Charolles, 1997 : 4).

C'est l'aptitude de ces expressions à regrouper plusieurs propositions dans leur portée qui leur confère un intérêt certain pour l'étude de la structuration et de la segmentation des textes. Elles sont décrites comme jouant

un rôle fondamentalement procédural et cognitif et cela à deux niveaux :
- d'une part, elles servent à régler les opérations de mobilisation de connaissances requises pour l'interprétation pas à pas des relations entre propositions. (…)
- d'autre part, elles servent à répartir les contenus propositionnels dans des blocs homogènes relativement à un critère spécifié par le contenu de l'introducteur. (Charolles, 1997 : 24)

En faisant référence aux travaux de Halliday (1985) et de Thompson (1985), nous parlerons d'un fonctionnement double : à la fois sur le plan de la *métafonction idéationnelle* – l'expression introductrice de cadre fournit un critère d'interprétation, elle concerne la mobilisation de connaissances pour l'interprétation – et sur le plan de la *métafonction textuelle* – constitution d'un segment homogène quant à ce critère. Ainsi dans l'exemple (1), les quatre assertions concernant *l'école primaire* (jusqu'au ♦ qui indique la frontière droite du cadre) doivent être interprétées en rapport avec l'expression temporelle initiale (en gras) :

(1) ***De la fin du siècle dernier jusqu'aux années 1950****, l'école primaire a été le pilier du système scolaire français. Elle inculquait les connaissances de base, lire, écrire et compter, qui serviraient toute la vie. Elle avait aussi pour mission de former les citoyens de la République. Elle délivrait le*

[3] http://www.isi.edu/~marcu/discourse/

certificat d'études qui, pour le plus grand nombre, attestait de la réussite des études et marquait l'entrée dans le monde du travail. ♦
Les sessions du certificat d'études n'ont plus lieu.

Ces quatre assertions sont, dans les termes de Charolles, *indexées* par la référence temporelle réalisée dans l'expression introductrice de cadre, et se trouvent de ce fait regroupées dans un *cadre de discours*. Ce processus d'indexation se distingue des autres procédés de cohésion par la particularité – importante sur le plan du traitement cognitif – de fonctionner d'amont en aval, et non d'aval en amont comme c'est le cas pour ces autres procédés (anaphore, relations de discours), que Charolles (2002) regroupe sous le terme de *connexion*.

Toujours en rapport avec la problématique de la segmentation et l'interaction entre modes d'organisation du discours, nos travaux sur l'encadrement temporel se sont axés essentiellement sur les questionnements suivants :

– le rôle structurant associé à la possibilité d'étendre leur portée sur plusieurs propositions est-il réservé aux adverbiaux en position préverbale ? (Le Draoulec et Péry-Woodley, 2001)

– peut-on identifier de manière opérationnelle des marques signalant la borne droite des cadres initiés par un introducteur de cadre temporel ? (Bilhaut *et al.*, 2003)

– l'hypothèse de l'encadrement, formulée initialement « en langue », doit-elle être modulée pour tenir compte du genre discursif ? (Le Draoulec et Péry-Woodley, 2003)

Ce dernier travail a fait apparaître que dans des textes narratifs, où la relation de *Narration* (Asher, 1993) est très prégnante, le potentiel cadratif des expressions temporelles préverbales est contrarié par l'avancement du temps d'événement. L'index temporel fixe fourni par l'introducteur de cadre se trouve dans ces configurations en compétition avec la progression de la référence temporelle liée à la successivité des événements, ce qui a pour résultat un affaiblissement du rôle structurant de l'encadrement. Cette étude met en évidence l'interaction entre ces deux processus majeurs en jeu dans l'organisation discursive que sont d'une part l'encadrement / indexation et de l'autre la connexion (*via* les relations de discours). Les cadres forment des segments dont les unités sont regroupées par un critère d'interprétation (index), mais à l'intérieur d'un cadre, les relations entre les propositions jouent également, pour éventuellement favoriser (exemple de l'*Elaboration*) ou au contraire contrecarrer (*Narration*) son rôle structurant (Le Draoulec et Péry-Woodley, 2003 ; 2004). L'examen en corpus de cette interaction, qui devrait nous permettre de préciser la notion de portée de l'introducteur de cadre, est en cours. Par ailleurs, une étude examine les interrelations entre introducteurs de cadres et titres, ou sur le plan des processus, entre segmentation en cadres et en parties titrées[4] ; une autre étude également en cours porte sur la fonction discursive de la position initiale à différents niveaux de structuration (phrase typographique, paragraphe, partie titrée : Ho-Dac, à paraître 2004). Signalons que ces travaux s'inscrivent dans deux projets aux objectifs plus vastes : a) « Adverbiaux spatio-temporels et discours »[5] et b) GEOSEM : « Traitements sémantiques pour l'Information Géographique : textes, cartes, graphiques »[6]. La visée applicative du second, et l'insertion des travaux sur l'encadrement dans cette visée, seront développés dans la section 5.4.2. Il est maintenant temps d'en venir à l'objectif premier de ce chapitre, et d'entamer, à partir des travaux qui viennent d'être évoqués, la réflexion méthodologique sur l'étude du discours en corpus.

5.2.2. Aspects méthodologiques : un regard (auto)critique

L'hypothèse de l'encadrement du discours telle qu'elle est formulée dans (Charolles, 1997) s'appuie sur des exemples attestés, et non sur une étude de corpus. Ceci n'est pas une critique puisque

[4] Mémoire de DEA de Sciences du Langage (M. Laignelet, 2004) : *Les titres et les cadres de discours : structuration des discours et organisation de l'information.*
[5] Etude soutenue par l'Institut de la Langue Française : http://www.ilf.cnrs.fr/fr/gabarits/01b_axrecherche.php. Etude soutenue par l'Institut de la Langue Française.
[6] http://infodoc.unicaen.fr/geosem/. Projet du programme CNRS « Société de l'information ».

là n'est pas son propos. Notre objectif, en revanche, est une étude en corpus de l'encadrement temporel et de son interaction avec d'autres modes d'organisation. Nous voyons dans l'hypothèse de l'encadrement l'élaboration d'une intuition forte sur un mode d'organisation distinct des procédés de cohésion « classiques » (anaphores, relations de discours), intuition que nous avions commencé à examiner, de façon indépendante, dans des travaux antérieurs (Péry-Woodley, 1993), et qui peut guider l'examen des données textuelles. Cet objectif général se subdivise en objectifs particuliers liés aux questionnements évoqués en 5.2.1., parmi lesquels, pour cette discussion, nous retiendrons les trois suivants :

1 - confirmation de la spécificité de la position initiale en ce qui concerne le potentiel de portée des expressions temporelles ;
2 - identification de marques de frontière droite des cadres temporels ;
3 - examen de l'impact du genre discursif sur leur fonctionnement.

La constitution du corpus est en partie liée au contexte du projet GEOSEM, et à la richesse des documents de géographie humaine en expression de la localisation spatio-temporelle. Il comprend trois atlas[7], dont le plus utilisé pour ces travaux est l'Atlas de la France scolaire (56 346 occurrences). Pour l'étude de l'impact du genre discursif, on a adjoint à ce corpus initial un recueil de résumés de films[8] (32 202 occurrences) et un ensemble de textes « historiques » (synthèses chronologiques, 41 384 occurrences)[9].

Nous avons travaillé sur les textes bruts ou pourvus d'étiquettes morpho-syntaxiques, en faisant appel principalement à trois outils d'exploration et d'analyse de corpus : les concordanciers Yakwa, et MonoConc Pro, et l'outil de statistique textuelle Lexico [10]. Dans le cadre du projet GEOSEM, nous avons également pu exploiter la plate-forme LinguaStream, qui conjugue des traitements lexicaux, syntaxiques et discursifs dans une visée d'aide à la navigation, et dont le développement fait partie intégrante du projet[11].

Le contexte linguistique et méthodologique étant posé, nous voudrions rapidement esquisser ce qui nous apparaît comme des limites de ces études, limites qui feront ensuite l'objet d'un examen à la fois plus systématique et plus général à la lumière d'un retour aux principes qui fondent les linguistiques de corpus.

L'objectif (1) concerne une propriété importante, potentiellement généralisable, des adverbiaux temporels initiaux. Notre étude suggère que ce que nous appelons « potentiel cadratif » – la possibilité de jouer un rôle structurant sur le plan textuel – est bien le propre des adverbiaux à l'initiale, mais sa nature exploratoire, illustrative, limite le statut et la portée de ce résultat. L'objectif (3) implique la constitution d'un corpus partitionné en termes de genre discursif, ce qui pose le problème de la caractérisation en genre des documents sélectionnés. Nous avons partiellement éludé cette question en adoptant une simple opposition narratif / non narratif, mais notre choix de textes ne se fonde pas sur une caractérisation rigoureuse. L'identification des marques de frontière droite des cadres (objectif 2) a fait l'objet d'une mise en œuvre partielle dans la plate-forme LinguaStream, et d'une tentative de validation manuelle. Mais celle-ci a fait émerger d'autres problèmes, qui ont fait apparaître la nécessité de préciser la notion de portée (à travers l'étude de l'interaction avec les relations de discours) avant de pouvoir revenir à la recherche de bornes droites. En vrac, donc, des problèmes liés à

[7] Hérin, R. et Rouault, R. (1994) *Atlas de la France Scolaire de la Maternelle au Lycée*. Paris : La Documentation Française ; Buléon, P. (2002) *Atlas politique*. Consultable à l'URL: http://atlas-politique.certic.unicaen.fr ; *Atlas Transmanche*, ouvrage collectif et évolutif, consultable à l'URL : http://atlas-transmanche.certic.unicaen.fr.

