Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Discours, pragmatique et interaction
DOI 10.1051/cmlf08155

Méthodologie exploratoire outillée pour l'étude de l'organisation du discours

Lydia-Mai Ho-Dac et Marie-Paule Péry-Woodley

CLLE, Université de Toulouse
{hodac, pery}@univ-tlse2.fr

1 Introduction

Nous présentons une méthodologie exploratoire « outillée » pour aborder l'organisation du discours à travers sa manifestation dans des corpus de textes écrits. Les résultats dont nous faisons état apportent des éléments empiriques à grande échelle à même de préciser ou de mettre en cause certaines hypothèses courantes. Notre propos dans cet article est double : il s'agit de définir un couple méthodologie-objet de recherche dans le domaine des études du discours aussi bien que de fournir des apports descriptifs susceptibles de faire évoluer les conceptions des fonctionnements discursifs.

Les questions que nous abordons ont trait au texte en tant qu'objet structuré complexe, et au rôle essentiel joué par la reconnaissance/reconstruction des structures discursives dans l'interprétation, dans la création de sens. Ni herméneutique, ni étude des processus cognitifs, ce champ de recherche s'intéresse à la délimitation des unités textuelles constitutives, à l'identification des relations qu'elles entretiennent, aux modes de réalisation de ces unités et relations dans la linéarité du texte. Les notions de *cohésion* et *cohérence* sont parmi les plus significatives dans ce champ où se côtoient des études théorico-descriptives (voir par exemple Charolles, 1994 ; Jaubert, 2006) et des travaux en lien avec des préoccupations didactiques ou psycholinguistiques. Pour notre part, nous situant dans la lignée de la Systémique Fonctionnelle de Halliday (1985), nous travaillons à partir d'une hypothèse forte de prégnance des structures de haut-niveau sur les choix de formulation locaux, ce qui nous amène à nous intéresser à des textes longs, et à des objets peu étudiés, tels que ceux de la structure de document (découpage en sections et paragraphes, titres de section). Plus spécifiquement, notre recherche s'ancre dans les questionnements sur les procédés de séquentialité dans les textes, procédés qui conduisent le lecteur à voir une continuité ou au contraire une rupture entre ce qui vient d'être lu et ce qui est en train de l'être.

2 Séquentialité et linéarité

La question centrale, posée par de nombreux linguistes du discours, est celle de l'articulation entre l'inéluctable linéarité des textes et la non-linéarité des modèles mentaux qui en sont l'origine et la destination. Les unités textuelles, de taille et de complexité limitées par les potentiels cognitifs, s'enchaînent au fil du texte alors que ce qui est exprimé/interprété ne correspond nécessairement ni à un découpage en phrases ni à une expression séquentielle. Cette problématique fondamentale conduit à poser deux questions ; la première concerne la segmentation, processus à deux faces : le fait que les unités *se regroupent* en fonction d'un critère particulier *découpe* le texte en segments, signalés ou non par des « marqueurs de segmentation » ; la seconde a trait aux relations – y compris relations de hiérarchisation – entre segments. On peut envisager le discours-texte, dans sa séquentialité, comme « une alternance périodique de zones de transition et de continuation » (Goutsos, 1996 : 501). Goutsos propose un modèle de la structure du discours à trois niveaux : un niveau cognitif, où la représentation mentale du scripteur est structurée en termes des deux stratégies de base : continuité *vs.* discontinuité ; un niveau linguistique concerné par les procédés à disposition du scripteur pour réaliser ces stratégies, lui permettant de construire, au niveau textuel, cette alternance de zones de continuation et de transition.

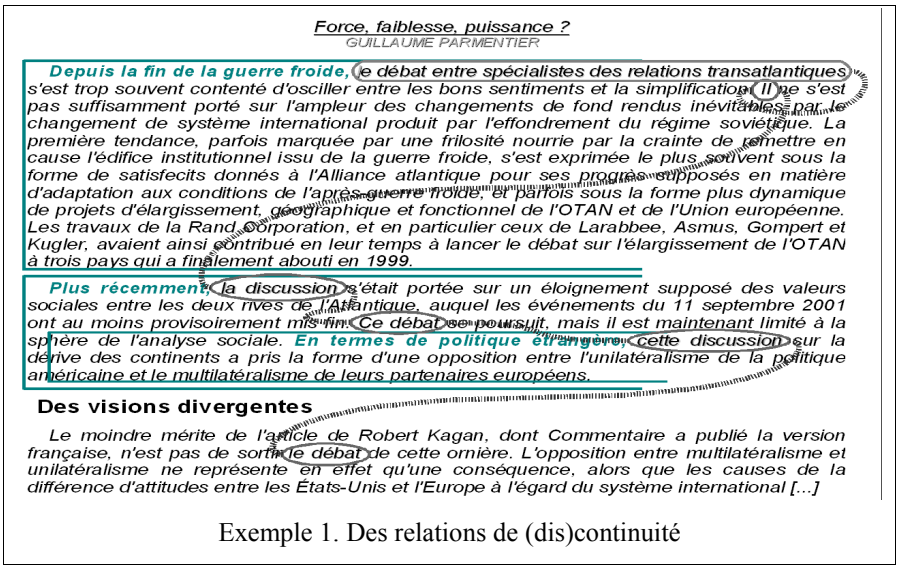
Selon nous, d'autres modes de structuration doivent être pris en compte pour rendre compte de la séquentialité du discours, en lien avec la nature du critère de regroupement des séquences. Les critères de regroupement peuvent se rapporter à « l'à propos » du discours (unités partageant un même référent, des

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Discours, pragmatique et interaction
DOI 10.1051/cmlf08155

circonstances similaires) ou à son organisation (unités participant à une même articulation rhétorique ou appartenant à une certaine structure de document, cf. le cas des structures énumératives). Dans le même ordre d'idée, la systémique fonctionnelle articule les composantes idéationnelle et textuelle. Nous ajoutons à cela un autre aspect organisationnel en distinguant le niveau *figure* – expression de « l'à propos » – du niveau *fond* – mise en place du décor[1] (circonstances, points de vue, cadre).

