La position initiale dans l'organisation du discours : une exploration en corpus

Lydia-Mai Ho-Dac

Thèse de doctorat nouveau régime
Université Toulouse le Mirail - département de Sciences du Langage
Laboratoire Cognition – Langues – Langage – Ergonomie
Équipe de Recherche en Syntaxe et Sémantique
Lydia-Mai Hồ-Đắc
automne 2007

LA POSITION INITIALE

DANS L'ORGANISATION DU DISCOURS :

UNE EXPLORATION EN CORPUS

Composition du jury :

Liesbeth Degand	Professeur, Université Catholique de Louvain la neuve (UCL), Belgique (rapporteur)
Benoît Habert	Professeur, école normale supérieure de Lyon (ENS LSH) (rapporteur)
Michel Charolles	Professeur, Université Paris III (examinateur)
Andrée Borillo	Professeur émérite, Université Toulouse II-le Mirail (examinatrice)
Didier Bourigault	Chercheur HDR, CNRS, Université Toulouse II-le Mirail, CLLE-ERSS (examinateur)
Marie-Paule Péry-Woodley	Professeur, Université Toulouse II-le Mirail (directrice)

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

CLLE ERSS UMR 5263
COGNITION, LANGUES, LANGAGES, ERGONOMIE

UNIVERSITE DE TOULOUSE LE MIRAIL

« Quand on montre la lune,
L'imbécile regarde le doigt. »
(Kong Fu Zi, Chine)

« Quand on représente le réel,
Le linguiste regarde ce qui sert à le représenter. »
(Lefebvre, Université Nancy 2, France)

Remerciements

Afin de ne pas orienter l'interprétation du lecteur en faveur de tel ou tel remercié (la nature de mes relations avec toutes ces personnes étant, comme la vie, impermanente), mes remerciements se feront sous forme d'abécédaire, type de texte dans lequel l'ordre des paragraphes importe peu, et où la stratégie textuelle consiste principalement à exprimer en position initiale la personne ou le groupe de personne que l'on souhaite remercier.

B **Andrée Borillo** j'ai souvent senti une présence rassurante derrière mon épaule

Didier Bourigault footballeur qui a « légué à la science » un sacré outil qui ne va pas cesser de se développer et encore plus si, un jour enfin, on arrive à échanger nos programmes

C **Michel Charolles** mes premiers pas en pragmatique du discours puis, plus tard, sa façon d'expliquer l'hypothèse de l'encadrement du discours et toutes les conséquences que cela a eu sur mes travaux

D **Liesbeth Degand** beaucoup de coïncidences ont fait notre rencontre. Je reste encore admirative devant cette mère marathonienne de linguistique (ça ne veut rien dire mais je trouve ça joli)

E **(mes) Enfants** (Edgar-Thành et Léo-Sanh) du travail, du courage, de la ténacité... il paraît que que je suis trop négative lorsque je parle de mes excessivement adorables petits bouts de bonheur!

ERSS un chouette labo. où l'on rencontre des gens 'qui n'en veulent' et qui le font, avec une petite pincée de rêve. Une troupe de doctorant soudée à laquelle on facilite énormément leur intégration et leur collaboration avec les « grands ». Une petite pensée personnelle pour (par ordre d'apparition) Anne Le Draoulec, Ludovic Tanguy, Josette Rebeyrolle, Marie-Paule Jacques, Christine Pernet, Pascale vergely, Cécile Frérot, Sylwia Ozdowska, Christophe Pimm, Marion Laignelet, Myriam Bras, Marianne Vergez-Couret, Aurélie Picton, Fred Saez & Christel LeBellec.

G Le **GREYC** le projet GeoSem et tous ces caennais ont été pour moi une grande source d'inspiration et de motivation, en particulier nos discussions avec Frédéric Bilhaut, Patrice Enjalbert et Nadine Lucas.

H **Benoît Habert** jeux de mots exigeants, quelqu'un qui donne beaucoup de son temps et de son intelligence aux autres.

I **Isa** gratte

L **Philippe Lefebvre** qui m'a montré l'intérêt des langues et de leur observation... entre Saussure et Antonin Arthaud, avec la 'gueule poétique' de Léo Ferré.

O **Orthographe du français** je ne la remercie pas.

P **(mes) parents** que c'est rassurant et motivant de sentir la fierté de ses parents! C'est à eux que je dédie principalement ce travail.

Marie-Paule Péry-Woodley notre rencontre, nos discussions, sa simplicité et sa volonté. Je ne saurai pas encore bien décrire notre relation, mais c'est quelqu'un avec qui je partage beaucoup de choses.

Les Poupées Barbares : une belle partie de ma vie, des amis, des tournées (internationales) passées à méditer et cogiter sur la vie et... la position initiale dans l'organisation du discours.

S **Franck Sajous** : une voie de niveau 6A (c'est le maximum qu'on ait réussi à faire jusque ici), une petite mousse sur la dalle, des transformations XSLT pour mettre en ligne le corpus GEOPO, des déménagements... quelle variété d'événements nous avons vécus ensemble!

T **Tatou** : arrivée à Toulouse le 6 novembre 1998, sur neuf années de vie commune il y a eu sept années de thèse... ça y est Tatou, je vais être docteur!

Table des matières

Table des figures

Table des tableaux

Table des graphiques

RÈGLES TYPOGRAPHIQUES

Dans le corps de texte -

Caractères italique_1 : en mention aux mots présents dans les exemples discutés.

Ex : Dans l'exemple x, le connecteur *mais* relie deux propositions entre elles.

Caractères italique_2 : entourent les emprunts à une autre langue ou les expressions issues de traductions.
Ex : en anglais, une distinction est faite entre : d'une part, les *circumstances* c'est-à-dire l'expression des circonstances au niveau sémantique (circonstants de temps, de lieu, de manière, de but, de condition, etc.) ; et d'autre part, les *settings* qui permettent la localisation du procès et/ou des entités par rapport à un repère spatial/temporel, un domaine de connaissance (une notion), un élément du « *text world* ».

Caractères gras : pour mettre en valeur une expression qui devient une sorte de mot-clef du passage dans lequel se situe cette expression.

« Guillemets doubles » : entourent les emprunts à une terminologie spécifique.
Ex : Pour Charolles (1997), un « cadre de discours » correspond à un segment dont les propositions successives sont toutes à interpréter selon un même critère sémantique exprimé explicitement en début de cadre par un « introducteur de cadre ».

'Guillemets simples' : entourent les expressions utilisées dans un usage déviant par rapport à la 'normale'.

<> : entourent les étiquettes données aux différents éléments repérés dans l'analyse en corpus. Ainsi, <*étiquette*> constitue une balise. Comme tout balisage, une étiquette délimitant la borne finale de l'élément est nécessaire et sera indiquée par une balise fermante : </*étiquette*>.

MAJUSCULES : les majuscules sont réservées aux sigles et aux noms de corpus.

Pour les exemples -

Les expressions focalisées sont en **gras** et en surligné gris s'il y a besoin de distinguer deux types d'expression.

Les nom du corpus ou sous-corpus et le numéro du texte dont est extrait l'exemple sont mis à la fin en gris et entre crochets : *[(SOUS-)CORPUS_NumTxt]*. Les titres des différents textes associés aux numéros sont donnés en annexe A. Si l'exemple provient d'une autre source que notre corpus d'étude, elle est mentionnée de la même manière : caractères gris entre crochets.

S'il y a un titre de section dans l'extrait de texte, il est souligné et accompagné (le cas échéant) de sa numérotation.

pour les tableaux -

Les mesures en nombre sont en caractères normaux.

Les mesures en pourcentage sont en caractères italiques.

Les mesures en écart réduit sont toujours précédées du signe + ou –.

Les notes de bas de tableaux apparaissent dans la dernière ligne du tableau.

Glossaire des Abréviations

EOD	étude de l'organisation discursive (p.20)
Modèles théoriques	
SF	Systémique Fonctionnelle (p.32)
FSP	*Functional Sentence Perspective* (p.88)
SDRT	*Segmented Discourse Representation Theory* (p.21)
RST	*Rhetorical Structure Theory* (p.21,58)
MAT	Modèle d'Architecture Textuelle (p.42)
MFM	Mise en forme matérielle du document (p.42)
Principes organisationnels	
CIF	*crucial information first* : principe de l'information cruciale en premier (p.19)
OIF	*old information first* : principe de l'information donnée en premier (p.88)
TSC	continuité texto-stratégique (p.38,50)
TP	progression thématique (p.73)
IC	introducteur de cadre (p.76)
caractéristiques de la position initiale	
PosTxt	position textuelle (S1, P1 ou P2) (p.100)
S1	Initiale de sections (p.100)
P1	initiale de paragraphes hors S1 (p.100)
P2	initiale de phrases hors P1 et S1 (p.100)
Connect	connecteur 'pur' en position initiale (p.174)
INIT	élément détaché en initiale de phrase (p.96)
INIT1	premier élément détaché en initiale (p.174)
INIT2	deuxième élément détaché en initiale (p.174)
APPO	apposition (p.111)
ARGU	argument inversé (p.111)
CIRC	adverbial circonstanciel (p.111)
CIRCtps	adverbial circonstanciel temporel (p.177,195)
CIRCspa	adverbial circonstanciel spatial (p.177,195)
CIRCnot	adverbial circonstanciel notionnel (p.177,195)
CIRCautre	autre type de CIRC (p.195)
MIL	Marqueur d'intégration linéaire (p.177)
MODA	adverbial modalisateur (p.111)
ST	Séquence thématique en initiale d'une construction pseudo-clivée (p.177)

TEXT	adverbial textuel (p.111)
TOP	élément disloqué en initiale (p.177)
ssINIT	absence d'élément détaché en initiale (p.227)
Iautre	INIT de forme ou de fonction indéfinie (p.211)
ThTop	Thème topical = sujet grammatical dans les constructions non spéciales (ThSpe) (p.96)
ProPoss	pronoms et SN possessifs (p.206)
PRO3	pronom personnel de 3e personne (p.100)
PROdemo	pronom démonstratif (p.179)
SN	syntagme nominal (p.37)
SNdef	SN définis (p.179)
SNdem	SN démonstratifs (p.179)
SNindef	SN indéfinis (p.179)
SNposs	SN possessifs (p.179)
NP	nom propre (p.179)
SP	syntagme prépositionnel (p.170)
PPA	proposition avec participe passé (p.177)
PPR	proposition avec participe présent (p.177)
REL	proposition relative (p.177)
ADJ	syntagme adjectival (p.177)
ADV	adverbe ou locution adverbiale (p.177)
INF	proposition infinitive (p.177)
FIN	propositions subordonnées finies (p.177)
_R	reprise lexicale (p.179)
ThSpe	Construction à Thème spécifique = construction spéciale (p.96)
Cliv	construction clivée (p.179)
Present	construction présentationnelle (p.179)
ILimp	construction impersonnelle (p.179)
On...	construction en « *On...* » (p.179)
SujInv	sujet inversé (p.179)
Disloc	dislocation à gauche (p.179)
z*(x)=+/-*n	l'écart réduit pour la variable *x* est égal à +/-*n* (p.159)
DegAccess_*n*	degré d'accessibilité *n* (p.180)

Chapitre I
Introduction

Sommaire

Cette thèse offre une approche et une méthodologie nouvelles pour l'**É**tude de l'**O**rganisation **D**iscursive – EOD, partant de l'hypothèse que la « position initiale » (de phrase, de paragraphe, de section, de texte) joue un rôle crucial dans le marquage de l'organisation du discours.

Notre méthodologie expérimentale propose des analyses quantitatives effectuées de façon automatique sur un corpus relativement important pour le domaine d'étude[1]. L'originalité de cette méthodologie est de permettre une analyse *data-driven* de l'organisation du discours (cf. VI.1.1). Elle se base sur une annotation automatique de ce qui compose la position initiale. Le corpus annoté constitué peut être réutilisé pour d'autres travaux dans le même domaine[2]. Le programme d'annotation peut également être intégré à un outil TAL plus complexe, ayant trait à l'accès à l'information dans des documents longs.

1 En effet, comme nous l'expliquons dans le Chapitre VI , les études en corpus sur la segmentation du discours consistent généralement en des analyses manuelles qui impliquent nécessairement une faible quantité de données et une interprétation, voire une explication des textes.

2 Le travail présenté dans Hồ-Đắc & Laignelet (2005) se base justement sur ce corpus annoté.

I.1. Problématique : rôle de la position initiale dans le discours

I.1.1. Un point de départ

La « position initiale » (Chapitre IV) correspond à la notion de « point de départ » d'un segment. Elle englobe tous les éléments par lesquels commence un segment textuel : les premiers éléments d'une phrase, d'un paragraphe, d'une section. L'étude de la position initiale relève du « problème de la linéarisation » (Levelt 1981). L'écriture et la lecture (ou leurs correspondants en communication orale) étant tributaires du temps, il y a inévitablement un élément qui commence le texte, la nouvelle section, le nouveau paragraphe, la nouvelle phrase. L'étude du phénomène de linéarisation et de la position initiale constitue un terrain d'investigation commun aux linguistiques cognitives et linguistiques du discours : par la notion de point de départ et de saillance pour les premières ; et par la notion de thème, de topique, d'information donnée ou de degré de dynamisme communicatif pour les secondes.

> "Linearization is an area of study where the interests of cognitive linguists and discourse linguists meet."
> (Virtanen 2004:95)

Dans cette étude, l'objet position initiale correspond à la zone préverbale des phrases. Cependant, nous ne limitons pas la position initiale au niveau phrastique et envisageons une extrapolation du concept de point de départ. Ainsi, la zone préverbale de la première phrase de paragraphe est considérée comme la position initiale du paragraphe. Au niveau des sections, la position initiale est plus complexe, étant constituée du titre de section et de la zone préverbale de la première phrase de la section. On peut cependant supposer que le titre de section joue à un niveau légèrement différent (voir V.2).