[8] Recueillis sur les sites Internet de la chaîne de télévision *Canal+* et du journal *Libération*.

[9] Textes recueillis sur le site Internet : *La résistance allemande au nazisme* (http://resistanceallemande.online.fr/).

[10] Yakwa est développé par L. Tanguy ; Lexico par A. Salem ; MonoConc Pro par M. Barlow. Informations sur ces outils sur les sites suivants (pages disponibles en septembre 2004) :
- Yakwa : http://www.univ-tlse2.fr/erss/textes/pagespersos/tanguy/Yakwa.html ;
- Lexico : http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/fsalem.htm ;
- MonoConc Pro : http://www.ruf.rice.edu/~barlow/mono.html#monopro.

[11] LinguaStream (F. Bilhaut) : http://users.info.unicaen.fr/~fbilhaut/linguastream.html.

la nature des données fournies par l'analyse, aux méthodes de construction de corpus, à la difficulté de constituer des observables pertinents par rapport à un objet construit de manière théorique quand on est confronté à la jungle textuelle.

5.3. Linguistique de corpus et discours : un constat

Dans un panorama récent des linguistiques de corpus (Biber *et al.*, 1998), on trouve le constat suivant en ce qui concerne la situation des études sur le discours :

Most discourse studies identify salient discourse structures and exemplify those structures with illustrative text excerpts – such as identifying turn-taking structures in conversation or tracking the 'themes' in a written text. However, it has proven difficult to apply these techniques to texts in a way that allows for generalizable results. Thus, although nearly all discourse studies are based on analysis of actual texts, they are not typically corpus-based investigations: most studies do not use quantitative methods to describe the extent to which different discourse structures are used, and relatively few of these studies aim to produce generalizable findings that hold across texts. (Biber et al.*, 1998 : 106)*

Les auteurs établissent ainsi une distinction nette entre le fait de fonder une étude sur des exemples attestés, ce que font la plupart des études sur le discours[12], et le fait de mettre le corpus au centre de l'étude, ce qui pour eux passe par la quantification des résultats et la possibilité de les généraliser à d'autres textes. On trouve un constat similaire dans le manuel d'introduction aux linguistiques de corpus de McEnery et Wilson (1996). Il se trouve par ailleurs conforté par la relative rareté des études sur le discours dans les principaux forums des linguistiques de corpus, tels les conférences *Corpus Linguistics* qui se sont tenues à Lancaster en 2001 et 2003 ou les récentes livraisons de l'*International Journal of Corpus Linguistics*, pour ne citer que ces quelques exemples. Cette rareté tendrait à indiquer que les linguistes travaillant sur le discours, bien qu'ils fassent appel à des textes attestés, ne sont pas perçus, ou ne se perçoivent pas eux-mêmes, comme faisant partie de la communauté des linguistes travaillant sur corpus.

Pour éclairer cette situation, il est utile de creuser un aspect du constat de Biber *et al.* (1998) : l'importance accordée dans les linguistiques de corpus à la quantification, et à son rôle dans la généralisabilité des analyses proposées.

5.3.1. Approches quantitatives *vs.* qualitatives

Le constat ci-dessus a trait à la difficulté d'appliquer aux études sur les structures discursives le type de méthode qui permettrait d'évaluer l'importance quantitative de ces structures, approche sans laquelle, selon les auteurs cités, il ne peut y avoir de comparaison ni de généralisation. C'est un précepte tout à fait propre aux linguistiques de corpus qui se trouve ainsi formulé, avec la mise en avant des caractéristiques de fréquence d'usage. Celles-ci sont en effet demeurées largement absentes des préoccupations d'une linguistique – descriptive aussi bien que théorique – qui d'une façon générale s'est montrée peu concernée par les données quantitatives, qu'il s'agisse de fréquence relativement à d'autres structures, ou de variations de fréquence d'un groupe de textes à un autre (cf. Sampson, 2001).

Cette problématique de la quantification est développée par McEnery et Wilson (1996), qui exposent les limites des approches qualitatives, même s'ils leur reconnaissent un rôle potentiellement important de « précurseurs » de l'analyse quantitative :

[12] A l'exception des approches relevant de la sémantique formelle, dont les objectifs de modélisation très fine des interactions entre les différents composants du discours (niveaux syntaxique, lexical, sémantique – phrastique et interphrastique –, pragmatique...) entraînent souvent une limitation drastique des données traitées (cf. Péry-Woodley, 2001).

Whereas in quantitative research we classify features, count them and even construct more complex statistical models in an attempt to explain what is observed, in qualitative research the data are used only as a basis for identifying and describing aspects of usage of the language and provide 'real-life' examples of particular phenomena. […]
[…] the main disadvantage of qualitative approaches to corpus is that their findings cannot be extended to wider populations with the same degree of certainty with which quantitative analyses can, because, although the corpus may be statistically representative, the specific findings of the research cannot be tested to discover whether they are statistically significant or more likely to be due to chance. (McEnery et Wilson, 1996 : 62)

Bestgen *et al.* (2003) identifient un autre problème lié à l'approche illustrative dominante dans les études sur le discours – non seulement le volume de données est trop limité, mais les analyses sont excessivement dépendantes de l'analyste qui les exécute :

While these empirical studies [hand-based studies of connectives] are useful from a qualitative point of view, they all suffer from the same quantitative drawback, namely the relatively small number of data (rarely more than 100 occurrences are analysed, mostly only 50). In addition, most of these analyses are still too analyst-dependent, making generalisations and replications difficult. (Bestgen et al.*, 2003 : 189)*

Pour résumer la situation telle que la décrivent les auteurs cités dans cette section, les études sur le discours se caractérisent actuellement par une approche qualitative, sur des données de faible volume, avec des méthodes manuelles et donc subjectives (« analyst-dependent »), ce qui fait obstacle à leur reproductibilité – et partant à leur validation –, et à la généralisation de leurs résultats. Sans quantification, pas de comparaison possible : l'usage mis en évidence est-il simplement un effet du hasard ou est-il réellement représentatif ? quel est son statut par rapport à d'autres usages ? est-il lié à un texte (ou groupe de textes) ou peut-il être généralisé à d'autres textes ou corpus ? Cette mise en cause m'amène à engager dans ce qui suit une réflexion sur les principes fondamentaux des linguistiques de corpus, de manière à y situer l'exigence de quantification et de reproductibilité, et à cerner les implications et les difficultés de la mise en œuvre d'une linguistique du discours en corpus.

5.3.2. Exigences des linguistiques de corpus : le point de vue du discours

L'insistance sur la quantification a des implications pour le volume de données considérées, et partant pour les méthodes utilisées : de faibles volumes de données rendent la quantification, ou tout au moins sa validation statistique impossible ; par ailleurs, ces petits effectifs vont de pair avec des méthodes qui restent dépendantes de l'analyste, et peu reproductibles ; le traitement de gros corpus pour l'acquisition de données plus riches implique quant à lui la mise en œuvre de méthodes (semi-) automatiques, i.e. informatiques, qui ont l'avantage d'être reproductibles. Ainsi l'union entre linguistiques de corpus et outils informatiques est-elle scellée, contribuant au mouvement de fond vers un empirisme qui, sous des formes plus ou moins radicales, peut être vu comme définitoire (cf. Sampson, 2001 ; Tognini-Bonelli, 2001).

Le lien entre approche en corpus et informatique apparaît en effet comme une partie intégrante de la démarche empirique prônée depuis une quinzaine d'années – avec des variantes – par les tenants de la linguistique de corpus. Dans un article qui a beaucoup contribué à préciser les implications de cette approche, Leech (1992) énonce quatre traits qui caractérisent le nouveau domaine qu'il nomme précisément « Computer Corpus Linguistics » (CCL). Ce domaine est défini comme étant centré :

(1) sur la performance plutôt que sur la compétence ;
(2) sur la description linguistique, plutôt que sur les universaux linguistiques ;
(3) sur des modèles quantitatifs, autant que sur des modèles qualitatifs du langage ;
(4) sur une vision empirique, plutôt que rationaliste, de la démarche scientifique. (Leech, 1992 : 107)

Leech ajoute un point de méthode essentiel à cette définition, point qu'on peut rapprocher de la critique formulée par Bestgen *et al.* (op. cit.) à l'encontre des méthodes dépendantes de l'analyste, en posant le principe suivant :

[the] data are used exhaustively: there is no prior selection of data which we are meant to be accounting for and data we have decided to ignore as irrelevant to our theory. This principle of 'total accountability' for the available observed data is an important strength of CCL. (Leech, 1992 :112)

Il précise que ce principe d'exhaustivité n'exclut pas que l'on puisse s'intéresser à un « niveau » ou à un « aspect » particulier plutôt qu'à la langue en entier, et que l'on puisse étudier un échantillon du corpus et non le corpus intégral. Le principe fondamental, reformulé dans la suite du texte de façon plus polémique, est que

no theoretically-motivated selection process intervenes to choose suitable data, as so often happens in other varieties of linguistics. (Leech, 1992 : 113)

On est bien loin du « picorage » en corpus à la recherche d'exemples attestés : le linguiste a pour mission de rendre compte de toutes les occurrences de la structure étudiée rencontrées dans le corpus. Cependant, cette structure à l'étude demeure quant à elle la projection d'une construction théorique. Toute exigeante qu'elle soit, la position de Leech est pourtant dépassée par certains auteurs, comme Tognini-Bonelli (2001) : ce que Leech rejette avec force, c'est en effet la sélection biaisée des données envisagées, mais l'analyse de corpus n'en reste pas moins un banc d'essai pour des modèles ou des parti pris théoriques préexistants, qu'elle ne pourra pas affecter en profondeur. Dans cette approche, dit Tognini-Bonelli, "the 'order' is superimposed on the evidence of language use, rather than derived from it." (Tognini-Bonelli, 2001 :182) Ainsi, le fait de rassembler dans une même classe fonctionnelle lemmes et formes fléchies, comme le font les catégories traditionnelles, va à l'encontre de ce qu'on observe dans les corpus. De même, la notion classique de ce qui constitue une unité de sens tend à être ébranlée par la mise au jour de patrons systématiques qui définissent de nouvelles unités fonctionnelles. Au delà d'une linguistique qui se fonde sur les corpus, « corpus-*based* linguistics », c'est une approche radicalement empirique qui est alors prônée : « a corpus-*driven* approach », qui pose l'observation du corpus comme point de départ, et pour laquelle les catégories théoriques doivent dériver d'événements langagiers « typiques, récurrents et observables de manière répétée ».