L'exemple 1 présente le début d'un article de géopolitique intitulé « Force, faiblesse, puissance ? ». Dans la première section (d'introduction), l'auteur recourt à une stratégie qui consiste à tisser l'ensemble autour de la notion de débat (continuité de la figure) et à répartir les contenus en deux périodes pertinentes par rapport au thème (discontinuité du fond) : il s'agit d'examiner la nature du débat entre spécialistes des relations transatlantiques dans la période suivant la fin de la guerre froide et dans la période actuelle.



Force, faiblesse, puissance ?

GUILLAUME PARMENTIER

Depuis la fin de la guerre froide, le débat entre spécialistes des relations transatlantiques s'est trop souvent contenté d'osciller entre les bons sentiments et la simplification. Il ne s'est pas suffisamment porté sur l'ampleur des changements de fond rendus inévitables par le changement de système international produit par l'effondrement du régime soviétique. La première tendance, parfois marquée par une frilosité nourrie par la crainte de remettre en cause l'édifice institutionnel issu de la guerre froide, s'est exprimée le plus souvent sous la forme de satisfecits donnés à l'Alliance atlantique pour ses progrès supposés en matière d'adaptation aux conditions de l'après-guerre froide, et parfois sous la forme plus dynamique de projets d'élargissement, géographique et fonctionnel de l'OTAN et de l'Union européenne. Les travaux de la Rand Corporation, et en particulier ceux de Larabbee, Asmus, Gompert et Kugler, avaient ainsi contribué en leur temps à lancer le débat sur l'élargissement de l'OTAN à trois pays qui a finalement abouti en 1999.

Plus récemment, la discussion s'était portée sur un éloignement supposé des valeurs sociales entre les deux rives de l'Atlantique, auquel les événements du 11 septembre 2001 ont au moins provisoirement mis fin. Ce débat se poursuit, mais il est maintenant limité à la sphère de l'analyse sociale. **En termes de politique étrangère**, cette discussion sur la dérive des continents a pris la forme d'une opposition entre l'unilatéralisme de la politique américaine et le multilatéralisme de leurs partenaires européens.

Des visions divergentes

Le moindre mérite de l'article de Robert Kagan, dont Commentaire a publié la version française, n'est pas de sortir le débat de cette ornière. L'opposition entre multilatéralisme et unilatéralisme ne représente en effet qu'une conséquence, alors que les causes de la différence d'attitudes entre les États-Unis et l'Europe à l'égard du système international [...]

Exemple 1. Des relations de (dis)continuité

Pour réaliser ces stratégies, l'auteur fait appel essentiellement à deux procédés linguistiques : les chaînes topicales et les cadres de discours. Pour le premier de ces procédés, les expressions co-référentielles participent à la continuité topicale du passage ; pour le second, ce sont les adverbiaux temporels qui indiquent la transition d'un segment à l'autre (une discontinuité), chacun étant homogène quant à la localisation temporelle des propos.

Il apparaît clairement dans l'exemple 1 que le fonctionnement de ces indices n'est pas limité à un niveau purement local, interphrastique. Les deux adverbiaux temporels ne font pas que signaler une discontinuité, ils participent également à la structuration globale du passage en étendant leur portée au-delà de la phrase d'accueil pour « indexer »[2] les contenus selon deux localisations temporelles différentes : *depuis la fin de la guerre froide* et *plus récemment*. De même, les expressions co-référentielles ne se contentent pas de relier leur phrase d'accueil au discours antérieur. Leur répartition au fil du texte, ajoutée à leur positionnement en initiale de paragraphe à la suite des adverbiaux temporels, leur confère un rôle à un niveau plus global : *le débat entre...* constitue le topique global de cette section d'introduction, et certainement du texte, la notion de débat étant reprise au début de la section suivante.

Pour résumer, nous proposons une représentation de la structure globale de l'exemple 1 sous la forme d'une énumération qui accorde le même poids aux adverbiaux temporels à l'initiale de paragraphe qu'au titre de section (Figure 1). Cette représentation sous forme de structure énumérative établit une analogie entre le phénomène de discontinuité et le passage d'un item au suivant dans une énumération. C'est une façon de souligner que la discontinuité s'apparente davantage à un déplacement qu'à une rupture. En effet, étant donné que différents niveaux d'organisation sont à l'œuvre, il est fréquent que la discontinuité

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Discours, pragmatique et interaction
DOI 10.1051/cmlf08155

concerne uniquement un niveau. Ainsi, dans l'exemple 1, le cadre temporel (niveau fond) se déplace tandis que la continuité référentielle (niveau figure) demeure.

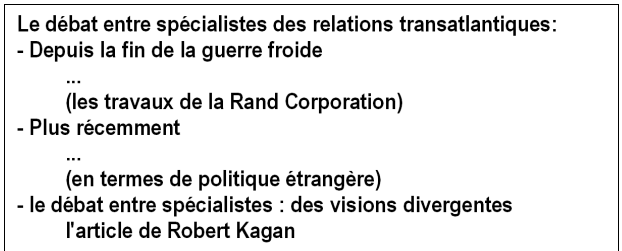


Figure 1. Représentation de la structure de l'exemple 1

L'approche de l'organisation du discours proposée ici vise donc essentiellement à mettre au jour les indices qui signalent la séquentialité du discours. En d'autres termes, nous cherchons à identifier les éléments textuels qui, comme les expressions co-référentielles et les adverbiaux temporels de notre exemple, participent à instruire d'une relation de continuité ou de discontinuité entre deux segments. Nous espérons ainsi nous donner des points d'entrée formels pour aborder et définir les structures de discours que nous étudions, points d'entrée qui pourront être exploités par la suite dans des applications de traitement automatique des langues ou dans des analyses automatisées.

2.1 Des configurations d'indices pour signaler la (dis)continuité

Nous proposons d'envisager le signalement des structures discursives en termes de « configurations d'indices » plutôt qu'en termes de marqueurs de segmentation (Bestgen & Vonk, 2000 ; Piérard & Bestgen, 2006), marqueurs dits « discursifs », ou « organisateurs textuels » (Adam & Revaz, 1989, Schneuwly *et al.*, 1989). Cette notion de configuration est en lien étroit avec notre approche en corpus, et émerge en quelque sorte des données elles-mêmes. Notre objectif est en effet de découvrir des procédés de signalement, et non d'examiner le fonctionnement d'éléments définis d'emblée comme des marqueurs. Cette approche nous amène à observer que l'identification d'une structure discursive repose rarement sur un élément lexical isolé, mais plutôt sur l'influence conjointe de facteurs multiples de nature parfois autre que lexicale (tels le type de texte, la structure de document, la position de la portion de texte en cours de lecture dans la hiérarchie du document, etc.).