Dans cette définition, nous gardons une délimitation relativement serrée de la position initiale. Nous aurions pu adapter la taille de cette délimitation à celle du segment textuel initié et alors considérer comme position initiale toute la première phrase pour le paragraphe, tout le premier paragraphe pour la section. Deux raisons ont justifié notre choix :

- D'un point de vue cognitif, le point de départ ne peut dépasser une certaine taille. Quand nous survolons un article écrit, nous regardons essentiellement le début des segments[3], notre capacité de lecture limitant cette zone à quelques mots[4].
- D'un point de vue méthodologique, travailler sur un objet de même taille permet une analyse comparée des différents niveaux de segmentation. Nous comparons ainsi la composition de la position initiale selon que l'on se place au niveau de la phrase, du paragraphe ou de la section.

Les théories sur le statut spécial de la position initiale sont nombreuses et variées, que ce soit dans une approche cognitiviste, discursive, informationnelle ou syntaxique (*cf.* le tableau proposé par Gómez-González 2001:50-51). Cette position est importante au niveau organisationnel puisqu'elle se situe à l'articulation de deux segments. Ainsi, elle fonctionne comme un pivot servant à la fois d'ancrage et de point de départ pour le nouveau segment. Notre conception s'accorde complètement avec celle exposée par Virtanen (1992, 2004) qui travaille depuis longtemps sur le rôle de la position initiale dans l'organisation des textes. Dans un article intitulé '*Point of departure : Cognitive aspects of sentence-initial adverbials*', Virtanen définit la position initiale par les trois fonctionnements suivants : lier le discours antérieur au discours à venir, orienter l'interprétation des segments à venir et donner une certaine importance dans la construction de la représentation mentale aux éléments mis en position initiale.

3 Si une mise en forme matérielle attire notre attention ailleurs, cette affirmation est à relativiser.

4 Les travaux en oculométrie montrent que notre regard peut englober entre 5 et 15 caractères typographiques en même temps : 5 en vision fovéale (centre de l'oeil – rétine : vision précise) et 15 en vision parafovéale (zone périphérique de la rétine : vision moins précise qui permet tout au plus de distinguer les caractères alphabétiques des autres caractères), *cf.* Foucambert (2003).

> "To start with, the element placed [in initial position] can be given the job of tying what is to come to what can be assumed to be present in the text world that readers are constructing on the basis of the text, its context, and their knowledge of the world. [...]
> Secondly, elements placed at the outset of a sentence also help readers anticipate what is to come as they pinpoint what the sentence is about and how it relates to the discourse topic. [...]
> Furthermore, it is occasionally profitable to start with what is regarded as 'crucial information' (Enkvist 1989) in the context, irrespective of whether this information is treated as given or not." (Virtanen 2004:80-81)

Cette dernière idée d'« information cruciale en premier » – CIF[5], nous permet de généraliser le fonctionnement de la position initiale en discours. Car ce qui est 'le plus important' varie selon la situation. Il est parfois plus important d'insister sur le lien à établir entre deux segments et parfois plus important de préciser le cadre dans lequel les informations à venir sont à interpréter. La partie V.1.1 concerne précisément cette idée de correspondance entre position initiale et information cruciale.

I.1.2. La position initiale : un indice de l'organisation discursive

Nous envisageons la position initiale comme un indice dans le marquage de l'organisation discursive. Dans le chapitre V, nous posons que le marquage de l'organisation discursive ne se fait pas par l'emploi d'expressions bien définies, mais par des ensembles d'indices – des configurations d'indices – qui, par combinaison, marquent telle ou telle relation de discours. Jacques & Rebeyrolle (2006:7) parlent d' « influence conjointe de divers facteurs (linguistiques) ».

Le fait de placer tel ou tel élément en position initiale confère à l'élément un rôle dans l'organisation du discours. Par exemple, un connecteur va jouer un rôle différent dans le discours selon qu'il est placé au milieu, en début ou en fin de phrase, ou encore en début de paragraphe. Cette différence a trait à l'organisation discursive. Dans le cas d'un milieu de phrase, le connecteur a pour unique rôle de connecter deux propositions, *i.e.* d'établir entre elles une relation de cause, de conséquence, de succession, etc. Dans les cas d'initiale de phrase ou de paragraphe, le connecteur peut être perçu comme un marqueur discursif : délimitant et articulant deux blocs informationnels. L'exemple I.1 nous montre un exemple typique de « connecteur propositionnel » *vs.* « connecteur discursif ».

(I.1) Les images vendues par les entreprises non-américaines restent de résolution plus grossière et présentent pour l'instant moins de danger en termes de sécurité. Lors de la campagne d'Afghanistan, par exemple, Spot Image a reçu peu de commandes de la presse, car les images disponibles ([...]) n'étaient finalement pas assez parlantes pour le public..... C'est pour cette raison que l'entreprise française se contente actuellement de respecter les décisions de l'ONU. Il peut s'agir d'embargos, comme celui qui s'applique à l'Irak, ***mais*** *surtout des résolutions plus spécifiques de 1986 sur l'imagerie. En accord avec ces résolutions, Spot Image garantit l'accès des pays observés aux photographies prises, dans des délais et pour un prix raisonnable. Le gouvernement français n'a pas adopté pour l'instant de mécanisme de contrôle plus strict.*
Mais *les satellites commerciaux non-américains progressent eux aussi vers des résolutions plus fines, aux alentours de deux mètres. Le satellite israélien Eros s'approche de la résolution métrique avec une résolution de 1,8 mètres. Spot 5 [...]. Rocsat pour la Corée et Alos pour le Japon[...]. La compagnie russe Sovinform Spoutnik [...]. L'Inde [...]. L'intérêt de l'imagerie à deux mètres reste plus limitée pour des utilisateurs hostiles ou pour les médias. [GEOPO_2]*

Dans cet extrait, nous avons un premier *mais* qui relie deux propositions par une relation rhétorique que l'on peut appeler « relation de préparation »[6]. Cette conjonction de coordination ne remplit pas d'autre fonction que celle-ci. Par contre, on ne peut pas dire que le deuxième *mais* sert également à mettre en relation les deux propositions alentours. Ce deuxième *mais* relie le dernier paragraphe aux paragraphes précédents. Plus encore, il organise l'extrait en indiquant que les propos qui le suivent vont dans une autre direction que celle suivie dans les paragraphes précédents.

5 Nous gardons les abréviations de la formulation anglaise : *Crucial Information first*.

La position initiale confère à ce dernier *mais* une **portée**, c'est-à-dire une capacité à étendre son influence au delà de sa phrase d'accueil. De ce fait, il acquiert la capacité à **indexer** toute une portion de texte (les notions de portée et d'indexation sont expliquées en II.3.2.). En ce sens, la position initiale est un marqueur discursif puisqu'elle consiste en un trait linguistique pouvant donner aux éléments qui s'y situent une portée.

I.2. L'Étude de l'Organisation Discursive - EOD

L'objet de cette étude concerne ce que l'on appelle « l'organisation discursive ». Ce terme répond à l'idée que les textes ne sont pas des sacs de mots, de phrases, de paragraphes, etc. Le texte constitue un objet structuré : il y a un ordre des mots, des phrases, des paragraphes, etc. A un niveau très général, une façon d'organiser son texte peut suivre le plan : introduction, développement, conclusion. À une niveau beaucoup plus local, une façon d'organiser son texte peut consister à subdiviser ses phrases en une partie thème et une partie rhème. La partie thème contient l'information donnée qui ancre la phrase dans le discours précédent, et la partie rhème contient l'information nouvelle qui fait avancer le discours, lui confère un certain « dynamisme[7] ».

Deux approches majeures sont généralement adoptées en EOD. D'un côté, nous avons les modèles logiques qui visent à modéliser le langage et à travers lui, la pensée (domaine de l'intelligence artificielle). De l'autre côté, nous avons les modèles cognitivistes qui visent à comprendre les processus mentaux à l'oeuvre dans la production et la compréhension des textes (domaine de la psychologie). Notre étude s'inscrit dans une troisième approche que nous qualifions d'approche fonctionnelle et qui vise principalement une description linguistique de l'organisation du discours. Avant de décrire plus précisément notre démarche, il nous a semblé nécessaire de nous situer par rapport aux deux grandes approches : l'approche logique et l'approche cognitive.

I.2.1. Une approche logique

L'approche logique est née avec l'intelligence artificielle et le langage informatique. Il s'agit de mettre sous une forme logique et donc compréhensible par une machine, les règles sous-jacentes au langage naturel. L'approche logique du discours cherche à formaliser les représentations sémantiques établies en discours afin d'élaborer une sémantique computationnelle du discours. La première étape de cette démarche consiste à représenter le langage naturel sous forme de règles de production. La visée applicative est de générer automatiquement un langage naturel, c'est-à-dire à exprimer sous la forme d'un langage informatique les processus propres au langage naturel.

Les modèles adoptant cette approche logique, dits « modèles logiques » (*model-theoric semantics*) cherchent à construire une représentation de l'information véhiculée par des expressions en langage naturel au niveau de la phrase ou au niveau d'une suite de phrases[8]. Cette représentation consiste à attribuer une forme logique à un sens. Par exemple la phrase suivante :

Un homme dort

---

6 Cette dénomination est emprunté à la Rhetorical Structure Theory qui établit la liste des contraintes pragmatiques et sémantiques impliquées dans les différentes relations de discours. Selon la RST, la relation de « préparation » met en relation deux propositions, l'une permettant de placer le lecteur dans de meilleurs conditions pour la lecture de l'autre. Un exemple de représentation RST est donnée en III.1.2.a et plusieurs relations rhétoriques sont définies à l'adresse : http://www.sfu.ca/rst/07french/definitions.html

7 Nous empruntons ce terme à la théorie du dynamisme communicatif de Firbas (1992).

8 Il s'agit généralement d'analyses *bottom-up* : on part d'une analyse des phrases pour 'monter' à l'analyse des suites de phrases... jusqu'à arriver à la représentation logique du texte entier.

peut être représentée par la forme logique suivante :

$\exists x$ (homme(x) $\wedge$ dort(x)) qui se lit : « Il existe un élément x tel que x est un homme et x dort. »

À un niveau plus discursif, au delà de la phrase, l'approche logique va particulièrement s'attacher à représenter les relations qui lient les propositions (relations rhétoriques, relations d'anaphore, etc.) La **R**hetorical **S**tructure **T**heory – RST (Mann & Thompson 1986) et la **S**egmented **D**iscourse **R**epresentation **T**heory – SDRT (Asher 1993) proposent une modélisation de ces relations. Par exemple, les deux propositions en (I.1) sont articulées par une relation d'élaboration.

(I.1) En juin, Julie est partie en Iran. Elle y visita de nombreux musées.

Les représentations de suite de phrases sont plus complexes qu'une simple addition des formules logiques de chaque phrase prise isolément. La figure I.1 montre une représentation de (I.1) selon la SDRT (Asher 1993). Cette représentation schématise trois propositions reliées les unes aux autres. K_1 et K_2 modélisent respectivement le sens véhiculé par la première et la deuxième proposition, qui constituent des propositions explicites, *i.e.* ayant une réalité textuelle. La 'boîte' K'_0 modélise une proposition d'un autre type : une proposition relationnelle, *i.e.* ayant une réalité non pas textuelle mais interprétative. La proposition relationnelle reliant les deux phrases de (I.1) a pour sens : K_1 est lié à K_2 par une relation d'élaboration. Le sens du discours (I.1) équivaut donc à la mise en relation de K_1 et de K_2, K_2 servant à élaborer les propos véhiculés par K_1.

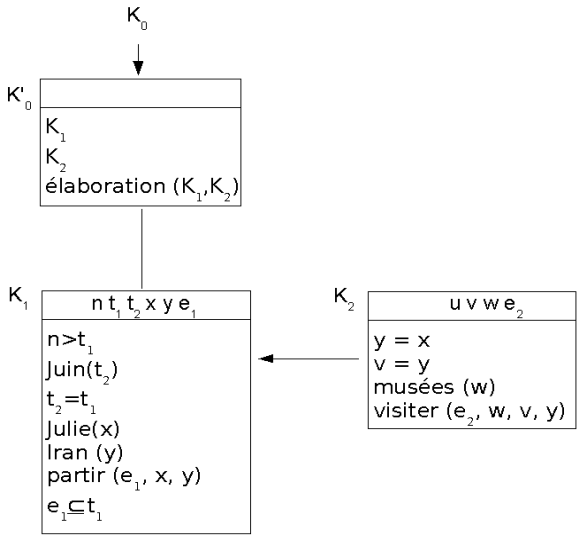


Figure I.1 : Représentation logique selon la SDRT (Asher 1993)

I.2.2. Une approche cognitiviste

Parallèlement à l'approche logique du discours se trouvent des modèles issus du domaine de la psychologie et plus particulièrement de la psychologie cognitive. Ces modèles de linguistique cognitive cherchent à comprendre et à décrire les processus mentaux à l'oeuvre dans un acte de communication verbale. Ici, l'objet d'étude a trait à la relation entre l'organisation discursive et l'organisation de notre pensée, ce qui suppose une relation entre notre façon d'organiser nos discours et notre façon de comprendre et d'appréhender le monde.