Pour caractériser les études linguistiques se référant à des données attestées, il serait donc possible de les situer en regard de plusieurs dimensions :

1 - la taille du corpus ;
2 - le volume des données envisagées : cette dimension est à distinguer de la précédente dans la mesure où la densité des observables varie selon le phénomène à l'étude (problème des données dites « sparse ») ;
3 - la prise en compte plus ou moins exhaustive des données ;
4 - la visée plutôt qualitative ou plutôt quantitative de l'étude ;
5 - l'approche plus ou moins ouverte (cf. « corpus-based » *vs.* « corpus-driven ») : p.ex. recherche de formes connues *vs.* procédures de découverte ;
6 - les méthodes d'analyse : analyse subjective ou appel à des procédures automatiques cherchant à « objectiver » le processus interprétatif.

Lorsqu'on les soumet à cette caractérisation, il apparaît clairement que beaucoup d'études sur l'organisation discursive – et les nôtres en particulier - ne peuvent se réclamer de l'appellation « linguistique de corpus ». Sans mettre en cause l'utilité des travaux théoriques et des études qualitatives, nécessaires « précurseurs » pour reprendre le terme de McEnery et Wilson, il semble important de mener les études de l'organisation discursive vers les conditions qui autoriseront leur reproduction, leur validation, la comparaison entre corpus. L'étape suivante dans cette réflexion est de s'interroger sur les difficultés spécifiques qui sont à l'origine de cette situation. Plusieurs auteurs ont

entrepris de les recenser. En reprenant McEnery et Wilson (1996) et Biber *et al.* (1998), on peut relever les problèmes suivants :

– la plupart des premiers corpus construits – pour l'anglais – dans le but de représenter la diversité des usages (tels les corpus Brown, LOB, etc.) sont constitués d'extraits et non de textes entiers, ce qui exclut toute possibilité d'analyse discursive au-delà d'un niveau très local ;

– la pragmatique et l'analyse du discours dépendent du contexte, et les corpus éliminent une bonne partie du contexte des énoncés ;

– les outils d'analyse de corpus disponibles sur le marché (les concordanciers par exemple) sont peu adaptés à la visualisation du contexte élargi. Ajoutons que le manque d'outils de visualisation dépasse ce simple problème d'accès au contexte : on aurait besoin d'outils facilitant la détection d'éventuelles régularités de distribution de phénomènes peu fréquents sur de vastes zones de texte – par exemple une tendance des introducteurs de cadre temporels à apparaître en début de paragraphe –, ou l'évolution de la distribution au fil d'un document.

Posent également problème le fait que le format des corpus, ainsi que les interfaces d'accès, entraînent souvent la perte de mise en forme originale (typographie, disposition) ; la difficulté d'arriver à une classification typologique du corpus qui soit pertinente par rapport à l'objet d'étude.

Une difficulté beaucoup plus fondamentale est évoquée par Biber *et al.* (1998 : 107) : de nombreux traits discursifs ne se prêtent pas facilement à un repérage automatique, dans la mesure où leur identification nécessite un examen détaillé des traits linguistiques présents dans leur contexte élargi. Dans de nombreux cas en effet, l'identification des caractéristiques formelles susceptibles de fonctionner comme marque d'un principe de structuration (marque de relation, borne de segment, etc.) fait partie intégrante de l'étude de tel ou tel mode d'organisation. C'est ainsi que Virtanen (1992), au début d'une étude des fonctions discursives de la position des adverbiaux en anglais, annonce d'emblée : « it is a well-known fact that textual phenomena are in general too fuzzy to be quantified ». Son objectif d'examiner le rôle discursif des adverbiaux dans la perspective du texte entier lui semblant incompatible avec une approche quantitative, elle tranche pour une approche résolument qualitative (Virtanen, 1992 : 31). Faut-il se résigner et se dire qu'on est, en ce qui concerne l'organisation du discours comme en ce qui concerne le sens des mots – selon le grand défenseur de la méthode empirique en linguistique qu'est Sampson –, en dehors du domaine de cette méthode[13] ?

La vogue vite délaissée des analyses quantitatives des procédés de cohésion qui a fait suite à l'étude de Halliday et Hasan (1976) sur la cohésion en anglais[14], et dont le but était le calcul d'indices de la cohérence textuelle, incite à la prudence. D'abord parce que les marqueurs répertoriés sont souvent polyfonctionnels, assumant un rôle phrastique ou un rôle discursif selon les configurations : ainsi un connecteur comme *mais* ou *et* peut relier deux constituants phrastiques ou deux unités textuelles (cf. Marcu, 2000) ; de surcroît, une même occurrence peut avoir à la fois un rôle phrastique et un rôle discursif : un adverbe de phrase par exemple peut signaler le premier item d'une structure énumérative s'étendant sur plusieurs paragraphes. Par ailleurs, les inventaires ne peuvent être clos, soit parce qu'il s'agit de classes ouvertes, soit – si, comme pour les connecteurs, on a affaire à une classe fermée – parce que la fonction structurante peut s'accomplir en leur absence (marques non lexicales – e.g. typo-dispositionnelles –, absence de marque). Un aspect fondamental des travaux sur le discours est précisément la recherche ouverte des corrélats linguistiques des principes d'organisation. On peut voir dans la recherche exploratoire, dans des corpus, de ces corrélats linguistiques une application spécifique du principe d'exhaustivité de Leech, pour autant qu'on cherche à identifier toutes les réalisations linguistiques associées à une fonction discursive. Mais il est clair qu'avec de tels objectifs, le recours à des méthodes quantitatives est infiniment plus difficile que dans l'étude de marqueurs lexicaux classiques déjà répertoriés et identifiables par simple filtrage. On aurait donc une tension

[13] "Formal theories about grammatical or phonological patterns in a natural language may at least potentially be testable predictions, but human lexical behaviour is such that analysis of word meaning cannot be part of empirical science." (Sampson, 2001: 181)

[14] Voir Morgan et Sellner (1980) pour une présentation critique.

entre exhaustivité – dans le sens de prise en compte de toutes les réalisations linguistiques dans un corpus donné – et méthodes quantitatives.

Clairement, l'adoption de méthodes quantitatives pour l'étude de l'organisation discursive ne saurait se limiter à des décomptes et à des calculs sur les marques connues. On a besoin de techniques permettant d'appréhender et d'articuler des fonctionnements souvent enchevêtrés, à différents niveaux de granularité ; on a besoin de pouvoir penser ces fonctionnements de manière probabiliste plutôt que binaire ; on a aussi besoin d'outils pour traiter, enrichir, visualiser des données volumineuses. La section suivante propose d'aller voir du côté du TAL. La dimension informatique a en effet été présentée – à la suite de Leech en particulier – comme constitutive des linguistiques de corpus actuelles parce qu'indispensable à la mise en œuvre de méthodes quantitatives. En même temps qu'un ensemble d'outils pour l'exploration de corpus et le traitement de données, l'informatique, à travers les applications du traitement automatique des langues, constitue un cadre qui impose des exigences méthodologiques, fournit des solutions techniques, et peut amener à poser différemment certaines questions linguistiques.

5.4. Approches du discours en traitement automatique des langues

Les applications du traitement automatique des langues qui vont retenir notre attention sont celles concernées par l'exploitation de bases de documents. D'une manière générale, les travaux sur l'indexation et la recherche d'information font peu référence à la structuration discursive des documents. En effet, tant que l'unité pertinente est le document, la structuration interne de celui-ci n'a pas lieu d'intervenir. Mais les choses changent dès que la visée applicative implique que l'on pénètre dans le document pour identifier des zones de texte jugées « importantes », ou pertinentes par rapport à une requête ou à un profil. Les méthodes évoquées ci-dessous concernent des applications distinctes mais entre lesquelles les frontières ont tendance actuellement à s'atténuer : synthèse de documents, certaines formes d'extraction d'information, navigation inter- ou intra-documentaire. Dans la mesure où elle se focalise sur des applications qui suscitent ou exploitent des travaux empiriques sur le discours, c'est-à-dire où TAL, discours et corpus sont tous trois présents, cette revue des recherches sera très partielle. De manière à préciser son objet, il est instructif de la contraster avec deux synthèses récentes : l'une (Nazarenko, ce volume) part des méthodes automatiques d'accès au contenu textuel dans diverses applications pour s'interroger sur la sémantique sous-jacente ; l'autre (Moore et Wiemer-Hastings, 2003) part au contraire d'une revue de différents modèles pour en arriver aux « techniques de traitement du discours » dans des applications.