La représentation proposée dans la figure 1 ne se fonde pas sur les seules occurrences d'expressions co-référentielles et d'adverbiaux temporels. Leur positionnement dans le texte et leur interaction jouent un rôle essentiel dans cette interprétation. Le fait que chaque paragraphe de la première section commence par un adverbial temporel confère à celle-ci sa structure temporelle. Le fait que chaque paragraphe (et le texte lui-même) commence par une référence au débat « donne » à l'extrait son « à propos » global.

La prise en compte du positionnement des éléments lexicaux apparaît comme essentielle pour la découverte des configurations d'indices. Enkvist (1985) défend une vision paramétrique de l'ordre des mots (*i.e.* le positionnement des éléments lexicaux), qu'il voit comme relevant d'un compromis entre différentes forces. La gestion de ces différentes forces se situe tant au niveau local de la construction phrastique qu'au niveau global de la construction du texte, et elle est informée, ou sur-déterminée (cf. Adam & Revaz, 1989), par des paramètres liés au type de texte. Enkvist suggérait que dans l'oral spontané la structure d'information l'emportait régulièrement sur la « syntaxe canonique ». Virtanen (1992), qui poursuit les travaux d'Enkvist, défend l'hypothèse que le placement des adverbiaux, notamment en initiale de phrase ou de paragraphe, dépend de stratégies textuelles et non phrastiques. La linguistique systémique fonctionnelle propose un modèle tripartite organisant ces différentes forces à travers les métafonctions idéationnelle, textuelle et interpersonnelle. Rapidement, nous nous sommes rendu compte que les éléments lexicaux pris dans des configurations signalant la séquentialité du discours participaient conjointement aux composantes idéationnelle et textuelle. Du point de vue idéationnel, les

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Discours, pragmatique et interaction
DOI 10.1051/cmlf08155

expressions sujets ou les adverbiaux détachés en initiale de phrase expriment ce à propos de quoi on parle et circonscrivent cet à propos par rapport à des localisations temporelles, spatiales, notionnelles, énonciatives, etc. D'un point de vue textuel, ces expressions organisent ce à propos de quoi on parle en participant à la répartition des contenus dans des segments de texte. Ce fonctionnement textuel nous apparaît comme fortement lié à la position initiale de tels éléments.

2.2 Un point d'entrée dans l'organisation du discours : la position initiale

Dans un article intitulé *Point of departure : Cognitive aspects of sentence-initial adverbials*, Virtanen (2004) définit la position initiale en termes de trois fonctions : lier le discours antérieur au discours à venir ; orienter l'interprétation des segments à venir ; conférer aux éléments initiaux une certaine saillance dans la construction de la représentation mentale, ces éléments étant alors associés à l'information « cruciale ». Cette dernière idée d'« information cruciale en premier » permet de rendre compte de fonctionnements différents – et apparemment contradictoires – de la position initiale en discours, car ce qui est « le plus important » peut varier selon la situation. Il est parfois plus important d'insister sur la relation de continuité à établir entre deux segments et parfois plus important de signaler une discontinuité dans le discours (changement de topique, déplacement de cadre, articulation rhétorique, etc.). Dans une vision phrastique de la structure d'information, la position initiale est associée à l'information donnée selon l'hypothèse que, dans les cas non marqués, l'information donnée (plus prévisible et donc moins lourde à traiter) est placée avant l'information nouvelle (moins prévisible et donc plus lourde à traiter). Dans une conception discursive, comme celle de Virtanen (1992, 2004), la position initiale est associée à l'information jugée nécessaire à l'interprétation du segment de texte qui suit. Cette information nécessaire peut correspondre à une information donnée en cas de continuité ou à une information plus ou moins nouvelle en cas de discontinuité.

Dans l'exemple 1, tout ce qui apparaît en initiale de paragraphe « continue » sur le paragraphe entier. De fait, ce qui y est exprimé peut être considéré comme important, et ce sur deux plans : celui de l'à propos (composante idéationnelle) et celui de l'organisation générale du texte i.e. de la répartition des contenus (composante textuelle). Au début de la section intitulée « des visions divergentes », on observe une discontinuité avec l'introduction d'un nouveau référent (l'article de R. Kagan) qui pousse le lecteur à attribuer à ce référent une certaine saillance dans la section ainsi introduite.

3 Une démarche exploratoire en corpus : des observables à l'objet

Au-delà d'un simple exposé de notre méthodologie, nous tentons d'expliciter dans cette section l'articulation entre les stratégies rendues possibles par les outils utilisés, les observables sélectionnés, et la définition de notre objet d'étude.

Péry-Woodley (2005) fait état des difficultés des études sur le discours à reprendre à leur compte les principes méthodologiques des linguistiques de corpus. Si la plupart des travaux se fondent sur des exemples attestés, rares sont ceux qui mettent en œuvre des méthodes permettant d'évaluer l'importance quantitative des structures étudiées. Pour les linguistiques de corpus, la quantification est la condition de toute tentative de comparaison ou de généralisation (Biber, 1998). Les méthodes principalement qualitatives qui dominent en linguistique du discours prêtent également le flanc à la critique pour leur approche subjective (« analyst-dependent ») qui fait obstacle à leur reproductibilité et donc à leur validation (Bestgen *et al.*, 2006). Il demeure donc souvent difficile de déterminer si l'usage mis en évidence est un effet du hasard ou s'il est représentatif, quel est son statut par rapport à d'autres usages, s'il est propre à un texte ou à un corpus spécifique. Nous cherchons pour notre part à nous situer pleinement dans une linguistique de corpus, nécessairement outillée pour tirer parti de gros volumes de textes, pensée pour prendre en compte la variation liée au type de corpus, associant une approche quantitative permettant de mettre en œuvre des stratégies de découverte et une démarche qualitative fondée sur un retour aux textes. Au cœur de cette méthodologie se trouvent l'analyse automatique et l'annotation des corpus, à travers lesquelles se réalise l'articulation entre modèle et corpus.