> "Human beings understand the world by constructing working models of it in their minds" (Johnson-Laird 1983:10)

La plupart de ces modèles étudient l'organisation discursive du point de vue de la compréhension plutôt que de la production. Ils s'intéressent davantage aux processus mis en oeuvre pour reconstruire une représentation mentale à partir d'un texte qu'à ceux mis en oeuvre pour construire un texte à partir d'une représentation mentale. Ces modèles s'appuient donc sur des expérimentations de lecture plus que d'écriture (en cas de discours écrit). Ainsi, nous avons le modèle de situation (Kintsch & Van Dijk 1983) et le modèle de construction-intégration (Kintsch 1988, 1995), les modèles mentaux (Johnson-Laird 1983), le 'Structure Building Framework' (Gernsbacher 1990), ou encore, pour le français, le modèle des espaces mentaux élaboré par Fauconnier.

L'idée générale et commune à ces modèles est que notre compréhension fonctionne par cycles. À chaque cycle correspond l'activation ou la construction d'une sous-représentation qui est une sorte de représentation temporaire. Au fil de la lecture, ces représentations 'tampon' (en référence à la mémoire tampon ou « *buffer memory* ») se connectent et se structurent pour former un réseau représentatif du texte entier. Sweetser & Fauconnier (1996) parlent d'un réseau constitué d'espaces mentaux, chaque espace mental correspondant à une représentation que l'on s'est construite au cours de notre expérience du monde.

> "As we think and talk, mental spaces are set up, structured, and linked under pressure from grammar, context, culture. The effect is to create a network of spaces through which we move as discourse unfolds. Because each space stems from another space (its 'parent'), and because a parent can have many offspring, the space network will be a two-dimensional lattice." (Sweetser & Fauconnier 1996:11)

Récemment, des liens entre l'EOD et les modèles cognitivistes se sont tissés autour de la problématique des marqueurs discursifs, *i.e.* autour de la recherche de corrélats linguistiques entre fonctions discursives et traits de surface. Il s'agit de déterminer la contribution de certains indices linguistiques à la construction du réseau des sous-représentations.

> "Text and discourse processing are dynamic processes during which the reader or listener constructs a cognitive representation of the information in the text or discourse. Even though readers' and listeners' representations are not identical to the information they read and hear, texts and discourses contain many linguistic signals that guide comprehension." (Sanders & Gernsbacher 2004:79)

I.2.3. Une approche fonctionnelle

Dans cette thèse, nous ne nous plaçons ni dans une approche logique de formalisation du sens, ni dans une approche réellement cognitive, même si nous nous inspirons de nombreux travaux appartenant au domaine de la psychologie cognitive. Notre point de départ n'est pas la sémantique du discours ou la pensée humaine, mais la structure même des textes : quel sens véhicule la structure des textes ou plus exactement quelle fonction remplissent les différents éléments constitutifs du texte dans la construction de sa représentation mentale ? Ainsi, nous cherchons, en corpus et de façon exploratoire, à mettre au jour des corrélats entre des traits linguistiques et des fonctions discursives.

L'organisation discursive est abordée par le biais de la segmentation textuelle et de la séquentialité (chapitre III). La **segmentation textuelle** est définie comme la subdivision du texte en portions ayant des propriétés à la fois fonctionnelles et structurelles. Ces propriétés se manifestent par deux phénomènes conjoints : un regroupement des éléments constitutifs de ce segment autour d'un critère spécifique et un découpage du texte en portions marquées. Les éléments regroupés sont alors dans une relation de continuité par rapport au critère de regroupement et les segments délimités dans une relation de discontinuité. La **séquentialité du discours** correspond à la gestion textuelle de ces zones de continuation entre éléments et de ces zones de transition entre segments.

Certains traits peuvent être considérés comme des **indices de séquentialité** car ils instruisent d'une discontinuité ou d'une continuité entre éléments. Nous supposons ainsi l'existence d'une sémantique de l'organisation du discours qui se réalise par des signaux dont une des fonctions est d'instruire comment ce qui est dit est à intégrer à ce qui a été dit et à ce qui va être dit. Ce sont ces indices de séquentialité qui nous intéressent ici, par une exploration en corpus de la position initiale, position clef dans l'enchaînement et la délimitation des différentes portions textuelles.

I.3. Méthodologie : exploration en corpus et analyses quantitatives

La structuration du discours est vraisemblablement instable et sensible au type de texte. Il semble évident que la façon d'organiser le contenu varie selon que l'on se situe dans des textes narratifs, procéduraux, expositifs, etc. Ce constat nous a amené à parier sur la variation langagière pour mettre au jour des confluences de traits, marqueurs de l'organisation discursive.

Nos analyses quantitatives reposent sur deux hypothèses : (1) la variation textuelle se trouve également au niveau de l'organisation discursive et (2) les stratégies textuelles varient selon le niveau d'organisation textuelle considéré (de phrase en phrase, de paragraphe en paragraphe et de section en section). Concernant la première hypothèse, nous réalisons une analyse comparée de trois sous-corpus sélectionnés *a priori* pour leur différence au niveau de l'organisation de leur contenu. Nous arrivons ainsi à distinguer des traits propres à un type d'organisation particulier de traits vraisemblablement communs au genre expositif (tous nos textes sont du genre expositif, voir I.4.2 infra). La deuxième hypothèse aboutit à l'analyse comparée des positions initiales de phrases, de paragraphes et de sections. Nous mettons ainsi en évidence des associations entre configurations de traits et position textuelle qui marquent des organisations discursives particulières ou communes aux différents sous-corpus constitués.

Encore peu de travaux en EOD se basent sur des analyses quantitatives. Il est effectivement difficile d'adapter une linguistique du discours aux méthodes des linguistiques de corpus. Le principal problème vient d'une caractéristique propre à l'EOD : l'absence d'indices précis et définis à partir desquels initier l'étude. Ce qui est recherché, ce sont justement les correspondances entre des fonctions discursives et des confluences de traits linguistiques associés à des traits extra-linguistiques. Les traits extra-linguistiques relèvent de l'acte de discours lié au texte (nous retrouvons ici les notions de genre discursif, voir VI.2.3) et du format physique du texte (sa longueur, sa mise en forme, voir I.4). Ainsi, selon par exemple le genre discursif et la longueur du texte, tel trait linguistique aura ou n'aura pas telle fonction discursive.

A cette sensibilité à l'extra-linguistique s'ajoute le caractère multi-échelle du texte. La construction du sens véhiculé par un texte va du niveau phrastique jusqu'au niveau du texte entier. De ce fait, les fonctions discursives sont difficilement limitées à un seul niveau. Ainsi, un connecteur peut faire l'objet d'une double annotation : au niveau phrastique entre deux propositions mais aussi au niveau discursif entre deux segments de taille plus importante (un ensemble de propositions, de phrases, de paragraphes, etc.) comme on l'a vu dans l'exemple I.1.

Troisième difficulté, le discours est un processus, un acte de communication, qui implique énormément de processus implicites et inférentiels. Les textes ne sont que la partie visible du discours et les phénomènes discursifs ne sont pas définissables de façon uniquement formelle. La part d'interprétation humaine dans la détection de segments textuels, de chaînes de référence, de cadres de discours est conséquente. Les analyses en corpus orientées discours sont de fait très souvent manuelles, ce qui restreint considérablement le nombre de données analysables et le nombre d'analyses envisageables. Face à ce constat, nous avons opté pour une démarche exploratoire basée sur des

analyses automatiques et quantitatives. Nous voulons alors tester ce que des méthodes automatiques peuvent apporter à l'EOD et proposons ainsi une méthodologie nouvelle pour le sujet d'étude.

Notre méthodologie ne remet pas en cause les méthodes traditionnelles qui s'appuient sur un repérage manuel des phénomènes à l'étude. Elle propose une alternative à ces travaux qui, en plus d'être fastidieux et de ne pouvoir porter que sur de petites quantité de données, présentent le risque de faire une analyse de texte et non une analyse du phénomène linguistique. En laissant un traitement automatique « lire » le texte, notre méthodologie offre la possibilité d'observer des régularités et des variations issues des textes eux-mêmes et non de notre interprétation des textes.

I.4. Des textes 'nécessitant' une organisation

I.4.1. Des textes longs

Il est frappant de constater que très peu de travaux en EOD s'appuient sur l'observation de textes longs. Il existe même des travaux portant sur des extraits de textes et non des textes entiers. Or, quelle organisation discursive peut être étudiée si l'on n'envisage pas le discours dans son entier ? À la base de nombreux travaux en EOD figure le précepte : un texte n'est pas un sac de mots. Mais n'est-ce pas le considérer un peu comme tel que de ne prendre que des bouts de textes ? C'est ne pas considérer qu'un texte est composé d'un début (une introduction), d'un milieu (un développement) et d'une fin (une conclusion), trois segmentations grossières et globales mais qui peuvent toutes trois prétendre à un mode d'organisation spécifique ou du moins, différent des autres.

Cette distinction entre différents modes organisationnels au sein d'un même texte n'est pas si évidente lorsque l'on travaille sur des textes courts. Notre hypothèse de portée associée aux éléments situés en position initiale n'est pas valable pour tous les types de texte. Nous sommes persuadée qu'une fonction discursive peut être 'inhibée' selon le type de texte étudié. Par exemple, la partie III.2.2 relate un travail qui montre que les textes narratifs présentent une organisation spécifique qui 'annule' la fonction 'cadrative' de certains éléments en position initiale. Mais en dehors de la notion de type de texte (précisée en VI.2.3) certaines caractéristiques physiques des textes peuvent atténuer ou supprimer la fonction discursive de certaines expressions.

La longueur et le découpage matériel des textes en chapitres, sections, paragraphes, etc. ont un impact évident sur l'organisation discursive. En effet, sans considérer les romans et autres textes narratifs, la longueur d'un texte implique une nécessité de l'organiser et de marquer cette organisation. Un texte d'une page (environ 500 mots), quelle que soit sa fonction sociale ne présente pas forcément un découpage en sections. Par contre, dès que l'on a un texte (non narratif, voir section suivante) de plus de quatre pages, ce qui correspond à un minimum de 2000 mots, la titraison devient une technique de structuration fréquemment utilisée. La longueur des textes et leurs caractéristiques typologiques constituent des facteurs de variation textuelle à prendre nécessairement en compte dans l'EOD vue leur influence sur l'organisation même des textes.

I.4.2. Des textes expositifs

Ce qui distingue les textes expositifs des autres catégories de textes (textes narratifs, textes procéduraux, etc.), c'est la fonction qu'ils remplissent dans notre vie : les textes expositifs 'exposent' des faits, des idées, des informations, etc. Comme le souligne Halliday (1971), le langage permet la construction d'objets (les textes) ayant une fonction

sociologique relative à certains besoins universels, c'est là la base de l'approche fonctionnelle définitoire de la Systémique Fonctionnelle, modèle dans lequel nous nous inscrivons pleinement.

> "[The functional approach is an approach] which attempts to explain linguistic structure, and linguistic phenomenon, by reference to the notion that language plays a certain part in our lives ; that it is required to serve certain universal types of demand" (Halliday 1971:331, cité par Trosborg 1997:11)

La fonction première que remplissent les textes expositifs est le partage et la transmission d'informations. L'auteur va publier ses connaissances en destination d'un public qui va chercher dans ce type de texte de l'information. Le terme « expositif » se retrouve dans la typologie définie par Longacre (1976), où le type expositif est opposé aux types narratif, procédural et exhortatif. Cette typologie ne s'appuie pas sur la fonction sociale des textes, mais sur leur organisation discursive. D'après Longacre, le discours expositif se caractérise par une stratégie rhétorique non persuasive et une structure non chronologique qui aboutit à une relation atemporelle entre les informations exprimées. Nous verrons que ces distinctions ne sont pas vraiment applicables à notre classification. Les textes expositifs peuvent répondre de différents genres tels que les biographies, les articles scientifiques, les éditoriaux, les articles de presse, les thèses, etc. qui tous ont pour fonction la communication d'information. La partie VI.2.3. présente notre distinction entre genres textuels et types textuels. La partie VII.1. présente en détail les caractéristiques fonctionnelles et structurelles de notre corpus d'étude.

Quelques études ont comparé le fonctionnement discursif de textes expositifs et narratifs, et particulièrement le travail de Le Draoulec & Péry-Woodley (2001, 2003 et 2005) et Berman & Nir-Sagiv (2007). Berman & Nir-Sagiv (2007) proposent une étude psycholinguistique de l'acquisition des techniques d'organisation discursive. Cette étude offre une passionnante comparaison de textes narratifs *vs.* expositifs produits, pour l'étude, par un public d'enfants, d'adolescents et d'adultes (80 locuteurs natifs anglais au total). La consigne de l'expérience est de produire deux textes sur un même sujet (les conflits entre personnes[9]) : un texte visant à raconter une histoire personnelle d'expérience de conflit interpersonnel (textes narratifs) et un autre dont le but est d'exposer sa conception des relations conflictuelles et ses idées sur le sujet[10]. Nous retrouvons ici les différences majeures entre les textes narratifs et les textes expositifs : raconter une histoire d'un côté et exprimer des idées générales de l'autre.

Le travail de Berman & Nir-Sagiv met en évidence que ces deux types de texte diffèrent non seulement au niveau du lexique et des constructions syntaxiques utilisés (ce qui rentre généralement dans la définition du registre de langue, voir VI.2.3), mais également au niveau des principes d'organisation sous-jacents. Les textes expositifs sont basés sur une organisation **hiérarchisée** autour de la notion de topique alors que les textes narratifs sont basés sur une organisation **chronologique** autour de la notion d'action.