Nazarenko (ce volume[15]) propose un panorama de méthodes actuellement utilisées dans quatre familles d'applications visant l'accès au contenu textuel : extraction d'information, systèmes de question-réponse, aide à la navigation, résumé automatique. Elle examine en particulier le rôle des techniques de repérage d'entités nommées (noms propres, dates, …), qui constituent selon elle l'élément clé d'analyses combinant des traitement très hétérogènes, dont l'assemblage répond davantage à des critères pragmatiques qu'à une motivation théorique. La conception du « contenu » sous-jacente dans ces méthodes participerait donc d'une sémantique largement référentielle (les entités du monde réel auxquelles il est fait référence), et par ailleurs d'une sémantique « éclatée » dans la mesure où on n'analyse que des « bouts » de texte, des « ilôts », qu'on ne cherche pas à mettre en relation [16]. Les techniques sur lesquelles nous allons nous arrêter dans ce qui suit, dont plusieurs sont d'ailleurs mentionnées par Nazarenko, s'adjoignent au repérage/typage d'entités nommées en cherchant précisément à les situer dans le document. On verra cependant que l'introduction de la dimension discursive ne signifie pas nécessairement que l'on échappe à l'« éclatement » dont parle Nazarenko. Pour reprendre sa formulation :

[15] Ce paragraphe se fonde sur une version préliminaire du chapitre rédigé par A. Nazarenko. Les références correspondent à la pagination du manuscrit

[16] Ce constat rejoint notre discussion de la notion de « topic » dans les méthodes de résumé automatique (5.4.1).

*La confrontation de ces différentes méthodes d'accès au contenu des documents montre qu'elles ne reposent sur aucune sémantique unifiée. Il est loin le temps où les modules d'analyse sémantique servaient à produire une représentation globale et intégrée du contenu d'un texte ou même d'une phrase, écrite sous la forme d'un formule logique ou dans un formalisme dédié, laquelle pouvait ensuite servir à différentes tâches (traduction, interrogation, résumé ou même indexation). (*Ibid. : 10)

Le second article de synthèse, intitulé « Discourse in Computational Linguistics and Artificial Intelligence » (Moore et Wiemer-Hastings, 2003), fait pendant à celui de Nazarenko – du point de vue de l'étude de l'organisation discursive –, en prenant la question par l'autre bout, par celui de modèles proposant une sémantique unifiée. Mais on ne peut qu'être frappé dans cet article par la très faible intersection entre l'exposé des modèles (DRT, Grosz et Sidner, RST, SDRT), suivi d'un long développement sur la génération automatique, et la courte section consacrée aux applications – résumé automatique et systèmes de question-réponse –, qui sous la dénomination « discourse processing techniques » ignore les modèles précédemment cités et fait principalement référence à des méthodes statistiques et à des analyses surfaciques. Cela dit, il est intéressant de noter que les auteurs signalent parmi les chantiers les plus prometteurs pour l'avenir l'intégration des approches statistiques dans la modélisation du discours, intégration sur laquelle nous reviendrons (5.5.1).

Ces deux synthèses pointent chacune à leur façon la difficulté de mettre en œuvre une sémantique discursive : la « multiplication des angles d'analyse » dans les méthodes d'accès au contenu débouche sur une « sémantique éclatée », qui ignore assez largement le rôle sémantique de l'organisation discursive (cf. connexion et indexation, section 5.2) ; les modèles conçus pour théoriser cette organisation résistent quant à eux à l'opérationnalisation. Les sections suivantes vont s'attacher, dans les mêmes familles d'applications, à examiner spécifiquement les techniques qui d'une façon ou d'une autre, et à différents niveaux de grain, ont trait à l'organisation discursive.

5.4.1. « Résumé automatique », synthèse de documents, « text zoning »

Les méthodes actuelles de « résumé automatique » (cf. Mani et Maybury, 1999 ; Desclés et Minel, 2000 ; Mani, 2001) sont pour l'essentiel fondées sur l'extraction : la sélection des « énoncés importants » qui seront assemblés pour construire le « résumé » se fait par l'attribution d'un score composite aux unités textuelles (généralement les phrases). Une première composante de ce score se fonde sur des calculs lexicaux (e.g. *tf.idf*) ; s'y ajoutent ensuite des pondérations en fonction de divers indicateurs. C'est dans cette étape de pondération qu'interviennent des considérations liées à la structuration des documents. Nous citerons en particulier a) la prise en compte de la position des énoncés dans le texte et b) le repérage de « cue phrases » ou « expressions prototypiques ».

Exemplifiant la première approche, le « module positionnel » du système SUMMARIST (Hovy et Lin, 1999) exploite le fait que « in some genres, regularities of discourse structure and/or methods of exposition mean that certain positions tend to carry important topics ». (Hovy et Lin, 1999 : 85) Des méthodes par apprentissage sont mises en œuvre pour déterminer les « positions optimales » dans différents genres, sur lesquelles se fonde l'attribution de poids aux phrases. Hovy et Lin (1999) donnent l'exemple du corpus Ziff-Davis (13 000 articles de presse annonçant des produits informatiques), dans lequel les positions optimales se déclinent comme suit : [T1, P2S1, P3S1, P4S1, P2S2, {P3S2, P4S2, P5S1, P1S2}, P6S1, …], c'est-à-dire 1er titre > 1ère phrase du 2ème paragraphe > 1ère phrase du 3ème paragraphe, etc. [17]

La notion de « topic », qui se trouve au cœur des diverses procédures de calcul de score, est ici très liée aux fréquences lexicales, ainsi que le montre l'évaluation en termes du taux de couverture par les phrases extraites des mots-clés fournis par un humain à partir de l'examen d'un texte. La conception sous-jacente, en ce qui concerne les principes d'organisation discursive, est dominée par la notion de

[17] Noter que. le premier paragraphe, dont on nous dit qu'il constitue invariablement une accroche dans ce corpus, n'est pas ici une position optimale.

topic comme « l'entité dont il est question » et par une vision a-rhétorique du texte essentiellement comme un lieu où sont introduits et traités plusieurs topics. On retrouve ces présupposés dans les techniques de segmentation automatique qui font appel à des mesures de similarité entre phrases pour déterminer des regroupements de phrases apparentées lexicalement (cf. *TextTiling* : Hearst, 1997 ; *Latent Semantic Analysis* : Choi *et al.*, 2001)[18]. Ces techniques, qui s'appuient sur l'hypothèse que la cohésion lexicale d'un segment de texte reflète sa cohérence thématique, sont utilisées par certains auteurs pour identifier des segments « thématiques » susceptibles d'intervenir dans le classement des énoncés importants.

La seconde approche pour la pondération des unités textuelles à sélectionner se fonde sur le filtrage de mots ou expressions pouvant signaler des « énoncés importants ». Parmi les « cue phrases » ou « expressions prototypiques » relevés par divers auteurs à partir d'analyses de corpus, certaines ont trait à l'organisation discursive dans la mesure où elles signalent des segments rhétoriques spécifiques : « en résumé », « en conclusion », etc. Cette approche, qui peut paraître un peu fruste à travers cette rapide présentation, a été considérablement affinée et enrichie par certains auteurs, en particulier dans le cadre de la méthode d'exploration contextuelle (Desclés et Minel, 2000. ; Minel *et al.*, 2001 ; Minel, 2003), méthode qui articule l'identification d'indicateurs linguistiques (ensembles de marqueurs associés à une « étiquette sémantique ») et d'indices complémentaires. Parmi les étiquettes sémantiques touchant à l'organisation discursive, mentionnons « annonce thématique », « récapitulation/ conclusion thématique ».

Les objectifs de ces méthodes de surface sont moins l'analyse ou la représentation des structures textuelles que le repérage d'épiphénomènes, de traces d'opérations discursives effectivement exploitables. Ce repérage peut rejoindre la recherche de corrélats linguistiques de fonctions discursives qui nous occupe. Toujours dans le cadre du résumé automatique, il existe également un courant qui, à partir de l'identification de marques superficielles, vise une représentation hiérarchisée de la structure des textes à différents niveaux de grain. Il s'agit du « parseur rhétorique » de Marcu (2000 ; 2001), qui construit automatiquement des arborescences représentant les segments textuels et leurs relations rhétoriques. Conçu à partir de la Rhetorical Structure Theory (Mann et Thompson, 1988), cet analyseur fait appel à des marqueurs de surface (connecteurs, ponctuation) pour identifier des unités et construire les arbres rhétoriques rendant compte des relations tant au niveau local (« clause-like units ») qu'aux niveaux supérieurs. Notons que si les procédures automatiques mises en œuvre dans l'analyseur de Marcu s'appuient à la fois sur un modèle du discours et une analyse de corpus, la méthode d'analyse du corpus elle-même reste marginale par rapport à notre préoccupation ici dans la mesure où elle fait appel à une extraction de contextes pour chaque marqueur (± 200 mots, cf. Marcu, 1997) – en d'autres termes, à une concordance étendue. Ces extraits sont analysés manuellement de manière à construire la base de données (environnement orthographique, position dans l'unité, relations rhétoriques signalées, types des unités reliées, etc.) utilisée pour mettre en place les procédures de l'analyse automatique. L'analyse de corpus n'est donc pas concernée par l'organisation discursive au-delà de chaque marqueur et de chaque relation individuelle. L'analyseur, en revanche, fait apparaître la structure rhétorique totale. La validation de ces analyses automatiques demeure toutefois problématique, puisqu'elle fait appel à la construction manuelle d'arbres rhétoriques, et n'a été effectuée que sur des textes courts.