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Discours, pragmatique et interaction
DOI 10.1051/cmlf08155

3.1 Position initiale : caractériser la zone préverbale

La position initiale correspond dans notre analyse à toute la zone préverbale. De nombreuses études associent un fonctionnement discursif à certains éléments pouvant apparaître dans cette zone. Les adverbiaux circonstanciels (notamment les adverbiaux de temps) sont ainsi souvent associés à la délimitation de segments de texte (Virtanen, 1992 ; Bestgen & Vonk, 2000 ; Piérard & Bestgen, 2006), de même que les expressions co-référentielles redondantes à des effets de transition (Schnedecker, 2003). À l'inverse, les pronoms semblent instruire d'une relation de continuité topicale (au niveau figure). Ces travaux nous intéressent dans la mesure où ils suggèrent que certains éléments possèdent une forte capacité à marquer une relation de (dis)continuité. Nous nous en distinguons toutefois en fondant notre analyse sur l'ensemble des éléments présents en position préverbale, de manière à mettre en œuvre une approche guidée par les données, dans l'idée que des fonctionnements discursifs encore peu connus peuvent émerger[3].

Les fondements de notre méthode d'analyse peuvent se résumer en trois assertions : les corpus doivent être constitués de textes où l'organisation discursive et sa signalisation sont nécessaires (textes longs) ; la démarche quantitative à partir d'une annotation systématique doit permettre la mise en œuvre de stratégies de découverte ; la signalisation du discours est envisagée comme la résultante d'une interaction entre différents modes de structuration et concerne également des traits non linguistiques.

La première assertion a trait à l'impact de la longueur sur l'organisation discursive : là où des textes courts peuvent fonctionner sur la seule base de la continuité référentielle, des textes plus longs, surtout s'ils sont non-narratifs et donc non structurés en termes de succession d'événements par défaut, nécessitent d'autres formes d'organisation. La seconde assertion est intimement liée à la troisième : partant de l'hypothèse que la signalisation de l'organisation du discours est réalisée par des configurations d'indices plutôt que par des « marqueurs » spécifiques, nous cherchons à faire « jouer » les différents indices les uns par rapport aux autres de manière à mesurer leur impact. Il s'agit donc, à partir de nos analyses, d'expérimenter différents rapprochements d'indices. Parmi ces indices figurent le type de texte et la position textuelle. La partition du corpus en trois types de texte est présentée dans la section 3.2.1. Pour la position textuelle, trois niveaux issus de la structure du document sont distingués : S1 = première phrase d'une section ; P1 = première phrase d'un paragraphe ; P2 = phrases à l'intérieur d'un paragraphe. À partir de ces indices, nous allons explorer l'impact, pour le signalement d'une continuité ou d'une discontinuité, de la cooccurrence d'un élément A avec un élément B dans une position textuelle donnée dans un type de texte particulier.

3.2 Mise en œuvre

3.2.1 Constitution du corpus

L'hypothèse de variation liée au type de texte est prise en compte à travers une partition de notre corpus de textes « expositifs » en trois sous-corpus qui se différencient en termes de contenu thématique et d'organisation rhétorique :

- ATLAS (~205 000 occurrences) : 3 textes de géographie sociale ;
- GEOPO (~250 000 occurrences) : 32 textes concernant des problèmes de géopolitique actuelle ;
- PEOPL (~220 000 occurrences) : 30 portraits de personnages célèbres.

Les 3 textes d'ATLAS, beaucoup plus longs que ceux des deux autres sous-corpus, se caractérisent par la fréquence des localisations spatiales et temporelles qui campent le décor, organisant sur cette base de longs segments dépourvus de continuité topicale forte. Les textes de PEOPL, au contraire, se structurent autour d'une continuité topicale évidente, concernant la même entité du début à la fin. Ces textes présentent également une forte structuration temporelle. Les textes de GEOPO, plus difficiles à caractériser rapidement, sont pluriréférentiels et présentent occasionnellement et par endroits seulement une organisation spatiale ou temporelle.

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Discours, pragmatique et interaction
DOI 10.1051/cmlf08155

3.2.2 Choix et procédures d'analyse et d'annotation

Une fois notre corpus construit, l'étape de constitution des observables consiste à analyser les éléments composant la zone préverbale. L'extraction et la caractérisation de nos observables se basent sur une analyse automatique réalisée par l'analyseur Syntex (Bourigault & Fabre, 2000 ; Bourigault, 2007), qui fournit une lemmatisation et un étiquetage morpho-syntaxique, ainsi que les relations de dépendance (gouverneur/dépendant) entre les différentes unités[4]. Cette première étape d'analyse se veut systématique et exhaustive de manière à permettre des procédures de découverte : tous les éléments présents en zone préverbale ont été identifiés et caractérisés. Nous distinguons des connecteurs simples, des éléments détachés et des sujets grammaticaux. Les éléments détachés répondent à une caractérisation morpho-syntaxique (adverbe, syntagme prépositionnel, etc.) et fonctionnelle (circonstants, appositions, modalisateur d'énonciation, etc.). À la caractérisation morpho-syntaxique des sujets grammaticaux (forme pronominale, définie, etc.) s'ajoutent des informations liées à la notion de co-référence : la tête lexicale du syntagme est-elle une reprise d'un élément présent dans le discours précédent (section en cours), et quelle est la taille du syntagme en nombre de mots (jusqu'à trois mots, un syntagme est dit « court », au-delà, il est dit « long »). La figure 2 montre l'exemple 1 enrichi de ces annotations.