> "In narratives, events are the basic components of text construction, and they are embedded in a narrative schema or action structure. Expository text construction, in contrast, is topic centered, with the discourse topic functioning as a *superordinate category* under which all the information needs to be subsumed, because expository text construction proceeds by reference to categories and concepts that are 'established, instantiated ... and related one to the other to form a system' (Bruner 1986:12)." (Berman & Nir-Sagiv 2007 : 91)

Les textes expositifs sont ainsi définis structurellement comme des textes dont le mode organisationnel se construit autour de la notion de topique. Ils mettent en jeu des concepts à propos desquels ils apportent des

9 Pour mettre au même niveau les connaissances du public sur le sujet, un court-métrage sur le sujet est projeté.

10 Il leur est également demandé de produire deux présentations orales de leur texte écrit, l'étude globale dans laquelle s'inscrit le travail de Berman & Nir-Sagiv (2007) portant à la fois sur la différence narratif/expositif et écrit/oral.

informations, à la différence des textes narratifs généralement organisés par des relations entre participants et événements.

I.5. Dans une visée computationnelle

Depuis la possibilité de numériser les informations et la naissance du World Wild Web, le nombre de données accessibles à tous (dont les linguistes) a explosé. On peut désormais récupérer à tout instant des textes de toute provenance, en toutes langues, sur n'importe quel sujet, de n'importe quel type, sous n'importe quel format. Nous avons ainsi à disposition une énorme masse de données en évolution permanente. Ce « réservoir » (Kennedy 1998) est pour l'instant encore difficile à gérer[11] et il est de plus en plus difficile d'y puiser des informations précises. Cet aspect quasi-illimité de la masse de données mises en ligne, en même temps qu'il est un avantage, en est son principal inconvénient. Plus il y a de données et plus il est difficile de s'y retrouver.

L'accès à l'information est complexe et comprend différents processus : (i) repérer les documents présentant l'information recherchée, *i.e.* évaluer si le contenu d'un document est pertinent par rapport à une requête, (ii) repérer à l'intérieur des documents sélectionnés l'information recherchée, (iii) présenter le document ou la partie du document de façon claire, fidèle et succincte afin d'offrir une liste de choix ergonomique à l'utilisateur.

> « Lorsqu'une personne recherche une information sur le web, que son intérêt soit clairement défini et spécifié avec des mots clefs précis, ou que son besoin soit vague et exprimé par des mots peu discriminants, les systèmes de RI (recherche d'information) renvoient généralement le même type de résultats : une liste ordonnée de documents, où seuls le titre et parfois un extrait comportant les mots de la requête permettent d'en évaluer la pertinence pour son besoin initial.
> Ces types de résultats conduisent irrémédiablement à devoir consulter le contenu des documents pour s'assurer de leur pertinence. » (Hernandez 2004:i)

Les besoins en TAL concernant la gestion de ce « monde sauvage » sont nombreux et de plusieurs ordres : classer des documents par type, par domaine, par thématique ou sujet ; trouver un document pertinent par rapport à une requête ; générer un résumé du document ; proposer une visualisation globale du document, pointer dans le document là où peut se trouver la réponse à la requête ; extraire ces informations ; permettre la navigation d'un segment pertinent à un autre ; etc.

Les premières applications faisant appel aux notions de l'EOD se basent sur des calculs de statistique textuelle. Ces travaux s'appuient essentiellement sur le lexique contenu dans les textes et notamment sur la fréquence d'apparition des mots. Ainsi, des méthodes de segmentation thématique (*i.e.* localisation des ruptures thématiques et de fait délimitation de segments thématiquement homogènes) ont été élaborées, basées sur des calculs de fréquences relatives ou de récurrence lexicale, comme le TextTiling de Hearst (1997) appliqué au français par Ferret & Grau (1998, 2000). Ces techniques consistent, de façon simplifiée, à calculer la récurrence lexicale au fur et à mesure du texte en prenant en compte des fenêtres de mots dont le nombre est variable selon la tâche visée. Tant que les résultats issus du calcul de récurrence lexicale sont identiques, l'algorithme considère qu'il y a continuité thématique. Par contre, lorsque le résultat dans une fenêtre diffère de celui de la fenêtre suivante, une rupture thématique est localisée.

11 Le net tout azimut est d'ailleurs de plus en plus contesté en tant que corpus. Ce sujet fait l'objet d'un workshop particulier dans le cadre des campagnes CLEANEVAL (http://cleaneval.sigwac.org.uk/). Le 3e Workshop « Web as Corpus » (WAC3) se déroule à l'UCL en Belgique : http://central.fltr.ucl.ac.be/wac3/.

Dans la même veine, le résumé automatique d'un document a généralement consisté en une extraction de phrases jugées saillantes par des calculs statistiques, ce que proposent par exemple les différents travaux basés sur l'Analyse Sémantique Latente (LSA) (*Latent Semantic Analysis – LSA*) que l'étude de Kintsch (2002) illustre bien.

L'avantage de ces méthodes, qui ne se basent aucunement sur une analyse linguistique des textes, est qu'elles sont robustes et tout-terrain. Cependant, certains chercheurs – dont nous faisons partie – pensent que le TAL gagnerait en précision s'il appliquait des analyses plus fines, prenant par exemple en compte la structure du document ou son appartenance à un certain type. Simone Teufel est une pionnière dans cette approche :

> "We are interested in a formal description of the document structure of scientific articles from different disciplines. Such a description could be of practical use for many applications in document management; our specific motivation for detecting discourse structure is quality improvement in automatic abstracting" (Teufel 1998:1)

L'idée que la connaissance de la structure d'un texte améliore le résultat de certaines tâches en TAL rejoint notre hypothèse d'une sémantique de l'organisation du discours (*cf.* Enjalbert 2005). L'évaluation d'une telle hypothèse fait d'ailleurs l'objet de travaux très récents et en cours (Zerida *et al.* 2006, Pimm 2006). Si, comme nous le pensons, certaines expressions servent à préciser le cadre dans lequel les informations relatées sont pertinentes, la connaissance de ce cadre peut se révéler très utile pour représenter le contenu du document. Mais les linguistes sont encore loin d'une description complète et fiable de l'organisation discursive. Pour cela il subsiste de grands besoins en méthodes et en outils pour aborder de gros corpus en EOD. Le chapitre VI détaille en partie cette problématique.

L'idée d'une visée computationnelle a fondamentalement influé sur notre analyse linguistique. Le choix d'étudier l'organisation discursive dans des textes expositifs longs est secondairement lié à cette visée[12]. Ces textes sont effectivement plus susceptibles que d'autres de faire l'objet d'une recherche d'information et de traitements automatiques. Le modèle de segmentation thématique développé par Hearst (1997) s'appuie sur ce type textuel pour les mêmes raisons.

> "text-tiling is geared towards expository text ; that is, text that explicitly explains or teaches, as opposed to, say, literary texts, since expository text is better suited to the main target of applications of information retrieval and summarization" (Hearst 1997:35)

C'est également par cette conception applicative que se justifie notre volonté d'automatiser les méthodes d'investigation. Ce désir d'automatisation implique alors des caractérisations opérationnelles des phénomènes linguistiques étudiés, ce qui permet ensuite d'appliquer les méthodes d'observation, d'exploration et d'analyse sur d'autres corpus, pour d'autres études. Il y a dans ce travail, en plus d'une analyse linguistique, la proposition d'une méthodologie pour étudier des phénomènes linguistiques dépassant l'unité phrase ainsi que l'élaboration d'un outil d'annotation et d'exploration de phénomènes discursifs.

Cependant, la description de cet outil ne fera pas l'objet d'un chapitre en soi, le contenu de cette thèse étant principalement orienté vers l'analyse linguistique et non vers la visée applicative. Il n'y a d'ailleurs pas d'application précise envisagée et donc pas d'évaluation de cet outil par rapport à une tâche quelconque. De plus, l'outil utilisé pour cette analyse est encore au stade de conception : premièrement, le programme est trop opaque pour être exploitable et adaptable à d'autres utilisateurs ; deuxièmement, de nombreux modules sont encore nécessaires pour permettre l'exploration de phénomènes discursifs. Enfin, il reste encore à effectuer une réelle évaluation de la précision et du rappel d'un tel programme (une petite évaluation a cependant été effectuée et figure en annexe J).

12 La raison principale est exposée en I.4. Des textes nécessitant une organisation.

Partie 1.

Connaissances d'arrière-plan

« Fond »

Chapitre II
L'Organisation du discours

"Cohesion is a process because discourse itself is a process. Text is something that happens, in the form of talking or writing, listening or reading. When we analyze it, we analyze the product of this process ; and the term 'text' is usually taken as referring to the product." (Halliday 1985:290)

Sommaire

Un texte est le résultat d'un acte de communication, tel que l'illustre très schématiquement la petite Laura van Beek.

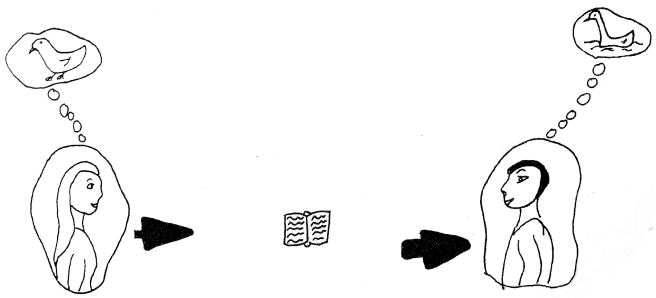


Figure II.1 : Communication through written text : From a cognitive representation to text, to cognitive representation (as seen by Laura van Beek, 9 years old) (Sanders & Spooren 2001:2)

En commentant ce dessin, Sanders & Spooren décrivent les processus de production et de compréhension des textes écrits.

> "There is a producer who has a cognitive representation of what she intends to communicate ; this is formulated in a linguistic code, called the text, and this text is decoded by the interpreter who can be said to understand a text once he has made a coherent representation of it. This view fits theories that describe the link between the structure of a text as a linguistic object, its cognitive representation and the processes of text production and understanding." (Sanders & Spooren 2001:2)

Dans cet échange de représentations, le texte est une étape, un intermédiaire, un messager entre un producteur : le **locuteur** ; et un destinataire interprétant : l'**allocutaire**. Dans cet échange, le texte, tout en véhiculant une représentation mentale, manifeste le contexte dans lequel il a été produit : sa construction (mise en forme du document, organisation globale, syntaxe) et le vocabulaire utilisé reflètent les paramètres extra-linguistiques à l'origine du texte. Nous avons évoqué en introduction la différence entre les textes narratifs et les textes expositifs. Cette différence provient clairement du contexte et non du texte lui-même ; et il est évident qu'il faut prendre en compte ces paramètres extra-linguistiques pour étudier l'organisation discursive. Notre étude étant une étude en corpus, nous reviendrons plus en détail sur la question des types textuels dans la présentation des linguistiques de corpus (chapitre VI).

Sans perdre de vue ces considérations, nous définissons la notion de texte comme une unité fonctionnelle intervenant dans un échange langagier. Nous envisageons l'unité texte selon le modèle de la **S**ystémique **F**onctionnelle – SF (Halliday 1985) qui distingue trois composantes définitoires, trois niveaux d'analyse :

- la composante *interpersonnelle* se rapporte à la manifestation des intentions discursives et de la considération de l'autre. Le texte est alors envisagé en tant que résultat d'un échange.

- la composante *idéationnelle* concerne l'expression du contenu : le fait qu'un texte est "au sujet" de quelque chose relatif au monde, à l'expérience du locuteur. En ce sens, le texte véhicule une représentation mentale.

- et enfin, la composante *textuelle* intègre tous les moyens linguistiques permettant d'organiser le texte en tant qu'objet physique, en tant que message.

Ces trois composantes sont autant de facettes par lesquelles étudier l'organisation du discours. On peut caractériser un texte par l'intention auteuriale qu'il dégage (sa visée discursive), par son sujet (sa thématique) ou par sa structure (sa forme). Pour interpréter un texte, le lecteur s'appuie sur des indices linguistiques et extralinguistiques émanant de chacune des trois composantes.

Avant même de commencer à produire un texte, le locuteur a défini sa thématique et le genre dans lequel il inscrit ses propos. De même il sait approximativement la forme que doit avoir le résultat de sa production, ainsi que les personnes auxquelles le texte s'adresse.

Dès la prise en main d'un texte, le lecteur associe ce texte à un genre discursif (il s'agit d'un roman, d'un article scientifique, d'un essai, etc.) et à une thématique. L'association à un genre et à une thématique est influencée par de nombreux facteurs extra-linguistiques (la source du texte ou comment il se l'est procuré, les personnes qui lui ont parlé de ce texte (ou simplement de l'auteur de ce texte), les autres textes qu'il a lu du même auteur, etc.) et linguistiques (le titre du texte, sa table des matières, sa longueur, sa structure visuelle : présence d'un résumé, présence d'illustrations, etc.)

II.1. Production et Interprétation

Le texte est le résultat (output) de processus de production tout en étant l'entrée (input) de processus d'interprétation. Ces deux types de processus, que l'on soit en discours écrit ou oral, sont confrontés à un défi commun : le *manque d'isomorphisme* (Heurley 1997:181) entre la linéarité du texte et la multidimensionalité de la représentation. Ce manque d'isomorphisme implique chez le locuteur un problème de linéarisation (Levelt 1981) et chez l'allocutaire, un problème de délinéarisation.

> "language production, whether oral or written, is strictly linear and time-dependent. This means that, at a given time, only one piece of information can be related. It follows that one of the main problems confronting speaker and writer is how to present information in a linear format (*i.e.* to linearize information) when such information is rarely stored in a linear structure in the mental model." (Fayol 1997:157)

Cette caractéristique – la linéarisation – est obligatoire à tout texte et oblige le locuteur à ordonner ses mots, ses propos. Les structures linguistiques envisagées dans cette thèse contribuent à l'organisation discursive en ce sens qu'elles sont à la fois :

- des instructions (ou des traces) laissées par le locuteur lors des processus de production (linéarisation de sa représentation),
- des indices utilisables par l'allocutaire pour la compréhension du texte (la délinéarisation de cette représentation).