Terminons ce rapide panorama par une application qui, plutôt qu'un « résumé », propose le repérage de zones rhétoriquement typées. La méthode de repérage de zones argumentatives développée par Teufel et Moens (1999), dite « text zoning », combine l'identification d'énoncés importants à partir de « cue phrases » avec une caractérisation rhétorique de l'énoncé identifié par rapport à la fonction communicative de l'ensemble (dans des articles scientifiques). C'est une méthode qui a été spécifiquement pensée pour l'exploitation d'articles scientifiques par des chercheurs, et qui vise à leur fournir un outil de recherche bibliographique intelligent. Elle s'appuie sur l'analyse d'un

[18] Pour le français, voir Ferret et al., 1998 ; Ferret & Grau, 2001 ; Sitbon & Bellot, 2004.

corpus de 200 articles, sur laquelle se fondent des méthodes par apprentissage permettant l'identification et l'annotation de sept types de zones telles que :

Background : arrière plan scientifique général ;

Others : descriptions neutres de travaux d'autres auteurs ;

Own : descriptions neutres de travaux de l'auteur de l'article ;

Aim : présentation de l'objectif spécifique de l'article ;

Text : indications métatextuelles concernant l'organisation de l'article (e.g. « au chapitre 1, nous introduisons… »).

L'organisation discursive, traitée principalement à travers des techniques de filtrage, est donc envisagée de façon très locale, soit pour contribuer à la sélection d'énoncés importants, soit pour caractériser des zones fonctionnelles. Les applications interactives examinées dans la section suivante l'abordent un peu différemment.

5.4.2. De l'extraction de zones de texte à la navigation sélective

De nombreux travaux récents se donnent pour objet l'élaboration de systèmes interactifs d'aide à la localisation de segments « pertinents » dans des bases documentaires ou dans des documents longs. Comme pour les applications mentionnées précédemment, le problème identifié a trait à l'exploitation de gros volumes de documents, mais plutôt que de passer par une extraction *a priori* (fondée soit sur une notion de saillance ou d'importance – résumé automatique –, soit sur une caractérisation rhétorique – text zoning), les solutions proposées sont interactives et doivent permettre à l'utilisateur de naviguer dans le ou les documents à partir de requêtes ou grâce à des procédés de visualisation sélective.

Dans le projet GEOSEM par exemple, conçu pour faciliter l'utilisation de documents géographiques par des professionnels, une requête va typiquement associer un « phénomène » avec une localisation spatiale, et/ou une localisation temporelle. Par exemple : *Retard scolaire dans l'Ouest de la France dans les années 1950.* Le système devra identifier des segments reliant les deux ou trois critères de recherche. La délimitation des segments dans lesquels des éléments textuels répondant à ces critères de recherche sont effectivement croisés va au-delà d'une simple cooccurrence : elle fait appel à la notion de cadre de discours (voir paragraphe 5.2.1) de manière à identifier des zones de textes correspondant à la portée d'une expression temporelle ou spatiale (Le Draoulec et Péry-Woodley, 2001 ; Bilhaut *et al.*, 2003).

Un deuxième type d'application a pour objectif la visualisation sélective de zones caractérisées par des marqueurs fonctionnels : la plate-forme ContextO (Minel, 2003) permet une mise en valeur visuelle d'éléments textuels associés à divers rôles discursifs, tels que des annonces thématiques (*au sujet de, à propos de, en ce qui concerne*, etc.), des énumérations, des définitions. Pour des documents longs, ce type d'application nécessite la mise en œuvre de techniques de visualisation sophistiquées (*TileBars* : Hearst, 1995 ; *typographie dynamique :* Small, 1999 ; *Perspective Wall* : Robertson et Mackinlay, 1993 *inter alia*), qui permettent une approche progressive, multi-échelle, du document[19]. Les documents sont « écrémés » de manière à ne conserver qu'une partie de l'information explicite tout en suggérant le contenu complet : l'objectif est que l'utilisateur puisse consulter le document à différents niveaux de grain, en zoomant sur des parties spécifiques sans perdre de vue les niveaux supérieurs. Ces objectifs dépendent fortement d'un travail préalable d'analyse de l'organisation discursive et de sa signalisation, y compris à des niveaux de structuration jusqu'à présent peu pris en compte par les linguistes du discours (le rôle des titres par exemple, cf. Jacques *et al.*, 2004).

[19] Cf. le projet *Visualisation dynamique de textes : extraction sélective, affichage spatial multi-échelle et observation des stratégies de lecture,* projet du programme CNRS « Société de l'information » (http://www.limsi.fr/Individu/jacquemi/COGNITIQUE02/).

5.4.3. Un objectif partagé : repérer des segments fonctionnellement « pertinents »

Un fil conducteur parcourt ces recherches à visée applicative : l'identification de segments fonctionnellement pertinents. Les plus anciennes – celles sur le résumé automatique remontent aux années 50 – partaient de l'hypothèse de l'existence d'énoncés intrinsèquement « importants » ou « saillants » qu'il fallait identifier de manière à les extraire pour construire une sorte de squelette du texte. D'abord presque exclusivement mise en œuvre par des métriques lexicales, cette approche s'enrichit ensuite à travers la recherche de segments spécialisés : présentatifs, conclusifs, ou porteurs de rôles rhétoriques beaucoup plus finement caractérisés (toujours cependant à partir d'une analyse de surface). L'analyseur rhétorique de Marcu combine sélection par saillance et caractérisation rhétorique. Les applications plus récentes font intervenir un degré croissant d'interactivité : résumé automatique à partir d'une requête (*query-biased summary*), navigation à partir d'une requête (GEOSEM), jusqu'à la navigation libre assistée par des interfaces bi- ou tri-dimensionnelles qui cherchent à rendre directement perceptibles les différents plans et niveaux de structuration des textes.

Une évolution importante – du point de vue d'une linguiste du discours – semble se dessiner au fil de ce parcours : avec la notion « d'énoncé important », dont l'identification repose principalement sur des métriques lexicales auxquelles s'ajoutent des critères positionnels et des marqueurs lexicaux, c'est une conception « plate » du texte qui transparaît (mis à part l'analyseur rhétorique de Marcu). Un pas est fait vers une plus grande prise en compte de l'organisation textuelle lorsque ces techniques sont associées à une étape d'identification des frontières entre segments « thématiques », qui conduit à situer les « énoncés importants » dans des segments. Mais même si certains algorithmes (Ferret *et al.*, 1998) autorisent des structures emboîtées (simples), on reste fondamentalement dans une vision plate, séquentielle, du texte. Sur le plan linguistique, ces techniques se rattachent à un modèle du discours assez intuitif dominé par la notion – très problématique – de thème ou topic discursif et sa manifestation à travers la cohésion lexicale. Les travaux axés sur l'identification de zones « argumentatives », d'objets textuels fonctionnels comme les définitions ou les énumérations, ou de cadres de discours, s'appuient quant à eux sur des hypothèses linguistiques spécifiques sur l'organisation des discours, mais là encore il s'agit d'identifier des « ilôts » considérés séparément. Le défi linguistique posé par la navigation et en particulier par la visualisation multi-échelle de documents est qu'il devient nécessaire non seulement d'identifier des segments présentant un intérêt discursivo-documentaire particulier, mais aussi de les situer dans la structure du document, et ce à différents niveaux de grain.

Les méthodes mises en œuvre pour cette recherche de segments pertinents associent diversement deux grands types de techniques : des techniques de filtrage, et des algorithmes de segmentation basés sur les distributions et les récurrences lexicales. Cette association constitue la première des pistes qui vont être proposées dans la section suivante, où nous chercherons à formuler pour les études de discours en corpus des directions tirant parti des expériences du TAL.

5.5. Perspectives : trois directions pour croiser discours, corpus et TAL

5.5.1. Intégration de techniques numériques

Il est intéressant de noter que si les travaux linguistiques sur le discours tendent très largement à privilégier les approches qualitatives et les faibles volumes de données, les techniques auxquelles font appel les applications du traitement automatique des langues privilégient au contraire le quantitatif. L'appel massif à des techniques numériques – métriques lexicales, calculs de poids – est bien entendu motivé par leur large couverture. Elles ont une très large applicabilité et fournissent toujours un résultat, même s'il y a bien sûr lieu d'être vigilant quant à la pertinence et le fiabilité de ce résultat. Les techniques fondées sur le repérage de marques linguistiques nécessitent quant à elles un investissement considérable pour la constitution de ressources (bases de marques, identification et représentation des contraintes d'usage) et elles ont souvent une couverture limitée. Des solutions

d'avenir sont certainement à chercher dans l'association de ces deux types de techniques. Pour les applications, des tentatives existent déjà : citons l'intégration de marques d'annonces thématiques et de techniques de segmentation automatique dans le repérage de structures thématiques proposé par Ferret *et al.* (2001). On trouve également des exemples d'association de ces techniques dans des études linguistiques sans visée applicative directe : Bestgen *et al.* (2003 ; cf. aussi Degand *et al*, 2004) font appel à des analyses de similarité lexicale (*Latent Semantic Analysis*) pour tester une hypothèse concernant la nature mono- ou polyphonique de connecteurs causaux : si les arguments d'une relation causale sont « proches », ils auront tendance à être reliés par un connecteur monophonique (pas de changement de perspective), alors que dans le cas contraire un connecteur polyphonique sera péférentiellement utilisé (changement de perspective). Méthodologiquement, cette approche leur permet de constituer des volumes d'observables qui seraient inatteignables manuellement, et de satisfaire aux exigences de reproductibilité et de quantification évoquées dans la section 5.2.