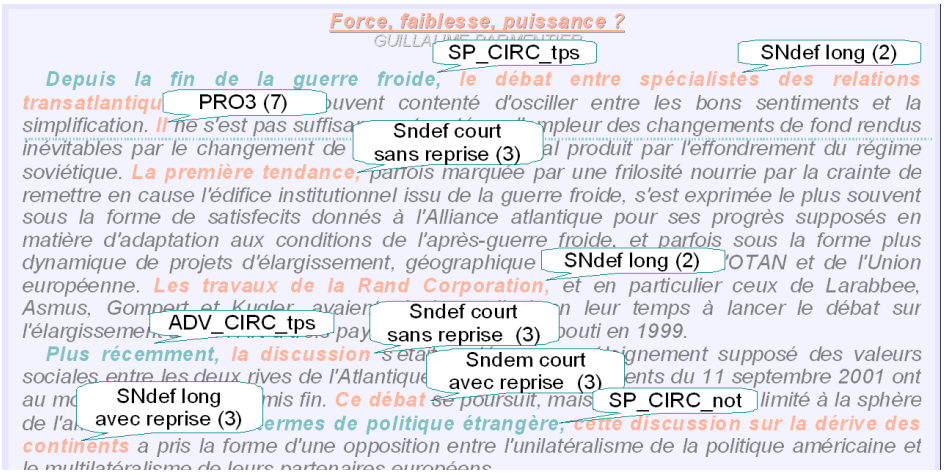


Figure 2. Illustration des annotations générées pour l'exemple 1

Les analyses effectuées permettent de générer des annotations en lien plus immédiat avec nos hypothèses sur la continuité/discontinuité. Les phrases commençant par un connecteur simple, une apposition ou un sujet de forme pronominale peuvent constituer des indices de continuité. À l'inverse, la présence d'un circonstant détaché en initiale est associée à l'indication d'une discontinuité. Il en va de même pour les sujets estimés peu accessibles selon les différentes échelles d'accessibilité (Ariel, 1990 ; Gundel *et al.*, 2000), comme les SN longs et/ou indéfinis. Les SN courts ainsi que les noms propres, les SN définis ou démonstratifs, associés à un degré d'accessibilité plus élevé, sont quant à eux susceptibles d'indiquer une continuité au niveau figure. Ces expressions peuvent toutefois faire plus qu'instruire d'une simple continuité référentielle : les cas de reprise lexicale permettraient d'assurer une continuité référentielle dans des situations « à risque » (cf. notamment Schnedecker, 2003). Dans notre corpus, elles semblent constituer des indices privilégiés de continuité au niveau figure en présence d'une discontinuité au niveau fond (signalée par exemple par un changement de paragraphe ou la présence d'adverbiaux circonstanciels détachés).

À partir de ces données, nous extrayons des configurations que nous examinons en fonction du type de texte et de la position textuelle. Cette extraction s'effectue en mesurant les variations pour les différents éléments lexicaux entre i) les distributions pour chaque corpus comparées avec les distributions globales ; ii) les distributions dans chaque position textuelle comparées avec les distributions globales. Nous pouvons ainsi faire « émerger » des configurations pertinentes.

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Discours, pragmatique et interaction
DOI 10.1051/cmlf08155

4 Des données aux résultats

Après analyse de tout le corpus, nous avons 23 217 phrases annotées dont 30% (7 021) présentent un élément détaché en initiale. Chaque sous-corpus contient entre 7 500 et 8 000 phrases. Le tableau 1 résume la composition des deux principaux constituants de la zone préverbale : éléments détachés en initiale et sujets grammaticaux. Les constructions spéciales repérées sont des constructions impersonnelles, des clivées, des présentationnelles, des inversées, etc. où le sujet grammatical ne peut constituer le topique du discours. En regard de chacun de ces éléments figure sa position textuelle associée (+) selon le test de l'écart réduit (z score). Par exemple, les sujets de forme pronom personnel de 3^{e} personne (désormais PRO3) sont associés à la position P2, c'est-à-dire aux phrases intraparagraphiques, avec un écart réduit équivalant à +9 (z = +9). Puisqu'il y a une probabilité de 0,01 (*i.e.* une chance sur 100) d'atteindre un écart réduit (positif ou négatif) de 2,5, nous considérons tout écart au-delà de ce seuil comme significatif et donc que la répartition des PRO3 selon les positions textuelles n'est pas aléatoire mais contrainte, peut-être par des phénomènes linguistiques.

Élément non sujet	**Fréquence**	**% de phrases**	**Position textuelle**		**Sous-corpus**	
			+	**z (• =+0,5)**	**+**	**z (• =+0,5)**
Connecteur simple	2220	9,6	P2	••••••••••	PEOPL GEOPO	••••••••••••••••••••••••• ••••••
Élément détaché en initiale	7021	30,2	P1	••••••••	GEOPO	••••••••••••••••
Circonstant	5406	23,3	S1 P1	••••••• ••••••••	ATLAS	••••••••••
... temporel	... 1641	7,1	S1	••••••••••		
... spatial	... 564	2,4	P1	•••••••	ATLAS	••••••••••••••••••••••••••
... notionnel	... 797	3,4			GEOPO	•••••••••
apposition	1055	4,5	S1	•••••••	PEOPL	•••••••••
Sujet grammatical	**23217**	**100**				
Construction spéciale	3857	16,6			PEOPL	••••••••••••••••••••
SN défini	10143	43,7	S1 P1	•••••••••••••••••• ••••••••••••••••••	GEOPO ATLAS	••••••••••••••••••• •••••••••••••••••••
Pronom de 3e pers. - PRO3	2221	9,6	P2	•••••••••••••••	PEOPL	••••••••••••••••••••••••••••••
Nom propre	1592	6,8	S1	•••••••	PEOPL	••••••••••••••••••••••••...+22
SN démonstratif	1538	6,6			ATLAS	•••••••
SN indéfini	1447	6,2				
SN défini court	4111	17,7	P2	••••••••	PEOPL	••••••••••
SN avec reprise	6022	26,0	P1	•••••••••••••••••••	ATLAS	•••••••••

S1 = premières phrases de section, P1 = premières phrases de paragraphe, P2 = intérieur de paragraphe

Tableau 1. Répartition et variations des éléments en position initiale

Les cases vides du tableau signifient qu'aucun écart significatif positif n'est observé pour cet élément, pour aucune des trois positions textuelles ou aucun des sous-corpus. Ainsi, la fréquence des circonstants notionnels et des SN indéfinis ne semble pas dépendre de la position textuelle de leur phrase d'accueil. De même, la présence de circonstants temporels semble être indépendante du type de texte.

Ce tableau résume les principales associations d'éléments en zone préverbale avec certaines positions textuelles et certains sous-corpus (seuls les écarts réduits positifs significatifs sont indiqués). La section 4.1 traite des corrélations identifiées entre certains éléments et certaines positions textuelles. Les résultats concernant les associations avec les différents sous-corpus sont commentés dans la section 4.2 qui s'intéresse aux variations des éléments en initiale selon le type de texte.