L'organisation du discours est ici envisagée en se plaçant plutôt du côté de l'interprétation, même si nous appuyons parfois nos idées sur des travaux portant sur la production de texte. Les indices textuels étudiés sont envisagés par rapport à leur rôle dans l'interprétation, du point de vue de l'allocutaire ; et non par rapport à leur rôle dans la production. La distinction entre des éléments placés intentionnellement par le locuteur afin d'être reconnus comme tels (les « *cue-phrases* » de Grosz & Sidner 1986) et des éléments laissés non-intentionnellement ne transparaît pas au niveau des processus d'interprétation. Les indices retenus ici nous intéressent essentiellement pour leur fonction au niveau des processus d'interprétation. D'un point de vue plus cognitif, ces indices aident l'interprétant à se construire une représentation mentale cohérente.

II.1.1. La construction d'une représentation mentale

Les modèles de compréhension actuels décrivent la compréhension comme une construction progressive d'une représentation mentale du texte, construction dans laquelle interviennent trois niveaux de représentation :

- La représentation de surface qui correspond à la forme linguistique du texte ;
- La représentation sémantique qui correspond au sens « littéral » véhiculé par les propositions du texte (nous retrouvons ici l'objet d'étude de la sémantique formelle : modélisation du sens, avec l'idée de découper le sens en propositions minimales) ;
- La représentation mentale qui correspond au sens personnel et actualisé que se construit l'allocutaire en associant à la représentation sémantique ses connaissances propres (rôle des inférences, largement étudié en linguistique cognitive).

Bien que cette étude aborde les textes par le premier niveau de représentation, quelques précisions sur la notion de « construction d'une représentation mentale » sont nécessaires pour justifier certaines orientations. Les termes choisis pour décrire les processus en jeu dans cette construction sont empruntés à Werth (1999) qui présente les processus de production et d'interprétation de façon très simple.

Selon Werth, la construction d'une représentation mentale est comparable à la construction d'un monde (imaginaire) dans lequel les propositions issues du texte sont cohérentes et forment un sens global : un *text world*. Ce « text world »[13] correspond aux « structures » de Gernsbacher (1990-1997, Gernsbacher & Robertson 2002), aux « espaces mentaux » de Fauconnier (1984, 1985), aux « modèles mentaux » de Johnson-Laird (1983), et aux « modèles de situation » chez Kintsch & Van Dijk (1983).

Dans ces trois dernières théories, le texte est considéré comme une micro et une macro structure qui active des « modèles » préexistants en dehors du texte, dans la mémoire de l'allocutaire. Ces modèles mentaux sont des représentations individuelles d'un état de chose, d'une expérience, d'une réalité... Ainsi, dès les premiers éléments d'un texte, le lecteur va activer un modèle, une représentation provisoire de ce que communique le texte. Cette représentation provisoire explique les relations implicites qui peuvent s'établir entre les entités exprimées dans le texte (phénomènes d'inférences, relation de causalité, etc.) En poursuivant sa lecture, le lecteur affine ou change sa représentation mentale, adapte son modèle à celui décrit dans le texte.

Un « text world» est d'abord défini par le texte qui donne les informations de base à sa construction : les fondations (*qui/quoi, quand, où*).

> "A text world is a deictic space, defined initially by the discourse itself, and specifically by the deictic and referential elements in it. […] The deictic information establishes the place, time and deployment of the text world […] while the referential items establish which entities are present. However, the latter have a second and very important function : that of evoking frames (areas of memory which relate to areas of experience and knowledge encoded as complex conceptual structures)." (Werth 1999:51-52)

L'étude de l'organisation du discours (EOD) par le biais de ces trois dimensions (*qui/quoi, quand, où*) correspond à l'analyse de sa composante idéationnelle et plus particulièrement de la représentation de notre expérience.

> "A fundamental property of language is that it enables human beings to build a mental picture of reality, to make sense of their experience of what goes on around them and inside them." (Halliday 1985:101)

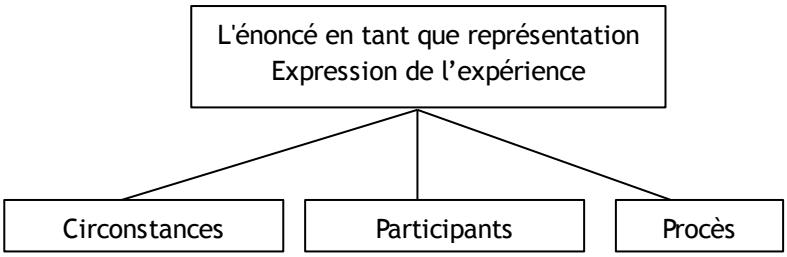


Figure II.2 : Composants pour l'expression de l'expérience en SF

D'après Halliday, l'expression de l'expérience se situe principalement au niveau de la proposition qui est l'unité grammaticale la plus signifiante pour représenter des procès propres à notre expérience d'être humain tels que des actions, des processus, des événements, des états, des phénomènes du monde réel, etc[14]. Pour exprimer ou interpréter ces procès, trois éléments entrent en jeu : le procès lui-même, le(s) participant(s) impliqué(s) dans ce procès et les circonstances associées à ce procès.

Nous retrouvons dans la figure II.2 ce que Werth nomme les entités (les « participants ») et les localisations spatiales et temporelles (les « circonstances ») qui constituent les éléments de bases du « text world » dans lequel se

13 Nous conservons le terme anglais pour plus de lisibilité, la traduction française nous obligeant à mettre une préposition : « monde du texte » qui peut, dans certains cas alourdir et/ou entraîner des ambiguïtés.

14 Le terme d'*éventualité* peut aussi être utilisé, traduction littérale de 'eventuality' que l'on trouve en sémantique formelle.

déroulent les différents procès. Werth parle alors de *World-Builders*, éléments qui permettent de situer, d'ancrer l'information dans le contexte courant. Ces *world-builders* sont à rapprocher des « introducteurs d'espaces mentaux » (*space-builders*) de Fauconnier qui sont décrits comme des signes linguistiques particuliers introduisant automatiquement de nouveaux espaces mentaux.

De façon plus générale, nous définissons ces *« world-builders »* par la fonction de *setting* que l'on peut traduire en français par « qui posent le décor », « qui situent le cadre », « qui fixent des repères ». Nous utiliserons fréquemment le terme anglais pour caractériser cette fonction, du fait de l'absence de terme efficace en français. Cependant, pour référer non pas à la fonction, mais aux expressions qui sont à la base de la construction des « text worlds », nous utiliserons les termes d'adverbiaux circonstanciels.

C'est principalement sur ces éléments adverbiaux circonstanciels que notre attention va porter, en nous focalisant sur la réalisation textuelle des entités/participants et des *settings* dans les textes.

II.1.2. Entités, référence et accessibilité

Les entités participant à un « text world » peuvent être de plusieurs natures : personnages, idées, phénomènes, lieux, temps, sentiments, etc. En linguistique, on parle plus communément de référents. Comme l'affirme Kleiber, « la référence est une notion à la fois claire et équivoque ». Claire, parce qu'on peut associer à la notion de *référence* une acception générale de type : « la fonction qui permet aux signes linguistiques de renvoyer à la réalité extra-linguistique » (Kleiber 1981:11). Équivoque, car cette fonction reste un mystère et donne ainsi lieu à de nombreuses théories divergentes sur la relation entre un signe et une représentation de la réalité : un référent. En linguistique, les différentes études de l'expression de la référence portent notamment sur la façon dont un référent peut arriver et être maintenu dans le discours. Nous parlons alors de son statut (informationnel) dans le discours qui se définit selon : (i) le mode d'introduction de l'entité dans le « text world » et (ii) le statut avec lequel elle participe à ce monde.

> The establishment of entities is one of the basic acts of text world building. A world is first defined by deictic expressions of place and time, and is then furnished with entities by reference establishment. Reference maintenance is then the process of keeping entities in the active register of discourse." (Werth 1999:158)

Beaucoup de travaux en psychologie cognitive concernent l'activation des référents en discours. La tendance actuelle consiste à mesurer des sortes de cartes topographiques d'activation au fil de la lecture pour chacune des entités présentes dans le texte, Van den Broek et al. (1996) parlent alors de paysages d'activation. Ces paysages permettent de déterminer les différentes entités sur lesquelles porte l'attention de l'allocutaire ainsi que la durée avec laquelle elles restent présentes en focus d'attention (Gernsbacher 1990, Kintsch & Van Dijk 1983, Grosz & Sidner 1986, Van den Broek *et al.* 1996 etc.) Dans ces différents modèles, la lecture est envisagée comme un processus cyclique. Chaque cycle correspond à un empilement[15] des informations arrivantes dans une mémoire temporaire appelée « mémoire tampon[16] ». Lorsque notre mémoire tampon sature ou qu'un indice linguistique nous en donne l'instruction (par exemple un changement de paragraphe), nous 'rangeons' notre mémoire tampon et un cycle de lecture s'achève. Les informations reçues et mémorisées temporairement sont intégrées dans la mémoire de travail (ou mémoire principale), l'interprétation du segment de texte est effectuée et la mémoire tampon est 'nettoyée', prête pour

15 Nous utilisons ce verbe en référence à la notion de « pile de focus » chez Grosz & Sidner (1986).

16 La notion de « mémoire tampon » (*Buffer Memory*) est issue du lexique de l'informatique pour désigner une mémoire dans laquelle sont stockés les fichiers temporaires avant d'être classés dans les dossiers adéquats. Elle correspond ici aux piles de focus chez Grosz & Sidner ou aux caches chez Walker (2000).

un nouveau cycle[17]. Ce nettoyage consiste à classer les informations stockées : certaines vont être archivées dans la mémoire principale, d'autres vont être retenues dans la mémoire tampon restant ainsi actives.

Pour mesurer le degré d'activation d'une entité, nous avons mentionné les notions de focus d'attention et de durée de focus d'attention. Ces notions correspondent à des mesures établies au niveau de la représentation mentale. Une autre mesure établie au niveau du discours lui-même s'ajoute à celles-ci : l'accessibilité du référent dans le texte. Une distinction est ici faite entre le statut du référent au niveau cognitif et son statut dans le texte. La Figure II.3 représente ce degré d'accessibilité selon le degré d'activation cognitive.

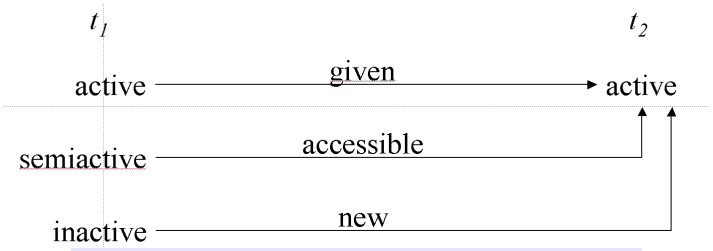


Figure II.3 : Activation States, Activation Costs, and Time (Chafe 1994:73)

Au niveau de l'interprétation, plus un référent est donné et donc déjà actif, moins la tâche est coûteuse pour l'allocutaire puisqu'il n'y pas de nouvelle entité à intégrer dans la mémoire tampon. En ce sens, on peut dire que la continuité référentielle allège le processus d'interprétation. À l'inverse, l'arrivée d'un nouveau référent entraîne une charge cognitive plus lourde. La charge cognitive est d'autant plus lourde que le référent est nouveau, non seulement pour la mémoire tampon, mais également pour la mémoire principale activée. Nous avons là une distinction entre un référent accessible car en relation lexicale avec les référents déjà évoqués et un référent complètement nouveau, *i.e.* qui n'appartient à aucun réseau sémantique activé.

L'augmentation du coût de la tâche d'interprétation lors de l'arrivée d'une entité inactive peut être prévue par le locuteur. Il peut faciliter la reconnaissance de ce nouveau référent en l'introduisant plus ou moins progressivement (utilisation de constructions syntaxiques spéciales, activation de connaissances d'arrière plan orientant l'interprétation, etc.)

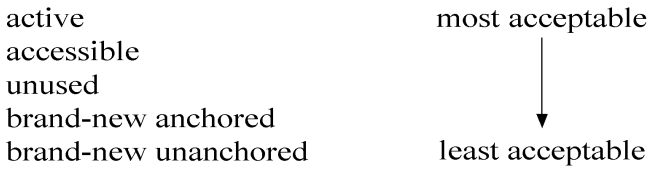


Figure II.4 : L'échelle d'acceptabilité topicale (Lambrecht 1994:165)

Lambrecht (1994) associe une échelle d'accessibilité similaire à un degré d'acceptabilité des expressions topiques. « Le topique d'une phrase est la chose *à propos* de laquelle est la proposition exprimée par la phrase » (Lambrecht 1994:118). Selon ces échelles parallèles, un topique est plus acceptable s'il est actif que s'il est inactif. Cette hypothèse constitue une des hypothèses de base de la théorie du centrage (Grosz *et al.* 1995, cf. III.3.2).

La distinction entre les trois types de référents nouveaux est issue de la terminologie développée par Prince (1981) qui mêle les statuts d'activation cognitive à ceux d'accessibilité discursive. Le degré d'accessibilité d'un référent revient alors à le situer sur une échelle allant des référents connus car donnés dans le discours aux référents inconnus et donc nouveaux pour le discours. Les référents dits 'brand-new' (*flambants neufs*) correspondent à des référents non

17 Les études oculométriques ont montré que l'oeil, lors des processus de lecture, fixait les espaces typographiques entre les mots, ce qui correspondrait à une 'pause de perception' permettant le traitement cognitif des informations précédemment perçues.

identifiables par le destinataire au moment où il les reçoit, dont il n'a pas connaissance ; contrairement aux référents encore non utilisés (*unused*) qui ne sont nouveaux que pour le « text world » en cours de construction. La notion d'ancrage permet de faire la distinction entre les éléments nouveaux qui peuvent ou ne peuvent pas être rattachés à une entité déjà existante dans le « text world ».