Sur un plan théorique, l'utilisation de techniques de segmentation faisant appel à des mesures de similarité lexicale trouve son fondement et sa justification dans la nécessité d'aborder conjointement, dans toute approche de l'organisation discursive, procédés lexicaux et procédés grammaticaux. Givón (1995) insiste sur l'interaction de ces processus dans la multiplicité de fonctionnements complexes (« strands of coherence ») dont la cohérence est un épiphénomène. La linguistique systémique fonctionnelle a toujours fait preuve d'une conscience aiguë de cette interaction : l'étude de la cohésion textuelle de Halliday et Hasan (1976) fait une grande place à la cohésion lexicale ; l'étude de Hoey (1991) est entièrement centrée sur le rôle des « patrons » formés par les chaînes lexicales dans la construction des textes[20] ; Martin (1992) articule son modèle de la « conjonction » avec une étude fine des procédés de cohésion lexicale ; Stuart-Smith (2001) place l'analyse rhétorique (RST) en regard de la cohésion lexicale. Etant donné la diversité des réalisations de la cohésion lexicale, cette articulation est extrêmement difficile à réaliser par des méthodes manuelles sur des volumes de données conséquents. Si les algorithmes existants sont effectivement capables de capturer – même partiellement – ces phénomènes, de nouvelles possibilités pourraient s'ouvrir à la recherche.

Pour en revenir à l'encadrement du discours, une hypothèse concernant l'interaction entre principes d'organisation est que l'importance du cadre comme principe structurant varie en fonction de la présence d'autres procédés de cohésion : plus il y a par ailleurs de « ciment cohésif » entre les propositions, moins l'introducteur de cadre joue un rôle structurant important. En revanche, s'il constitue le seul point de ralliement pour des propositions très hétérogènes, il est fortement structurant et peut même fonctionner comme un topic. Ainsi, dans l'exemple (1) l'expression temporelle cadrative *De la fin du siècle dernier jusqu'aux années 1950* est à mettre en relation avec un chaînage topical très évident : *l'école primaire […]. Elle […]. Elle […]. Elle […]*. Son rôle semble être d'ouvrir un cadre à l'intérieur duquel s'applique un critère d'interprétation restrictif. En l'absence d'un tel chaînage, on peut concevoir le paragraphe comme « étant au sujet » de la période temporelle en question. Nous cherchons à mettre en place une méthode faisant appel à la segmentation automatique pour tester cette hypothèse.

5.5.2. Profilage de textes

La question de la variation entre textes ou groupes de textes concerne le TAL comme les linguistiques de corpus. Les travaux visant le profilage automatique de textes (cf. Habert *et al.*, 2000 ; Folch *et al.*, 2000) se situent à l'intersection des deux domaines, tant par les techniques utilisées que par les objectifs : ils se fondent sur une approche inductive à la Biber (1988), à partir d'indices linguistiques (vocabulaire, catégories, patrons morpho-syntaxiques), pour construire un profil des parties d'un corpus permettant de regrouper ces parties en sous-ensembles homogènes en termes des indices utilisés[21]. Cette méthode doit aussi permettre de situer un nouveau texte par rapport aux

[20] Comme en témoigne son titre : « Patterns of Lexis in Text ». Berber Sardinha (2001) présente un algorithme de segmentation thématique basé sur les travaux de Hoey.
[21] Voir aussi Malrieu et Rastier (2001).

regroupements existants. Les auteurs envisagent et illustrent deux types d'enjeu : d'une part pour le TAL, où la fiabilité des traitements (étiquetage, parsage, recherche d'information) peut dépendre de l'homogénéité des données ; d'autre part pour l'analyse textuelle, où l'on a besoin de connaître les proximités d'un texte/corpus avec d'autres textes/corpus si l'on veut « pouvoir étendre la portée des constats effectués sur ce texte et ce corpus » (Habert *et al.*, 2000).

Toute étude en corpus suppose en effet une caractérisation de l'ensemble de textes étudiés, de manière à se donner les moyens d'évaluer la portée des descriptions ou des propositions avancées ; a fortiori, toute étude qui vise la comparaison de corpus à partir de l'hypothèse d'une différence spécifique en rapport avec le phénomène étudié doit pouvoir justifier le choix des sous-corpus comparés.

Dans notre cas, nous faisons l'hypothèse d'une variabilité dans le fonctionnement de l'encadrement temporel : en présence de relations de Narration (par opposition à d'autres relations de discours), il y aurait conflit entre le cadre temporel – référence fixe – et l'avancement narratif (voir 5.2.1), et l'expression temporelle initiale perdrait son potentiel cadratif. Cette hypothèse de conflit concerne un niveau local : la cooccurrence d'une expression temporelle susceptible d'ouvrir un cadre et d'une relation de Narration entre deux ou plusieurs propositions. Le choix de textes qui permettront de tester cette hypothèse peut se faire en fonction d'un jugement intuitif sur l'appartenance des textes candidats à un genre discursif communément dénommé narratif, et dans lesquels on peut faire l'hypothèse d'une prégnance particulière de la relation de Narration. Pour le corpus de contrôle, dans un premier temps au moins, nous pouvons nous contenter d'une différenciation binaire entre narratif et non-narratif. Reste à valider ce choix initial par l'identification d'indices linguistiques associés à la relation de Narration, et par la comparaison des fréquences d'occurrences de ces traits dans les sous-corpus destinés à la comparaison. Cela revient à procéder à un « profilage sélectif », focalisé sur les indices linguistiques qui semblent spécifiquement pertinents. Deux possibilités s'offrent à nous : sélectionner des textes dans un corpus existant « profilé » en utilisant la partition opérée par les indices en question (selon par exemple les deux pôles de la dimension « Orientation narrative *versus* non-narrative » de Biber (1988)), ou étayer un choix indépendant de corpus en comparant les parties du corpus sur ces points. C'est ce que nous avons fait, de façon très approximative et préliminaire, en comparant les fréquences des verbes de 3^{e} personne au présent, passé simple et imparfait dans les deux sous-corpus utilisés pour l'étude de l'interaction entre encadrement temporel et relation de Narration. Le tableau 5.1 présente les résultats du calcul des spécificités réalisé par le logiciel Lexico à partir de l'étiquetage fourni par Cordial Universités[22] :

Forme	Fréquence totale	Corpus « non narratif »	Corpus « narratif »
Vmip3s (présent)	*1 405*	***1 151 +E50***	*254 -E50*
Vmip3p (présent)	*1 391*	***1 305 +E50***	*86 -E50*
Vmis3s (passé simple)	755	2 -E50	**753 +E50**
Vmis3p (passé simple)	297	0 -E50	**297 +E50**
Vmii3s (imparfait)	470	47 -E50	**423 +E50**
Vmii3p (imparfait)	256	52 -E39	**204 +E39**

Tableau 5.1. *Verbes au présent et au passé dans les deux sous-corpus (calcul des spécificités)*

Ce petit bout d'analyse suggère une bonne différenciation de nos deux sous-corpus : sans entrer dans le détail du calcul des spécificités – ou formes caractéristiques – fourni par Lexico (modèle probabiliste, cf. Lebart et Salem, 1994), on notera la sur-représentation des formes au présent dans le corpus dit « non narratif » (indices de spécificité très forts : +E50), et inversement la sous-

[22] http://www.synapse-fr.com/Cordial_Analyseur/Presention_Cordial_Analyseur.htm

représentation des formes au passé. Pour le corpus dit « narratif », ce sont au contraire les formes au passé simple et à l'imparfait qui font l'objet d'un suremploi massif. Il y aurait bien sûr lieu d'enrichir cette caractérisation par l'analyse d'autres indices (cf. la distinction Narratif/Théorique de Bronckart (1985) ; les traits caractérisant la dimension « orientation narrative/non narrative » de Biber (1988)). Par ailleurs, l'éclatement des indices dû à la présence de traits non pertinents pour nous – ici la personne et le nombre – pose problème : il serait avantageux de pouvoir, pour reprendre l'expression de Habert *et al.* (2000), « manipuler des traits à géométrie variable ». Mais l'approche par profilage représente un pas vers une meilleure visibilité des principes gouvernant la sélection des corpus et leur partition en sous-corpus pour l'étude du discours. Et l'idée d'un profilage sélectif, en lien direct avec une question spécifique, est séduisante au vu de la difficulté de toute caractérisation satisfaisante d'un « type ». Reste le problème de la définition des indices linguistiques susceptibles d'être pertinents par rapport à la question à l'étude : le cas de la relation de Narration est sans doute parmi les plus simples que l'on puisse imaginer…