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Discours, pragmatique et interaction
DOI 10.1051/cmlf08155

Élément en position initiale		ATLAS			GEOPO			PEOPL		
		z-	0	z+	z-	0	z+	z-	0	z+
Circonstant spatial	S1	•••••			••				•	
	P1		•••••			•			••	
	P2	•						•		
Apposition	S1		•••••			•••••			••••	
	P1	•			••••				•••••	
	P2	••						••••		
SN défini	S1		••••••••••			••••••••••			••••••	
	P1		••••••••••			••••••••••			••••	
	P2	••••••••••••			•••••••••			••••		
Nom propre	S1		•			••			••••••••••	
	P1		•						•••••	
	P2	•				•		••••••		

S1 = premières phrases de section, P1 = premières phrases de paragraphe, P2 = intérieur de paragraphe

Tableau 1. Écarts réduits selon la position textuelle dans les sous-corpus (• = 0,5)

4.1 Corréler des éléments de la zone préverbale et des positions textuelles

Nombre de données présentées dans le tableau 1 vont dans le sens des hypothèses générales associées aux différents éléments. Ainsi, les connecteurs et les pronoms apparaissent significativement plus à l'intérieur d'un paragraphe alors que les éléments détachés (en général) et les sujets non pronominaux se trouvent préférentiellement en initiale de paragraphe ou de section, exception faite des SN démonstratifs. La différence observée au niveau des circonstants temporels et des circonstants spatiaux peut se lire de la façon suivante : les localisations temporelles jouent à un niveau d'organisation plus global que les localisations spatiales. S'il est fréquent de trouver une section encadrée par des localisations temporelles, les localisations spatiales semblent encadrer de préférence des unités de niveau inférieur : les paragraphes (voir l'exemple 2 infra).

Certains résultats sont plus inattendus, telle l'association entre les appositions (peu étudiées sur le plan discursif) et les débuts de section. Nous proposons en 4.2 une interprétation de ce résultat. On peut également s'étonner de la fréquence des SN indéfinis sujets, et du fait que leur utilisation ne semble absolument pas dépendre de la structure logique du texte, alors qu'on s'attendrait à les trouver en initiale de section ou de paragraphe pour introduire les nouveaux référents.

Les données concernant les sujets de forme courte ou avec reprise lexicale méritent également qu'on s'y arrête. Les SN définis courts (tout comme les démonstratifs) apparaissent significativement plus en position intraparagraphique qu'ailleurs, alors que les SN longs sont significativement associés aux positions S1 et P1. Les cas de reprise lexicale dans les SN définis, démonstratifs et les noms propres s'associent très préférentiellement à l'initiale de paragraphe (z = +11). Ces résultats vont dans le sens de travaux récents sur les redénominations (Schnedecker, 2003) : l'auteur aurait recours à une redénomination pour maintenir une continuité dans des situations où celle-ci est en danger (début de paragraphe, ou après un changement de circonstance, comme dans l'exemple 1.

4.2 Valider la caractérisation des sous-corpus

Nous avons vu dans la section précédente que certains éléments de la zone préverbale sont associés de façon significative à certaines positions textuelles. Lorsque les mêmes mesures sont reproduites cette fois à l'intérieur de chaque sous-corpus, les variations dans les écarts constatés suggèrent des différences marquées dans l'organisation globale des textes de chaque sous-corpus (cf. tableau 2).

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Discours, pragmatique et interaction
DOI 10.1051/cmlf08155

Tout d'abord, on remarque que certaines corrélations sont communes aux trois sous-corpus. Ainsi, dans les trois sous-corpus, les connecteurs simples, les SN courts et les PRO3 sont associés à P2, les cas de reprises lexicales à P1, et les circonstants temporels à S1. Les éléments pour lesquels les variations sont inexistantes ou limitées (SN indéfinis, circonstants notionnels et constructions spéciales) restent neutres dans les trois sous-corpus. Le tableau 2 représente les degrés d'association (écarts réduits) pertinents de certains éléments de la zone préverbale avec différentes positions textuelles dans chaque sous-corpus. Ces résultats confortent notre classification des sous-corpus.

Du côté de PEOPL, les portraits s'organisent principalement autour d'une continuité topicale globale qui se traduit par une très forte proportion de PRO3 (cf. tableau 1) et de noms propres avec reprise en P1 comme en P2. Les circonstants temporels ont comme dans GEOPO et ATLAS une fréquence significativement plus forte en S1 (l'association affichée dans le tableau 1 concerne les trois sous-corpus confondus). La position P1 est ici significativement associée à la présence d'appositions, alors que celles-ci se situent d'une manière générale préférentiellement en position S1 (cf. tableau 1). Nous revenons sur le cas des appositions dans la section suivante pour tenter d'expliquer cette particularité.

On s'aperçoit qu'ATLAS est seul responsable de la prédilection des circonstants spatiaux pour la position P1 (z = +3,5 pour les circonstants spatiaux en P1, cf. tableau 1), ce qu'illustre l'exemple 2.

Un élève étranger sur trois en Île-de-France [titre niveau 3]
Plus du tiers (38%) des élèves étrangers vit en Île-de-France, [...]
Les académies de Lyon et Grenoble accueillent près de 150 000 enfants étrangers, [...]. L'Est, de Montbéliard à Strasbourg et Nancy, forme le troisième ensemble à forte proportion d'élèves étrangers : [...]
Dans les académies méditerranéennes qui comptent également une centaine de milliers d'élèves étrangers, les taux d'élèves étrangers avoisinent 10% dans [...].
Dans le Nord, les élèves de nationalité étrangère sont surtout nombreux dans [...].
À l'ouest d'une ligne Le Havre-Montpellier, la population scolaire comprend généralement moins de [...].

Exemple 2. Énumération spatiale marquée par des circonstants en P1 dans ATLAS

ATLAS est d'ailleurs le seul sous-corpus à associer significativement les circonstants à P1 et non à S1, position préférée pour les circonstants dans GEOPO comme dans PEOPL. Du point de vue des continuités topicales, ATLAS semble recourir plus que les autres sous-corpus aux reprises lexicales, qui réalisent des progressions thématiques à thèmes dérivés. Les continuités se construisent essentiellement par des processus de cohésion lexicale, comme on le voit dans l'exemple 2.

GEOPO ne montre pas de spécificités claires si ce n'est des variations plus importantes concernant l'usage des SN longs (en S1 et P1) *vs.* courts (en P2), ce que l'on peut observer dans l'exemple 1. Ce sous-corpus rassemble des textes sans doute plus hétérogènes (présentant une variation intra-textuelle plus grande) que les textes d'ATLAS ou PEOPL. On y repère des zones construites autour de continuités topicales fortes comme dans PEOPL et d'autres autour d'« énumérations circonstancielles » (surtout temporelles) comme dans ATLAS. L'exemple 1 montre précisément ce type de zones construites à la fois autour d'une continuité topicale forte et d'une énumération temporelle, où l'on note l'utilisation de SN définis longs en S1 *vs.* plus courts en P1 et P2.