Les travaux linguistiques portant sur la mesure de l'accessibilité des référents consistent pour la plupart à découvrir des indices permettant de mesurer le degré d'accessibilité d'un référent ainsi que son maintien dans le « text world ». Le degré d'accessibilité d'un référent peut s'exprimer selon la structure informationnelle des énoncés (Lambrecht 1994), la forme et fonction syntaxique des syntagmes nominaux référentiels (Ariel 1990, Cornish 1996-2003, De Mulder 1994-2000, Gundel *et* al. 2000, Lambrecht 1994, Walker *et al.* 1998, Kleiber 1981-1994), l'inscription – ou plutôt la récurrence – des expressions des référents dans les textes (Givón 1990 cité par Chafe 1994:180, Hearst 1997). Les indices les plus étudiés dans ces travaux concernent la forme des Syntagmes Nominaux – SN – et notamment leur détermination. Cela permet alors d'établir des échelles de correspondance entre degrés d'accessibilité et formes morpho-syntaxiques.

> "I (will) argue that all referring expressions in all languages are arranged on a scale of Accessibility. Although actual Accessibility marking systems are to some extent language-specific, for the most part they are all based on a principled connection between marker form and degree of Accessibility. The more informative, rigid (unambiguous), and unattenuated the marker, the lower the Accessibility it is specialized for, and vice versa." (Ariel 1990:29)

Nous reviendrons sur certains de ces travaux dans la partie V.4.3 et présenterons l'échelle d'Accessibilité d'Ariel à cette occasion.

II.1.3. Des circonstances : le temps et l'espace

> "Time and space are the basic categories of our experience and our cognition, and without efficient communication about them, no well-coordinated collective action, hence no human society, would be possible. Therefore, all natural languages we know of have developed a rich repertoire of means to express temporality and spatiality." (Klein 1994:1)

L'expression de la temporalité et de la spatialité est indéniablement liée à celle des circonstances dans lesquelles évoluent les entités d'un monde. Comme le dit Werth, les référents sont introduits et maintenus dans un monde préalablement défini spatio-temporellement. Cependant, il ne faut pas confondre expression de la temporalité et de la spatialité et expression des circonstances. Si l'on reprend la trilogie circonstants – participants – procès qui constitue selon la systémique fonctionnelle la base pour l'expression de l'expérience (figure II.1 p.31), chacun de ces trois composants peuvent être de nature temporelle ou spatiale. Les études sur les temps et l'espace en discours montrent d'ailleurs une forte attraction pour l'étude du procès, surtout au niveau de la temporalité (Reichenbach 1947, Vendler 1967). En effet, excepté pour les langues n'ayant pas de système de conjugaisons, la temporalité est exprimée de façon intrinsèque dans le temps que portent les verbes.

Dans des langues isolantes n'ayant pas de morphèmes flexionnels de temps, les verbes (et donc le procès) n'expriment pas la temporalité de l'énoncé. D'autres moyens d'ancrer l'énoncé dans le temps sont alors utilisés, comme les adverbiaux spatio-temporels qui expriment les circonstances du procès[18] et qui nous intéressent spécialement ici.

18 La thèse de Do-Hurinville (2004) intitulée «Temps et aspect en Vietnamien : étude comparative avec le Français » propose une étude détaillée de l'expression du temps dans une langue isolante dépourvue de temps verbaux.

L'expression de la spatialité aussi est largement étudiée au niveau de l'expression du procès, notamment par l'étude des verbes de mouvement et apparaît alors comme étant tributaire du système de la langue utilisée. Une autre façon d'aborder l'expression de temporalité et de spatialité consiste en l'étude des relations exprimées par les prépositions – surtout pour l'espace – et les connecteurs – essentiellement pour le temps. Mais là aussi, à chaque système correspond des listes plus ou moins fermées. De plus, tout comme pour l'étude des verbes de mouvement ou des temps verbaux, l'observation de telle forme concerne essentiellement le niveau phrastique. De ce fait, la plupart des travaux sur le sujet sont orientés vers la détermination fine des différents fonctionnements de telle ou telle forme plutôt que vers la caractérisation formelle de tel ou tel fonctionnement discursif.

En dehors des verbes et des prépositions, le temps et l'espace peuvent être exprimés par des adverbiaux spatiaux et temporels, c'est-à-dire des éléments qui expriment les localisations spatiales ou temporelles du procès, ou en d'autres termes, l'expression des circonstances. Ces éléments ne dépendent pas nécessairement de la structure de la langue étudiée et, ce faisant, sont présents dans toutes les langues, ce que souligne Klein (1994) et ce qui constitue l'objet d'une typologie établie par Thompson & Longacre (1985).

> "There is much less research on temporal adverbials or particles, although they are not only ubiquitous – not all language have tense or aspect, but all languages have a wealth of temporal adverbials – they are also much more refined and richer in their expressive power. Their analysis is often considered to be a part of lexical semantics, whereas tense and aspect are deeply rooted in the structural organisation of language, and hence are more prone to excite the linguist's attention ." (Klein, 1994:2)

De plus, comme le remarque Klein, les adverbiaux sont moins intégrés dans des structures propres au système de la langue, ce qui leur confère une relative indépendance, et sont nécessaires à certaines fonctions discursives jouant au niveau de la segmentation textuelle. En effet, selon l'hypothèse de l'encadrement du discours (Charolles 1997, cf. partie III.3.3), ces 'introducteurs de cadre idéaux' permettent d'indexer une portion de texte dépassant la taille de la phrase en regroupant plusieurs propositions par un même critère d'interprétation. Par ailleurs, selon l'hypothèse des continuités texto-stratégiques, *« Text-strategic continuities* » – TSC (Virtanen 1992, cf. III.3.4), ces expressions, lorsqu'elles apparaissent en séquence, organisent la portion de texte selon une TSC spatiale ou temporelle.

Pour parler de localisation spatiale/temporelle, nous avons utilisé le terme « circonstance » qui est plus général et en même temps plus réduit que celui de l'« expression du temps et de l'espace ». Plus général, car des rôles sémantiques autres que le temps et l'espace peuvent exprimer les circonstances d'un procès (la manière, la cause, le but, etc.) Plus réduit, car, comme nous l'avons déjà signalé, il ne s'agit plus de l'expression mais de la localisation spatio-temporelle, c'est à dire du repérage dans le temps et l'espace du procès.

> « Le complément circonstanciel précise l'idée du verbe en marquant la connexion de l'action avec un repère (temps, lieu, etc.) situé autour d'elle dans le monde des phénomènes. » (Grévisse 1993:477)

On retrouve approximativement cette notion de circonstance ou de (complément) circonstanciel dans le concept de *setting*. Cependant, en anglais, une distinction est faite entre : d'une part, les *circumstances* c'est-à-dire l'expression des circonstances au niveau sémantique (circonstants de temps, de lieu, de manière, de but, de condition, etc.) ; et d'autre part, les *settings* qui permettent la localisation du procès et/ou des entités par rapport à un repère spatial/temporel, un domaine de connaissance (une notion), un élément du « text world ».

> "A setting can be defined as "a global" inclusive region within which an event unfolds or a situation obtains" (Langacker 1991:553)

Ainsi, d'un point de vue structurel, un *setting* a une fonction extra-prédicative[19] (indépendante de la prédication principale), ce qui n'est pas définitoire d'une *circumstance.* Un *setting* constitue un élément permettant de construire le « text world », de planter le décor.

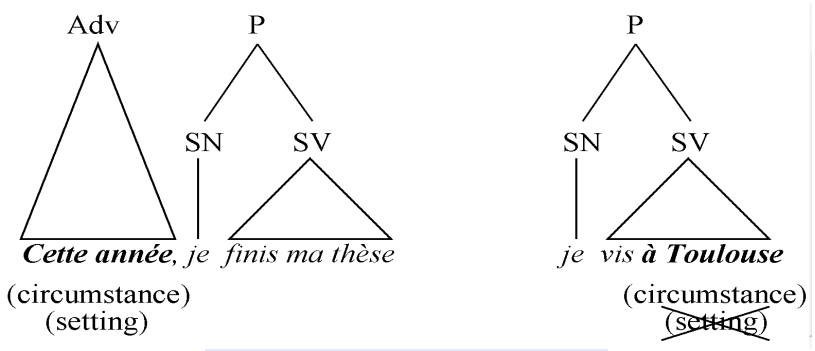


Figure II.5 : Distinction entre circumstance et setting

Le manque de distinction qui existe en français est d'ailleurs problématique : « Dans les grammaires, classiques et modernes, le concept de complément circonstanciel – aussi bien, d'ailleurs, que celui de circonstant – paraît écartelé entre deux définitions : l'une, positive, associe ce type de complément à la catégorie sémantique des circonstances ; l'autre négative, consiste à tenir pour circonstanciel tout complément dont on ne sait quoi faire » (Gosselin 1990:37)

Nous ajouterons à la dernière phrase une petite précision : « dont **le syntacticien** ne sait quoi faire » ; car les circonstanciels, par leur situation périphérique, acquièrent la capacité d'étendre leur portée au delà de la phrase d'accueil, ce qui leur confère un statut des plus intéressants pour l'EOD.

Dans cette étude, le temps et l'espace seront abordés :

- Par les adverbiaux extra-prédicatifs, c'est-à-dire des éléments périphériques et non intégrés à la proposition ;
- Par les **adverbiaux extra-prédicatifs circonstanciels**, c'est-à-dire la localisation spatio-temporelle et non l'expression générale du temps et de l'espace ;
- Pour leur rôle dans la définition du « text world ».

Par définition, un adverbial extra-prédicatif circonstanciel a une fonction de *setting*. On retrouve cette notion chez Firbas (1986:47-50) qui distingue les « *adverbials of setting »* des « *adverbials of specification ».* Les premiers sont extra-prédicatifs et généralement situés en position initiale, ce qui leur confère un très faible degré de dynamisme communicatif. Les seconds sont intra-prédicatifs, généralement situés après le verbe principal, ce qui leur confère un degré de dynamisme communicatif plus élevé. D'un autre point de vue, Enkvist (1976) distingue les « *setting adverbials »* des « *valency adverbials »* selon le degré de liberté de mouvement qui les caractérise. Les premiers sont plus mobiles et plus facilement topicalisables alors que les seconds sont obligatoires, leur lien très fort avec le verbe les forçant à rester en fin de phrase. Lambrecht (1994) parle lui de « *scene-setting adverbials »* qui correspondent bien à notre idée de « planter le décor »[20]. De façon générale, et en accord avec ces trois points de vue, nous parlerons d'adverbiaux circonstanciels[21].

19 Cette fonction correspond à la fonction « exophrastique » selon les termes de Guimier (1996). Guimier définit par cette fonction les adverbes qui restent à l'extérieur de l'énoncé et s'appliquent à toute la proposition, contrairement aux adverbes à fonction endophrastique qui sont intégrés à l'énoncé en s'appliquant à un élément de la proposition, généralement le prédicat. Les termes « extra/intra-prédicatif » qui sont équivalents seront préférentiellement utilisés au fil de cette thèse.

20 Nous renvoyons également à la définition des « *world builders* » donnée par Werth (1999:186-188)

La localisation spatiale et temporelle est d'autant plus importante à la construction d'un « text world » lorsque l'on se situe à l'écrit. En effet, l'écrit implique un déplacement spatio-temporel entre la situation de production, la situation du « text world » et la situation d'interprétation. Nous renvoyons ici à la notion de *« displacement* » (Chafe 1994:195-295). Il est alors indispensable que le locuteur définisse les fondations temporelles et spatiales du « text world » dans lequel 'vivent' les entités et se déroulent les procès relatés, comme le montre l'exemple II.1. Cet exemple montre le titre et premier paragraphe d'une section par lesquels l'auteur présente les référents majeurs dont il va parler (le marché pétrolier américain et l'intervention publique) dans un contexte particulier (entre les années 1920 et 1980). L'article dont est extrait cet exemple est intitulé ***Les Etats-Unis et le pétrole : De Rockefeller à la Guerre du Golfe***.

(II.1) ***SOIXANTE ANS D'INTERVENTION PUBLIQUE*** [titre niveau 1]
A partir des années 1920 et jusqu'au début des années 1980, *le marché pétrolier américain a vécu sous un régime de très forte intervention publique. On distinguera les mesures dites de proration, le contrôle des importations, et le contrôle des prix. [GEOPO_12].*

Nous notons la forte cohésion lexicale entre le titre de section et ce premier paragraphe (*Soixante ans -> à partir des années 1920 jusqu'au [...] 1980*, et les deux occurrences d'*intervention publique*). Le deuxième référent majeur *(le marché pétrolier américain*) est actif durant tout l'article, depuis le titre de l'article même : *les États-unis et le pétrole*. Nous voyons bien dans cet exemple comment les titres peuvent poser les fondations d'un « text world » et comment l'adverbial temporel reprend et précise la localisation temporelle fondatrice.

II.1.4. La construction de sous-représentations

Avant de refermer cette partie sur la construction d'une représentation mentale, il est important d'aborder la notion de sous-représentations. Il s'agit de concevoir la construction du « text world » non pas de façon monolithique, mais comme une succession de créations de sous-mondes qui contribuent à la construction du « text world ». Ces « petits mondes »[22] peuvent présenter une définition spatiale/temporelle différente de celle du « text world » général, ainsi qu'une palette d'entités spécifiques ayant entre elles des relations spécifiques.