5.5.3. Annotation discursive

L'annotation discursive de corpus constitue un objectif central, qui rassemble les études du discours en corpus et les applications concernées par ce niveau de fonctionnement des textes. Force est de constater qu'elle est actuellement loin du stade atteint par l'annotation morpho-syntaxique et même syntaxique. Voyons l'évaluation qu'en faisaient Leech *et al.* en 1997 :

(…) we have taken a journey which began with the most secure and agreed form of annotation (grammatical word tagging), and ended with probably the last secure type of annotation (stylistic annotation), where the absence of clear and concrete criteria to identify categories inevitably leads to a considerable degree of indeterminacy. (Leech et al., 1997 : 101)

Les raisons de cet état de choses sont clairement apparentes : diversité des fonctions discursives mises en avant par les travaux théoriques, difficulté d'identifier les corrélats linguistiques de ces fonctions, caractère labile de marqueurs discursifs souvent polyfonctionnels, nature interprétative de l'analyse au niveau discursif, etc. Pour les niveaux morpho-syntaxique et syntaxique, un énorme travail collectif a été réalisé pour atteindre un accord sur les catégories à prendre en compte. La perspective d'un semblable consensus pour l'annotation discursive est à présent peu envisageable. Il est intéressant de noter le parti qu'ont pris tout récemment sur cette question les organisateurs d'un atelier sur l'annotation discursive[23] dans leur appel à communication : « the workshop is neutral as to whether consensus annotation is possible for every type of discourse phenomenon ». Mais la suite de l'appel dit en substance : que cela ne nous empêche pas de travailler… et effectivement une communauté de chercheurs se constitue depuis quelques années (Webber et Byron, 2004).

Des réalisations substantielles existent déjà, pour l'anglais surtout mais aussi pour le français. Beaucoup concernent les anaphores (Tutin, 2002, cf. Figure 5.1 ; Salmon-Alt, 2002 ; Salmon-Alt et Durand, 2003 [24] ; les travaux de l'UCREL[25]), et plus largement les procédés de cohésion (Garside *et al.*, 1997). Signalons également le corpus arboré RST (Carlson *et al.*, 2000)[26], le projet Penn Discourse TreeBank pour l'annotation de connecteurs et de leurs arguments (Miltsakaki *et al.*, 2004 a et b, voir Figure 5.2).

> <s> Si **<exp id="f35">la CGT</exp>** pousse à l'élargissement, <exp id="f36"><**ptr type="coref" src="f35"/>elle**</exp> ménage en même temps l'opinion publique. </s> <s> C'est ainsi qu'<exp id="f39"><**ptr type="coref" src="f35"/>elle**</exp> a marqué <exp

[23] *Workshop : Discourse Annotation* (en conjonction avec *ACL'04*, Barcelone, juillet 2004). Appel disponible (septembre 2004) sur la page : http://www.cllt.osu.edu/dbyron/acl04/.
[24] Projet *ANANAS* : http://www.atilf.fr/ananas/
[25] http://www.comp.lancs.ac.uk/computing/research/ucrel/annotation.html
[26] http://www.isi.edu/~marcu/discourse/

id="f40"><**ptr type="coref" src="f35"/>ses</exp>** réserves face au blocage de voies [...].
</s>

Figure 5.1 : *Annotation des anaphores et de leurs « antécédents » (Modern French Corpus including Anaphors Tagging*[27]

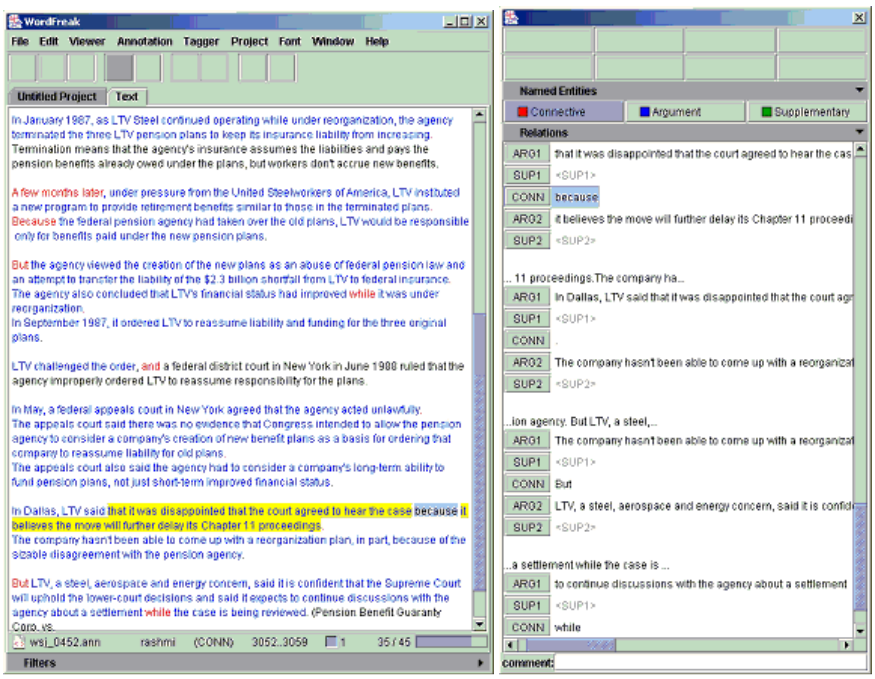

Figure 5.2 *: Penn Discourse TreeBank : outil d'annotation des connecteurs et de leurs arguments.*

Ces annotations peuvent être réalisées pour des objectifs applicatifs : on a vu dans la section précédente que plusieurs applications du TAL mettent en œuvre des procédures automatiques nécessitant des corpus d'apprentissage. D'autres visent la validation empirique d'hypothèses théoriques. C'est là que se trouve la première des clefs qui peuvent permettre aux études sur le discours de passer du stade de « précurseurs » qualitatifs au statut d'études de corpus. L'annotation systématique est en effet le préalable nécessaire à l'emploi de méthodes quantitatives. L'objectif de faire déboucher les études de l'organisation discursive sur des corpus enrichis d'annotations mis à la disposition de la communauté des linguistes entraînerait en outre les avantages suivants :

– la mise en train d'une réflexion collective plus focalisée (sur les catégories et les méthodes) visant l'élaboration de principes, sinon de normes, pour l'annotation ;

– le développement d'outils d'annotation (outils interactifs pour une annotation semi-automatique, cf. Orasan, 2004 ; Webber et Byron, 2004) et de visualisation[28] ;

– une meilleure prise en compte de l'exigence de reproductibilité ;

– et surtout l'acheminement vers un processus cumulatif qui tend à faire défaut dans ce domaine.

Ces perspectives pour une linguistique à la fois du discours et de corpus impliquent toutes trois que soient étroitement associés les trois termes du titre de ce chapitre : discours, corpus et traitements automatiques. Les études de l'organisation discursive ont besoin de s'outiller pour devenir une linguistique de corpus à part entière, mais les outils de l'établi standard ne sont guère adaptés. En revanche, les applications du TAL dans le domaine de l'exploitation de documents permettent d'imaginer des plate-formes pour l'analyse discursive : idéalement, elles combineraient des outils de

[27] Disponible par ELDA : http://www.elda.fr/catalogue/en/text/W0032.html

[28] Une réflexion de cet ordre a lieu dans le cadre du projet GEOSEM : la plate-forme Linguastream est aussi développée comme outil d'exploration, d'annotation et de visualisation de corpus pour les linguistes (cf. Bilhaut *et al.*, 2003).

profilage automatique, de filtrage, de statistiques lexicales, d'annotation interactive et de visualisation multi-échelle.

6. Bibliographie

Asher, N. (1993). *Reference to Abstract Objects in Discourse*. Dordrecht/Boston/London : Kluwer.

Berber Sardinha, T. (2001). Lexical segments in text. In M. Scott & G. Thompson (Eds.), *Patterns of Text - In honour of Michael Hoey* (pp. 213-237). Amsterdam : John Benjamins.

Bestgen, Y., Degand, L., & Spooren, W. (2003). On the use of automatic techniques to determine the semantics of connectives in large newspaper corpora: an exploratory study. In L. Lagerwerf, W. Spooren, & L. Degand (Eds.), *Determination of Information and Tenor in Texts: Multidisciplinary Approaches to Discourse 2003* (pp. 189-202). Amsterdam/Münster : Stichting Neerlandistiek & Nodus Publikationen.

Biber, D. (1988). *Variation across speech and writing*. Cambridge : Cambridge University Press.

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

Bilhaut, F., Ho-Dac, L-M., Borillo, A., Charnois, T., Enjalbert, P., Le Draoulec, A., Mathet, Y., Miguet, H., Péry-Woodley, M.-P., & Sarda, L. (2003). *Indexation discursive pour la navigation intradocumentaire : cadres temporels et spatiaux dans l'information géographique*. Actes de TALN'03, Batz-sur-Mer, 315-320.

Bronckart, J.-P. (1985). *Le fonctionnement des discours*. Neuchâtel : Delachaux et Niestlé.

Carlson, L., Marcu, D., & Okurowski, M. E. (2002). 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.

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

Charolles, M. (2002). *Les adverbiaux cadratifs : fonction et classification. Notes de synthèse*. Disponible septembre 2004 sur le site de LATTICE-CNRS : http://www.ltm.ens.fr/siteACFT/SiteLATTICEACFT4.htm

Choi, F., Wiemer-Hastings, P., & Moore, J. D. (2001). Latent semantic analysis for text segmentation. In L. Lee & D. Harman (Eds.), *2001 Conference on Empirical Methods in Natural Language Processing*, (pp. 109-117).