4.3 Évaluer le pouvoir structurant des « marqueurs »

Notre méthodologie nous permet également d'évaluer le pouvoir structurant des différents indices impliqués dans une structure discursive. Cette évaluation du pouvoir structurant de certains éléments peut bien entendu se circonscrire à un type de texte particulier. Pour cette évaluation, nous comparons la composition de la zone préverbale des phrases contenant un indice spécifique à celle de la zone préverbale des phrases ne contenant pas l'indice à l'étude. Notre interprétation des résultats suit le principe

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Discours, pragmatique et interaction
DOI 10.1051/cmlf08155

suivant : si la présence d'un indice dans une phrase change la donne par rapport au modèle général fourni par le tableau 1, alors son rôle dans le signalement de l'organisation textuelle est à creuser et à analyser par des observations plus qualitatives. Sinon, on peut supposer que l'indice en question n'a pas de pourvoir structurant en soi et que ce sont les configurations dans lesquelles il apparaît qui lui confèrent un semblant de pouvoir.

Pour illustrer cette analyse, prenons l'exemple des appositions qui sont fortement associées aux positions S1 ou P1 dans les trois sous-corpus. Cette étape de l'analyse pose ainsi la question du rôle discursif des appositions en dehors des initiales de paragraphes ou de section *i.e.* à l'intérieur des paragraphes.

Dans l'ensemble de notre corpus, les appositions sont généralement suivies d'un PRO3 ou d'un SN défini court ou présentant une reprise pouvant fortement prétendre à être co-référentielle. Il s'agit là des types de sujets les plus fréquents de notre corpus. Lorsqu'on compare les écarts réduits des phrases avec et sans apposition, on constate que les dernières ont significativement plus de sujets de type pronom ou redénomination de nom propre. La présence d'une apposition est donc associée à des sujets marquant une continuité topicale. En examinant ce qui se passe pour chaque sous-corpus, on observe certaines spécificités.

L'association avec les noms propres est particulière à PEOPL. PEOPL rassemble 60% des noms propres de notre corpus (cf. tableau 1), ceux-ci se situant essentiellement en initiale de section ou en initial de paragraphe pour le cas précis des noms propres répétés (cf. exemple 3).

Il s'impose donc d'explorer d'abord cette conscience esthétique pour comprendre la genèse, la raison d'être et la signification de l'œuvre poétique. C'est peut-être pourquoi l'influence de **Baudelaire** a été finalement plus esthétique que poétique : [...].
Critique d'art et critique littéraire, Baudelaire pratique, selon sa propre formule, une critique « partiale, passionnée, politique », dont il pense qu'elle est la plus « juste » : [...]

Exemple 3. Apposition et redénomination en initiale de paragraphe dans PEOPL

Si l'association entre une apposition et un sujet pronominal se retrouve dans tous les sous-corpus, la position textuelle privilégiée par cette association mérite d'être examinée. Dans ATLAS, les appositions suivies d'un pronom sont significativement plus présentes en S1 et en P2 ; alors que dans PEOPL, les écarts significatifs s'observent en P1 et P2. Ces préférences correspondent à notre sens à deux rôles différents des appositions. En P2, l'apposition joue son rôle phrastique de prédication d'arrière-plan sans réel pouvoir discursif structurant. En initiale de section ou de paragraphe en revanche, l'apposition marque une continuité topicale globale, comme en 3 ou encore comme dans l'exemple 4 ci-dessous, issu d'ATLAS, où l'apposition rend possible une reprise pronominale du titre de section.

SeaFrance-Sealink [titre de niveau 2]
Pavillon récent, apparu le 1er janvier 1996, il est l'héritier de diverses alliances entre la S.N.C.F. et des compagnies étrangères, depuis la privatisation de British Rail en 1984 par le gouvernement britannique. [...]

Hoverspeed [titre de niveau 2]
Appartenant au groupe britannique Sea Containers, la compagnie Hoverspeed est présente sur les lignes Calais-Douvres et Boulogne-Folkstone. Elle a concentré sa cible [...]

Exemple 4. Apposition et pronominalisation en initiale de section dans ATLAS

Les différences mesurées entre ATLAS et PEOPL sont relatives à la structuration générale des textes de ces deux sous-corpus et notamment au rôle du découpage en section. Dans ATLAS, le référent topique est introduit par un titre de section, alors que dans PEOPL le référent topique est donné par le titre du texte lui-même. En d'autres termes, dans ATLAS, seules certaines sections sont susceptibles d'être gouvernées par une continuité topicale globale. Dans PEOPL, la forte fréquence de PRO3 et de noms propres indique

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Discours, pragmatique et interaction
DOI 10.1051/cmlf08155

le caractère mono-référentiel des textes. Le positionnement des appositions en initiale de paragraphe permet alors de tisser plus fortement cette continuité topicale en dépit des déplacements qui se produisent au niveau du fond (les appositions montrent une forte propension à être précédées d'un circonstant) ou des articulations rhétoriques.

5 Conclusion

L'analyse et l'annotation systématiques de la zone préverbale des phrases dans un corpus partitionné constituent la base de notre approche exploratoire. Le choix de cette zone préverbale tient au fait que des travaux antérieurs tendent à lui attribuer un rôle clé dans l'organisation du discours, tout en laissant de nombreuses questions ouvertes quant au fonctionnement des éléments la composant. C'est là en effet que sont fournies des instructions essentielles concernant la relation de continuité ou au contraire de discontinuité qu'entretient la phrase en cours avec le discours alentour. En fait, comme la zone préverbale est constituée essentiellement d'éléments circonstants et d'éléments sujets, on constate qu'elle permet fréquemment de combiner continuité (au niveau figure) et discontinuité (au niveau fond). Les données présentées ici visent d'abord à mettre en évidence l'importance de la position textuelle dans laquelle apparaît un « marqueur » potentiel. Nous proposons que la signalisation de l'organisation discursive n'est pas le fait de « marqueurs » discrets mais plutôt de configurations de traits incluant la position dans la structure du document et le type de texte. À ce premier parcours des résultats nous ajoutons un regard plus ciblé sur la comparaison des sous-corpus, qui nous permet d'appuyer par des données quantitatives la caractérisation initiale que nous en avions faite.