L'exemple II.1 ci-dessus illustre tout à fait l'idée des sous-mondes. En effet, dans l'histoire du marché pétrolier américain, il y a un sous-monde qui correspond à la période allant des années 1920 aux années 1980, et dans lequel *l'intervention publique* joue un rôle important. D'ailleurs, si l'on regarde les différents titres de sections de cet article, nous nous apercevons que la plupart ouvrent un sous-monde à l'intérieur du vaste monde que représente l'histoire du marché pétrolier américain de Rockfeller à la guerre du golfe. Le plan donné en (II.2) présente la liste des titres de sections de cet article, les titres exprimant des éléments fondateurs de sous-mondes y sont surlignés.

(II.2) *1. Une histoire d'entrepreneurs*
a. Au commencement était le droit
b. Ascension et chute de John D. Rockefeller
2. Soixante ans d'intervention publique
3. Le " moment Reagan " et l'option libérale
4. L'Amérique et le pétrole mondial
a. Quelques données sur l'approvisionnement pétrolier américain
b. Sécurisation et construction du marché mondial
c. L'Amérique a-t-elle le choix ? [GEOPO_12]

21 Nous pensions au départ que le terme « adverbiaux de setting » était plus approprié. Cependant, l'expression « adverbiaux circonstanciels » semble activer plus rapidement la notion à laquelle nous souhaitons référer. En anglais, nous aurions sans soucis utilisé le terme *setting adverbials*.

22 Nous ne faisons ici aucune référence aux « petits mondes » en théorie des graphes (Gaume 2004)

Dans le Structure Building Framework, Gernsbacher (1990, 1995, 1997, Gernsbacher & Robertson 2002) rejoint la notion de sous-mondes en utilisant les termes de sous-structures mentales mises en place lors des processus de compréhension :

> "According to the Structure Building Framework, the goal of comprehension is to build coherent mental representation or *structures*. At least three component processes are involved. First, comprehenders lay foundations for their mental structures. Next, comprehenders develop mental structures by mapping on new information when that information coheres or relates to previous information. However, when the incoming information is less coherent or related, comprehenders employ a different process : they shift or build a new substructure. Thus, most representations comprise several branching substructures." (Gernsbacher 1997:3)

Cette notion de sous-monde est importante pour l'EOD puisque le changement d'un sous-monde à un autre a de fortes chances d'être signalé dans le texte pour être perçu et interprété correctement. Les marqueurs de segmentation jouent un rôle majeur dans ce signalement. Si un sous-monde a une définition spatiale/temporelle différente, elle doit être explicitée ; si le sous-monde est hypothétique ou relève d'un point de vue spécifique ou d'un domaine de connaissance particulier, cela doit être précisé. Tout comme pour le « text world » général, la construction d'un sous-monde commence par l'installation du décor dans lequel les référents (actifs ou inactifs) vont être mis en relation et évoluer. Nous retrouvons ici la notion de cadres de discours définie dans Charolles (1997) comme des ensembles regroupant plusieurs propositions qui ont en commun un même critère d'interprétation[23].

> « La notion d'univers de discours […] est apparentée à celle d'« espaces mentaux » de G. Fauconnier (1984). Dans des publications ultérieures, G. Fauconnier parle de « domaines » et de « subdivisions cognitives », notions qui nous conviennent parfaitement mais qui anticipent […] sur la dimension cognitive du phénomène. Nous soulignerons en cours de route cette dimension qui est essentielle et qui n'est en rien discordante avec notre propos ; mais, pour commencer, l'idée d'univers de discours, plus neutre et peut-être plus grammaticale, nous suffira largement. » (Charolles 1997:6-7)

Il y a dans l'idée des cadres de discours une sorte de textualisation des sous-mondes de Werth, des subdivisions cognitives de Fauconnier, des sous-structures mentales de Gernsbacher. Alors que les modèles cognitifs abordent le phénomène de subdivision par son impact sur la construction d'un représentation mentale, le modèle ou plutôt l'hypothèse de l'encadrement du discours aborde ce phénomène de segmentation par son inscription dans le texte, par le fait qu'il existe à la surface du texte des marqueurs qui ont pour principal rôle le marquage d'une segmentation. Ce phénomène de segmentation textuelle est longuement développé dans le chapitre suivant et est étudié ici dans un type de discours particulier : le discours écrit.

II.2. L'écrit : un échange en différé

L'écrit est une forme de discours pour laquelle le texte n'est pas produit et reçu en direct : « la communication écrite est différée et hors situation » (Riegel *et al.* 1994:30). Le locuteur a donc tout le temps de revenir sur ce qu'il a écrit et l'allocutaire a tout le temps de revenir sur ce qu'il vient de lire. Mais parallèlement à ce confort, le discours écrit ne permet pas d'interaction entre le locuteur et l'allocutaire, les processus de production et d'interprétation ne se faisant pas simultanément. Lorentz (1999) considère la différence entre discours écrit et discours oral dans la façon de « construire une cohérence » : alors que l'oral permet une « négociation », l'écrit ne permet pas aux participants au discours de s'entraider dans la construction d'une représentation mentale.

23 Les différentes propriétés et natures de ce critère seront exposées dans la partie III.3.3.

> "Among other characteristics, writing differs from face-to-face interaction in the way coherence is constructed. In written communication [...] coherence cannot be explicitly negotiated. Instead, there is an implicit co-construction of meaning, and writers therefore have every reason for trying to be unequivocal and to make their ideas, intentions and arguments unmistakably clear. One way of doing this is to carefully signal logical relations and thereby 'signpost' the path to coherence for the reader. Consequently, when looking at the fabrication of coherence in written discourse, we need to pay special attention to those explicitly signposts of coherence, *i.e.* the text's cohesive ties." (Lorentz 1999:55)

Les processus de production et d'interprétation impliqués dans un discours écrit nous semblent différents de ceux impliqués dans un discours oral[24]. De son côté, le locuteur doit expliciter au maximum les référents qu'il aborde ainsi que les propos qu'il tient afin d'anticiper les incohérences de son discours et d'en assurer la bonne interprétation. L'allocutaire, lui, doit trouver dans le texte des indices lui permettant de décoder les intentions qui ont poussé le locuteur à écrire ce texte, et de se construire sa propre représentation. Lorentz (*op. cit.*) utilise l'image des panneaux de signalisation que l'auteur doit disposer le long de son texte, afin d'indiquer au lecteur le '*chemin de la cohérence'.* Ceci suppose évidemment qu'il existe une sorte de 'code tu texte' et que le lecteur connaisse ce 'code du texte'.

La partie II.3 revient sur les notions de cohérence et de liens de cohésion *(« cohesive ties »*) qui ne sont pas spécifiques au discours écrit. Ce que nous voulons souligner ici, c'est la grande importance qui est accordée, à l'écrit, aux signalisations de texte. Pour construire sa représentation mentale du discours, le lecteur se raccroche à des indices qu'il rencontre à la surface du texte. Ces indices lui permettent de se repérer dans le flot du texte et l'orientent dans son interprétation. Certains marqueurs sont spécifiques à l'écrit dans la mesure où ils n'existent que par la matérialité du texte. D'autres ne sont pas spécifiques à l'écrit, mais sont davantage utilisés à l'écrit du fait de cette distanciation entre le locuteur et le lecteur.

Comparé à Lorentz, et à l'instar de Heurley (1994, 1997), nous soulignons que les signalisations utilisées par le lecteur ne sont pas toujours laissées délibérément par le locuteur et peuvent parfois correspondrent à de simples « traces » des processus cognitifs de production. C'est précisément le cas des changements de paragraphes qui ne correspondent pas toujours à une intention bien définie de marquer quelque chose et qui pourtant sont toujours perçus comme une discontinuité dans le flot du texte.

II.2.1. Une mise en forme matérielle – MFM

> "La spécificité de l'écrit se traduit par la création de concepts originaux (page, ligne, marge, titre, énumération, note de bas de page, etc.)" (Luc *et al.* 2001:264)

Le **M**odèle d'**A**rchitecture **T**extuelle – MAT (Virbel 1986, Pascual 1991, Luc *et al.* 2001) définit le texte comme « un énoncé inscrit sur un support matériel, le procès d'inscription livrant des marques destinées à être appréhendées visuellement » (Luc & Virbel 2001:104). Ces marques peuvent être typographiques (police, gras, etc.) ou dispositionnelles (retraits, espaces verticaux, sauts de page, etc.) Toute structure de texte mise en forme par ces marques est alors appelé « objet textuel ». Bien évidemment, seuls les objets produits selon une intention auteuriale sont retenus – à distinguer de ceux causés par le support lui même*, i.e.* les limites de la page (*cf.* Lemarié 2006). Le MAT conçoit tout objet textuel comme résultant d'un acte de discours. La formulation de cet acte est appelé « contrepartie discursive développée » (Virbel 1986), ce qui permet de formuler et répertorier toutes les intentions auteuriales à la base de la mise en forme matérielle d'un texte. Les contreparties discursives correspondent à des

24 Cette position n'est absolument pas défendue par les modèles de compréhension (Gernsbacher 1990, Kintsch & Van Dijk 1983) qui considèrent que les processus cognitifs impliqués dans la compréhension ne dépendent ni du support ni de la modalité.

propositions construites autour de verbes dits « performatifs », dans la lignée des travaux d'Austin (1970) sur la définition des actes de discours. Ainsi, l'aspect matériel d'un texte correspond à un « métadiscours », tel que l'illustre la Figure II.6 où L désigne le locuteur-auteur du texte.

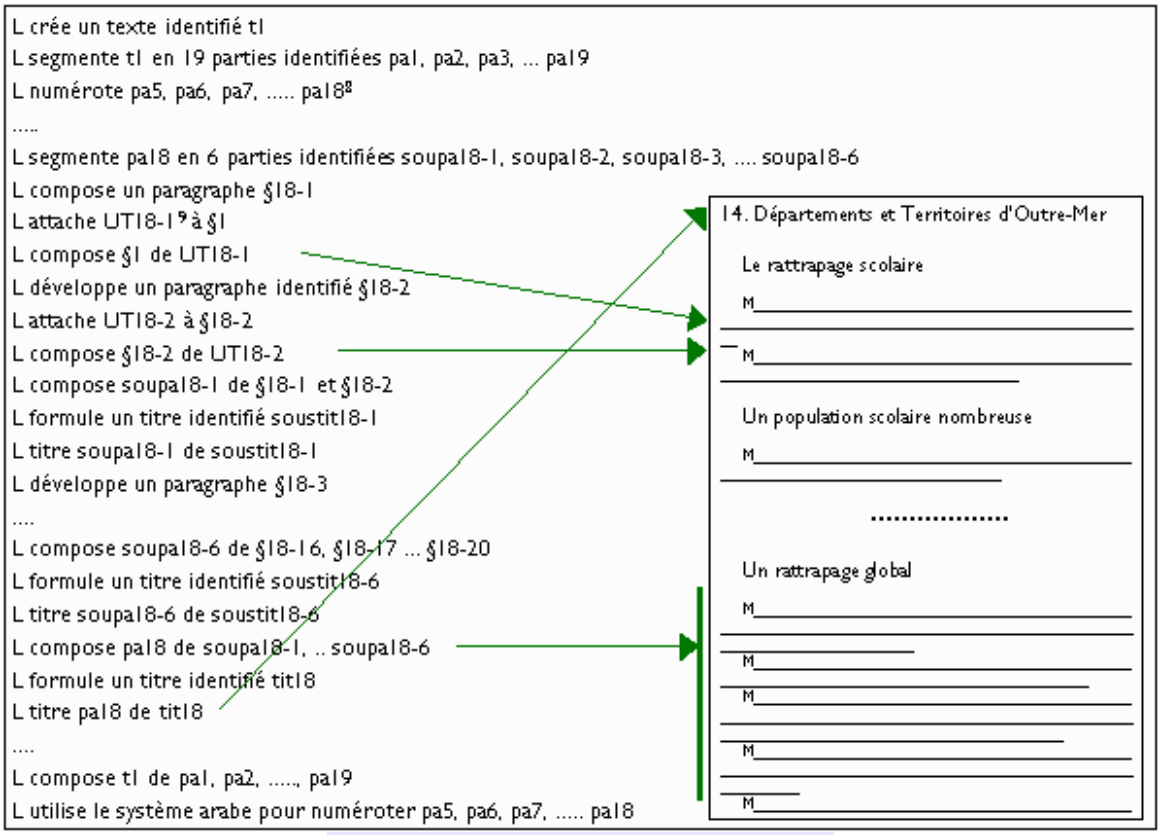


Figure II.6 : Image de texte et son « métadiscours » selon le MAT

L'écrit offre au locuteur la possibilité de marquer certaines parties, certains éléments ou unités du texte. Cette mise en forme confère au texte écrit la capacité à être abordé de façon non linéaire, par une vision globale, qui peut être décrite/résumée dans une table des matières ou dans une portion de texte présentant l'organisation de l'ouvrage (voir l'exemple ci-dessous)

Le lecteur peut également survoler un texte écrit en 'sautant' de titres en titres, en n'arrêtant son regard que sur les mots en gras, ou les débuts de paragraphes, etc. Les objets textuels énumérés ici n'ont pas de réelles contreparties à l'oral, l'écrit présentant certaines unités textuelles propres[25].