Cori, M., & Léon, J. (2002). La constitution du TAL. Etude historique des dénominations et des concepts. *TAL, 43*(3), 21-55.

Degand, L., Spooren, W., & Bestgen, Y. (2004). On the Use of Automatic Tools for Large-scale Semantic Analyses of Causal Connectives. In B. Webber & D. K. Byron (Eds.), *ACL 2004 Workshop on Discourse Annotation*, Barcelona, Spain, 25-32.

Desclés, J.-P., & Minel, J.-L. (2000). Résumé automatique et filtrage sémantique de textes. In J.-M. Pierrel (Ed.), *Ingénierie des Langues* (pp. 253-270). Paris: Hermès.

Ferret, O., Grau, B., & Masson, N. (1998). Thematic segmentation of texts: two methods for two kinds of texts, *ACL-COLING'98*, (pp. 392-396). Montréal, Canada.

Ferret, O., & Grau, B. (2001). Utiliser des corpus pour amorcer une analyse thématique. *TAL, 42*(2), 517-546.

Ferret, O., Grau, B., Minel, J.-L., & Porhiel, S. (2001). Repérage de structures thématiques dans des textes, *TALN 2001*, (pp. 163-172). Tours, France.

Folch, H., Heiden, S., Habert, B., Fleury, S., Illouz, G., Lafon, P., Nioche, J., & Prévost, S. (2000). TyPTex: Inductive typological text classification analysis for NLP systems tuning/evaluation. In M. Gavrilidou & G. Carayannis (Eds.), *Second International Conference on Language Resources and Evaluation*, (pp. 141-148). Athens, Greece.

Garside, R., Fligelstone, S., & Botley, S. (1997). Discourse annotation: anaphoric relations in corpora. In R. Garside, G. Leech, & A. McEnery (Eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora* (pp. 66-84). London : Addison Wesley.

Givón, T. (1995). Coherence in text vs. coherence in mind. In M. A. Gernsbacher & T. Givón (Eds.), *Coherence in Spontaneous Text* (pp. 59-115). Amsterdam : John Benjamins.

Habert, B., Illouz, G., Lafon, P., Fleury, S., Folch, H., Heiden, S., & Prévost, S. (2000). Profilage de textes : cadre de travail et expérience. In M. Rajman (Ed.), *Journées d'Analyse des Données Textuelles (JADT),*) Lausanne.

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

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

Hearst, M. (1995). TileBars: Visualization of term distribution information in full text information access, *ACM SIGCHI Conference on Human Factors in Computing Systems.* Denver, Colorado ; démo : http://www.sims.berkeley.edu/~hearst/tb-overview.html.

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

Ho-Dac, L-M., Le Draoulec, A., & Péry-Woodley, M.-P. (2001). Cohabitation des dimensions temps, espace et "phénomènes" dans un texte géographique. *Cahiers de Grammaire 26,* 125-142.

Ho-Dac, L-M. (à paraître 2005). *L'encadrement du discours : délimitation et articulation des univers de discours*. Thèse de doctorat de linguistique. Toulouse : Université de Toulouse- Le Mirail.

Hoey, M. (1991). *Patterns of Lexis in Text*. Oxford : Oxford University Press.

Hovy, E., & Lin, C.-Y. (1999). Automated Text Summarization in SUMMARIST. In I. Mani & M. T. Maybury (Eds.), *Advances in Automatic Text Summarization* (pp. 81-94). Cambridge, MA : MIT Press.

Jacques, M-P., Rebeyrolle, J., & Ho-Dac, L-M. (2004). Quelques aspects méthodologiques d'une étude de la fonction discursive des titres en corpus. *Modéliser et décrire l'organisation discursive à l'heure du document numérique. ATALA/Semaine du Document Numérique.* La Rochelle, juin 2004.

Lebart, L., & Salem, A. (1994). *Statistique textuelle*. Paris: Dunod.

Le Draoulec, A., & Péry-Woodley, M.-P. (2001). Corpus-based identification of temporal organisation in discourse. In P. Rayson, A. Wilson, T. McEnery, A. Hardie, & S. Khoja (Eds.), *Corpus Linguistics 2001* (pp. 159-166). Lancaster, UK.

Le Draoulec, A., & Péry-Woodley, M.-P. (2003). Time travel in text: temporal framing in narratives and non-narratives. In L. Lagerwerf, W. Spooren, & L. Degand (Eds.), *Determination of Information and Tenor in Texts: Multidisciplinary Approaches to Discourse 2003* (pp. 267-275). Amsterdam/Münster : Stichting Neerlandistiek & Nodus Publikationen.

Le Draoulec, A., & Péry-Woodley, M.-P. (2004). L'encadrement temporel : Interaction avec les processus de connexion. *Colloque Regards croisés sur l'unité texte*. (Exposé), Nicosie, Chypre, mars 2004.

Leech, G. (1992). Corpora and theories of linguistic performance. In J. Svartvik (Ed.), *Directions in corpus linguistics* (pp. 105-122). Berlin, New York : Mouton de Gruyter.

Leech, G., McEnery, T., & Wynne, M. (1997). Further levels of annotation. In R. Garside, G. Leech, & A. McEnery (Eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora* (pp. 85-101). London : Addison Wesley.

Malrieu, D., & Rastier, F. (2001). Genres et variations morphosyntaxiques. *TAL, 42*(2), 547-577.

Mani, I. (2001). *Automatic Summarization*. Amsterdam : John Benjamins.

Mani, I., & Maybury, M. T. (Eds.). (1999). *Advances in Automatic Text Summarization*. Cambridge, MA : MIT Press.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text,* 8(3), 243-281.

Marcu, D. (1997). The rhetorical parsing of natural language texts: *ACL/EACL'97,* (pp. 96-103). Madrid, Espagne.

Marcu, D. (2000). The Rhetorical Parsing of Unrestricted Texts: A Surface-based Approach. *Computational Linguistics, 26*(3), 395-448.

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

Martin, J. R. (1992). *English Text: System and Structure*. Amsterdam : Benjamins.

McEnery, T., & Wilson, A. (1996). *Corpus Linguistics*. Edinburgh : Edinburgh University Press.

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

Miltsakaki, E., Prasad, R., Joshi, A., & Webber, B. (2004b). The Penn Discourse TreeBank, *Language Resources and Evaluation Conference*. Lisbon, Portugal.

Minel, J.-L. (2003). *Filtrage sémantique. Du résumé automatique à la fouille de textes*. Paris : Lavoisier.

Minel, J.-L., Desclés, J.-P., Cartier, E., Crispino, G., Ben Hazez, S., & Jackiewicz, A. (2001). Résumé automatique par filtrage sémantique d'informations dans des textes. *Revue Technique et Science Informatique*, 3.

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

Morgan, J., & Sellner, M. (1980). Discourse and Linguistic Theory. In R. I. Spiro, B. C. Bertram, & W. F. Brewer (Eds.), *Theoretical Issues in the Study of Reading* . Hillsdale, N.J.: Erlbaum.

Nazarenko, A. (ce volume). Sur quelle sémantique reposent les méthodes automatiques d'accès au contenu textuel.

Orasan, C. (2004). PALinkA: A highly customisable tool for discourse annotation, *4th SIGdial Workshop on Discourse and Dialogue*, Sapporo, Japan.

Péry-Woodley, M.-P. (1993). *Les écrits dans l'apprentissage*. Paris : Hachette

Péry-Woodley, M.-P. (2001). Cohérence et relations de discours à l'écrit. Présentation. *Verbum, 23*(1), 3-9.

Robertson, G., & Mackinlay, J. D. (1993). The document lens, *ACM Symposium on User Interface Software and Technology*, (pp. 101-108) .

Salmon-Alt, S. (2002). Le projet Ananas : Annotation Anaphorique pour l'Analyse Sémantique de Corpus, *Workshop sur les Chaînes de référence et résolveurs d'anaphores, TALN 2002*. Nancy, France.

Salmon-Alt, S., & Durand, G. (2003). Représentation normalisée de corpus linguistiques : de l'étiquetage morpho-syntaxique et structurel à l'annotation anaphorique, *3èmes Journées de linguistique de corpus – JLC 2003*. Lorient, France.

Sampson, G. (2001). *Empirical Linguistics*. London & New York : Continuum.

Sitbon, L., & Bellot, P. (2004). Evaluation de méthodes de segmentation thématique linéaire non supervisées après adaptation au français, *TALN'04*, Fès.

Small, D. L. (1999). *Rethinking the Book*. Unpublished Ph.D. Dissertation. Cambridge, MA : MIT.

Stuart-Smith, V. (2001). *Rhetorical Structure Theory as a Model of Semantics: a Corpus-Based Analysis from a Systemic-Functional Perspective*. Unpublished Ph.D. Dissertation. Sidney : Macquarie University.

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

Thompson, S. A. (1985). Grammar and Written Discourse: Initial vs. Final Purpose Clauses in English. *Text, 5*(1-2), 55-84.

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

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam : John Benjamins.

Virtanen, T. (1992). *Discourse functions of adverbial placement in English*. Abo : Abo Akademi University Press.

Webber, B., & Byron, D. K. (2004). Proceedings, *ACL 2004 Workshop on Discourse Annotation*, Barcelona, Spain.