Enfin, nous proposons un zoom rapide sur l'apposition, dont le rôle dans l'organisation discursive, peu étudié jusque là, est apparu lors de cette étude. Là où les appositions en position intraparagraphique se bornent au rôle de prédication d'arrière-plan, le fait de placer une apposition en début de section ou de paragraphe lui confère un rôle structurant au niveau discursif. Cette même approche nous a conduit dans une autre étude à moduler le rôle des circonstants temporels, souvent considérés comme d'évidents marqueurs de segmentation, selon leur position textuelle (Ho-Dac & Péry-Woodley, 2008).

Deux aspects de notre approche en corpus ont permis ces observations : d'une part la prise en compte d'aspects positionnels dans une vision paramétrique ou configurationnelle de la signalisation de l'organisation discursive ; d'autre part la mise en œuvre de procédures de découverte à partir d'un étiquetage et d'une annotation exhaustives de la zone préverbale.

Références

Adam, J.-M. & Revaz, F. (1989). Aspects de la structuration du texte descriptif: Les marqueurs d'énumération et de reformulation. *Langue Française, 81*, 59-98.

Ariel, M. (2004). Accessibility Marking : Discourse functions, discourse profiles and processing cues. *Discourse processes, 37(2)*, 91-116.

Bestgen, Y. & Vonk, W. (2000). Temporal adverbials as segmentation markers in discourse comprehension. *Journal of Memory and Language, 42*, 74-87.

Bestgen, Y., Degand, L. & Spooren, W. (2006). Toward automatic determination of the semantics of connectives in large newspaper corpora. *Discourse Processes, 41(2)*, 175-193.

Biber, D. (1998). *Variation across speech and writing*. Cambridge : Cambridge University Press.

Bourigault, D. & Fabre, C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaire, 25,* 131-151.

Bourigault, D. (2007). *Un analyseur syntaxique opérationnel : SYNTEX*, mémoire d'HDR en sciences du langage, CLLE-ERSS, Toulouse, France.

Charolles, M. & Péry-Woodley, M.-P. (2005). Introduction. *Langue Française, 148*, 3-8.

Charolles, M. (1994). Cohésion, cohérence et pertinence du discours. *Travaux de Linguistique, 29*, 125-151.

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Discours, pragmatique et interaction
DOI 10.1051/cmlf08155

Charolles, M. (1997). L'encadrement du discours : univers champs domaines et espaces. *Cahier de Recherche Linguistique, 6*, 1-73. LanDisCo, Université Nancy2.

Enkvist, N.E. (1976). Notes on valency semantic scope and thematic perspective as parameters of adverbial placement in English. In N.E. Enkvist & V. Kohonen (eds) *Reports on Text Linguistics: Approaches to Word Order*, 51-74. Abo : Publications of the research institute of the Abo Akademi Foundation.

Enkvist, N.E. (1985). A parametric view of word order. In E.Sözer (ed) *Text Connexity Text Coherence : Aspects Methods Results*, 320-336. Hamburg : Helmut Buske.

Goutsos, D. (1996). A model of sequential relations in expository text. *Text, 16(4),* 501-533.

Gundel, J.K., Hedberg, N. & Zacharski, R. (2000). Statut cognitif et forme des anaphoriques indirects. *Verbum, 22*(1), 79-102.

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

Ho-Dac, L-M. & Péry-Woodley, M-P. (2008). Temporal adverbials and discourse segmentation revisited. In W. Ramm & C. Fabricius-Hansen (eds.) *Linearisation and Segmentation in Discourse (Multidisciplinary Approaches to Discourse 2008)*, 65-77. Oslo: University of Oslo.

Jaubert, A. (2006). Les ordres du discours en perspective : cohérence et pertinence. In F. Calas (ed) *Cohérence et* discours, 15-21. Paris : Presses universitaires de Paris-Sorbonne.

Lambrecht, K. (1994). *Information structure and sentence form. Topic focus and the mental representation of discourse referents*. Cambridge, Massachusset : Cambridge University Press.

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

Péry-Woodley, M.-P. (2005). Discours, corpus, traitements automatiques. In A. Condamines (ed) *Sémantique et corpus*, 177-210. Paris : Hermès.

Piérard, S. & Bestgen, Y. (2006). Validation d'une méthodologie pour l'étude des marqueurs de la segmentation dans un grand corpus de textes. *TAL, 47(2)*, 89-110.

Rayson, P.E. (2002). *Matrix : a statistical method and software tool for linguistic analysis through corpus comparison*. Ph.D. in Computer Science, Lancaster University, UK.

Schnedecker, C. (2003). La question du nom propre répété dans la théorie dite du centrage et ses problèmes. *French Language Studies, 13*, 105-134.

Schneuwly, B., Rosat, M-C. & Dolz, J. (1989). Les organisateurs textuels dans quatre types de textes écrits : étude chez des élèves de dix, douze et quatorze ans ». *Langue française, 81,* 40-58.

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

Virtanen, T. (1992). *Discourse Functions of Adverbial Placement in English: Clause-Initial Adverbials of Time and Place in Narratives and Procedural Place Descriptions*. Abo : Abo Akademi University Press.

Virtanen, T. (2004). Point of departure : Cognitive aspects of sentence-initiale adverbials. In T. Virtanen (ed) *Approaches to cognition through text and discourse*, 79-97. Berlin : Mouton de Gruyter.

---

1 Nous rendons ainsi en français la notion de « *setting* », distinguée des « *circumstances* » par plusieurs auteurs dont nous nous inspirons (Enkvist, 1976; Lambrecht, 1994, *inter alia*).

2 Cette notion d'*indexation* est développée dans (Charolles, 1997) et (Charolles & Péry-Woodley, 2005).

3 Cf. « data-driven corpus linguistics » par opposition à « hypothesis-driven corpus linguistics » (Tognini-Bonelli, 2001; Rayson, 2002).

4 Ce type d'approche n'est possible que grâce à l'existence d'analyseurs robustes et fiables. Nous remercions Didier Bourigault, de CLLE-ERSS, Toulouse, pour son aide dans l'utilisation de Syntex.