25 Certains travaux ont justement pour objectif la mise en correspondance entre des énoncés verbaux et les différentes mises en forme d'un document écrit (Maurel *et al.* 2003, Maurel *et al.* 2002). Ces travaux ont une visée principalement applicative. Par exemple, l'Institut de Recherche en Informatique de Toulouse – IRIT et le laboratoire Jacques LORDAT de Toulouse (laboratoire de psycholinguistque) collaborent au développement de techniques d'oralisation des pages internet, pour en faciliter l'accès à des personnes dont la vision est sévèrement dégradée. Ainsi, ils cherchent à établir des « méthodes, des modèles et des outils pour l'étude de communication/interaction dégradée - que celle-ci soit induite par l'usager et/ou l'environnement - en vue de la spécification d'outils de suppléance ou des stratégies palliatives au service des personnes handicapées et âgées. » (page web de Nadine Vigouroux : http://www.irit.fr/recherches/MODEL/DIAM/Membres/page.php3?name=vigourou)

II.2.2. Des unités textuelles spécifiques

Lorsque l'on parle de titre, de sections, de paragraphe, voire de phrase, il est nécessaire de se situer dans le discours écrit. En effet, un aspect de la définition de ces différentes unités textuelles concerne la matérialité du texte. Ce sont des objets textuels. Ainsi, un titre est généralement entouré de blancs verticaux, avec ou non une typographie particulière ; une section est délimitée, par deux titres, un double saut de ligne ou des objets graphiques (ex : °,°,°,°) ; un paragraphe est délimité par des alinéas ; etc. N'importe quel lecteur qui pose un premier regard sur un texte peut y voir ces différents objets. C'est d'ailleurs notre premier contact avec le texte. Mais la définition d'un titre n'est pas juste une question de typo-disposition. Sa mise en forme matérielle est le résultat d'un acte de discours.

Virbel (2002) définit l'objet titre comme la mise en forme matérielle associée à la contrepartie discursive : « dans cet objet titré, l'auteur (traite de + parle de + a pour (sujet + objet) + est relatif à +) *titre* ». Dans Hồ-Đắc *et al.* (2004) nous avons montré que cette définition discursive des titres est à nuancer selon le statut discursif des référents exprimés dans les titres. Cependant, nous ne trouvons pas à l'heure actuelle de définition précise des titres de sections. Si l'on regarde l'extrait (II.3) : seul le soulignement distingue *pôle de base*, *pôle d'équilibre* et *pôle de référence* du reste du texte. Peut-on affirmer qu'il s'agit de titres*, i.e.* des objets textuels que tout lecteur identifiera comme tels ?

(II.3) *Le SROSS de Bretagne préconise d'organiser cette offre de soins des secteurs sanitaires autour de pôles.*
• Pôle de base : c'est un ou plusieurs établissements publics ou privés qui assurent une offre de soins polyvalente de proximité en médecine, en chirurgie, éventuellement en obstétrique, et un service d'accueil des urgences.
• Pôle d'équilibre : en son sein sont réunis les établissements publics et privés qui assurent, en plus de l'activité du pôle de base, des soins de spécialité. L'obstétrique est toujours assurée, ainsi qu'une prise en charge des nouveau-nés sur place pour les pathologies communes.
• Pôles de référence : en plus des pôles précédents, l'offre de soins comporte des spécialités plus nombreuses. Il possède, outre un service de néonatalogie, des équipements lourds multiples. Il comprend obligatoirement un Centre Hospitalier. [ATLAS_1]

Selon le MAT, nous avons ici des objets qui correspondent effectivement à la contrepartie discursive citée précédemment. En poussant la définition de Virbel (2002) à l'extrême, des expressions du type : « Concernant *X* » peuvent constituer des titres. Nous aurions donc le même métadiscours pour les trois images de textes présentées en Figure II.7. Nous avons alors un sorte de continuum allant d'une mise en forme visuelle très marquée à une mise en forme visuelle quasi-nulle. Selon cette idée, les introducteurs de cadres du type « Concernant *X* » peuvent constituer des titres à mise en forme visuelle nulle[26].

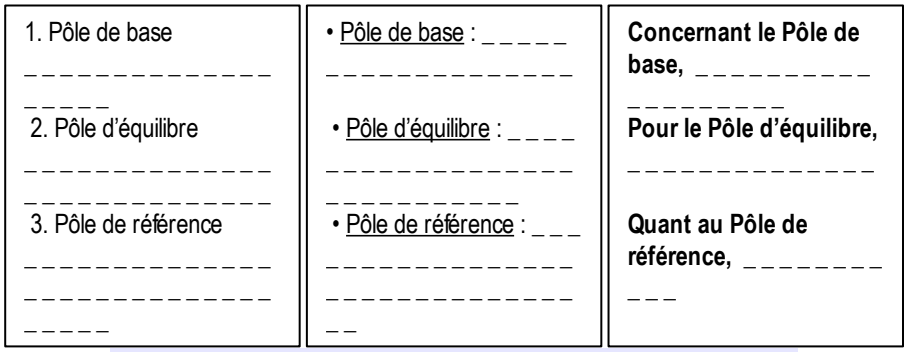


Figure II.7 : Trois mises en formes possibles relatives au métadiscours définissant le titre

26 On peut alors se poser la question de la reconnaissance, lors des processus de lecture, de ces titres sans mise en forme particulière.

Ce rapprochement entre introducteurs de cadre et titres, et donc entre cadres et sections peut également se faire au niveau des paragraphes. Charolles & Péry-Woodley (2005) compare les paragraphes à des cadres sous-spécifiés, *i.e.* qui ne sont pas spécifiés par un critère d'interprétation.

> « les unités typodispositionnelles (paragraphes, tirets, puces, etc.) sont des sortes de cadres sous-spécifiés sémantiquement (le critère de regroupement des propositions n'étant pas signalé sauf quand il y a titraison). » (Charolles & Péry-Woodley 2005:5)

En effet, comme pour les sections et les cadres, les paragraphes réalisent un découpage du texte en segments. Ces segments sont censés être homogènes relativement à un critère d'interprétation. Pour les sections et les cadres, ce critère est spécifié par le titre ou l'introducteur de cadre. Pour le paragraphe, aucune expression marquée ne signale ce critère, ce qui confère à cette unité textuelle une caractérisation très subjective (*cf.* partie V.3.2). Le MAT ne fournit pas de contrepartie discursive satisfaisante pour qualifier les paragraphes, ce qui est normal vu l'absence de correspondance entre cet objet et une intention définie. On pourrait peut-être définir le paragraphe par la contrepartie : « l'auteur marque un changement ou une pause dans le flot du texte » ; car ce qui est certain, c'est que l'organisation en paragraphes correspond à l'acte de segmenter le texte.

Du point de vue de la production comme de l'interprétation, le découpage en paragraphes peut signaler une multitude d'actes de discours : changement de point de vue, de référent, de thème, de style… Il peut marquer une variété d'articulations différentes et donc être l'indice d'une variété d'instructions ; il peut même n'être parfois qu'une trace dénuée d'instruction, ce qu'évoque Heurley (1997) au vu de ses résultats expérimentaux. Nous supposons cependant que tout lecteur qui rencontre un paragraphe le lit comme une rupture dans le flot du texte, un pause durant laquelle il peut 'nettoyer' sa mémoire tampon (cf II.1.2).

Définir les unités textuelles propres à l'écrit est encore plus délicat lorsque l'on cherche à définir le découpage en phrases. En effet, l'unité phrase est aussi une notion essentiellement intuitive :

> « Une phrase est d'abord une séquence de mots que tout sujet parlant non seulement est capable de produire et d'interpréter, mais dont il sent aussi intuitivement l'unité et ses limites. » Riegel *et al.* (1994:103)

En fait, seules les marques visuelles (marques typographiques de ponctuation finale, de changement de paragraphes et majuscules) nous permettent de délimiter ces unités, même s'il est encore délicat de découper de façon automatique un texte en phrases (*cf.* Mourad 2001:222-257). Il n'y a pas réellement de définition autre que typographique de cette unité, ou alors ce sont des définitions sémantiques ou syntaxiques qui rapprochent la phrase de la proposition, et qui restent :

- soit obscures, impliquant l'idée d'une complétude de sens : « La phrase est l'expression plus ou moins complexe, mais offrant un sens complet, d'une pensée, d'un sentiment, d'une volonté » (Mauger 1968, cité in Riegel *et al.* 1994:103) ;
- soit trop rigides pour couvrir toutes les réalités langagières, impliquant l'idée d'une complétude syntaxique : « la phrase minimale est formée de deux constituants, SN appelé sujet et SV prédicat » (Dubois 1969, cité dans Mourad 2001:220).

D'un point de vue syntaxique, tout indique que la phrase, comme le paragraphe, n'est pas une unité syntaxique. Ce qui est certain, c'est que les textes sont constitués physiquement de sections, de titres, de paragraphes, de phrases, etc. Nous voulons justement observer ce qui se passe à l'orée de ces unités textuelles, et notamment en initiale de celles-ci. Le but de ce travail n'est pas de définir ce que sont les phrases, les paragraphes, les sections. Notre intérêt pour ces différents découpages réside justement dans leur délimitation visuelle qui constitue à la fois des traces du processus de production et des indices forts pour le processus d'interprétation.

> « Les propriétés relatives de la réalisation typographique et de l'organisation spatiale de certains objets participent à la composante sémantique du document : l'architecture d'un texte, perceptible par le biais de ces propriétés de mise en forme matérielle, est directement partie prenante de la construction du sens de ce texte » (Pascual 1991:46)

En reformulant les termes de Pascual, nous considérons les différents découpages matériels des textes écrits comme participant à la cohésion d'un texte, jouant au niveau de la segmentation textuelle. En effet, ces découpages ne peuvent être niés. Et même si l'on ne sait pas exactement ce qu'ils signifient du point de vue de la production, il est indéniable qu'ils constituent des indices, au niveau de l'interprétation, impliquant un certain degré de cohésion – d'homogénéité – à l'intérieur des segments ainsi délimités.

Selon Goutsos (1996:503), l'appréhension première d'un texte écrit se fait par son architecture visuelle (sa taille, son titre, ses titres de sections, la justification des lignes, les sauts de paragraphes, etc.) Par ce premier 'contact' le lecteur se construit des prévisions sur le contenu du texte. Ces prévisions confèrent une unité au texte, unité précédente et indépendante de n'importe quel indice de cohésion. Ces suppositions sont du type : c'est un texte organisé et non un amas désordonné de phrases, ou encore la phrase qui arrive est en relation de continuité avec ce qui a déjà été lu[27].

II.3. Cohérence, Cohésion

Lorentz (1999)[28] explique qu'à l'écrit la cohérence ne peut être négociée entre le locuteur et l'allocutaire. Cela signifie que la construction du « text world » ne se fait pas en collaboration. Dès lors, si l'allocutaire n'arrive pas à rattacher une nouvelle information au « text world » déjà construit, il n'a comme ressource de clarification que ce que lui a laissé le locuteur sur le support matériel : le texte. La notion de cohérence englobe alors tout ce qui fait que l'allocutaire peut se construire une représentation mentale qui se tienne, dans laquelle chaque information reçue tient une place justifiée. La SF parle alors de « texture ».

> « l'interprétation des discours est soumise à un **principe général de cohérence** (Charolles 1983, 1994) ou de pertinence (Sperber & Wilson 1986) qui est de nature fondamentalement sémantique et pragmatique. Confronté à une séquence d'énoncés produits à la suite, le destinataire ne peut en effet que chercher à établir des relations entre ces énoncés, vu que, précisément, ils sont énoncés à la suite. L'établissement de ces liens fait appel à des opérations intellectuelles de haut niveau dans laquelle interviennent toutes sortes de compétences linguistiques et non linguistiques. Pour guider l'interlocuteur dans le processus de résolution de problèmes, le locuteur a à sa disposition un vaste ensemble de **marques de cohésion** qui codent des instructions relationnelles plus ou moins spécifiques. » (Charolles 1997:3)

Charolles présente ici la distinction claire et généralement admise, notamment en SF, entre les notions de cohérence et cohésion. Alors que la cohérence concerne le niveau cognitif des processus d'interprétation, la cohésion est propre au texte et se mesure par rapport à l'organisation du texte, à son marquage.

> "Cohesion appears on the textual surface in the form of distinct linguistic elements. Coherence, again, is not an inherent quality of a text but rather a matter of text interpretation, of the text receiver's ability to build a text world and a universe of discourse around the text." (Virtanen 1992:90)

Le principe général de cohérence pousse le lecteur à mettre en relation les différentes unités d'un texte. Cette mise en relation aboutit à la construction d'une représentation mentale et est fortement influencée par les attentes

27 Nous parlons de continuité par défaut (voir III.2.1)

28 Nous rappelons ici la citation présentée en partie II.2.

issues de facteurs tant extra-textuels (domaine, auteur, époque de parution, etc.) que intra-textuels (titre du document, présentation visuelle, et tout ce qui a déjà été lu). C'est par ce principe qu'un texte est interprété comme une unité et non comme une collection aléatoire de phrases, de paragraphes, de sections.

La cohérence, selon Givón (1995:343), résulte de la continuité ou la récurrence de certains éléments dans une portion (ou des portions) de texte.

> "For the text analyst, tracking recurrent elements through the text is facilitated by their predictable association with grammar. […] For the text comprehender, overt grammatical signals – syntactic constructions, morphology, intonation – *cue* the text processor, they *guide* him/her in the construction of a coherent mental representation of the text." (Givón 1995:343)

Mais les indices de cohésion d'un texte ne sont pas aussi faciles à 'traquer' que semble l'affirmer Givón. Les relations sémantiques entre les éléments du « text world » sont complexes et font autant appel à des connaissances langagières qu'à des connaissances du monde (*background knowledge*). Pour les premières, il y a des indices à la surface du texte qui guident l'interprétation ; mais pour les secondes, établir une continuité entre deux unités se fait par des jeux d'association qui relèvent de la culture, de l'expérience des allocutaires. On peut ainsi distinguer deux types de cohérence que Givón nomme : « *knowledge-driven coherence* » *vs.* « *grammar-cued coherence* » (1995:358-362). Nous retrouvons ici la distinction faite entre l'approche de l'EOD par les modèles cognitifs tels que ceux de Kintsch et l'approche fonctionnelle adoptée ici.

II.3.1. Indices de cohésion et de texture

Nous avons d'un côté des principes cognitifs qui orientent une interprétation des textes et, de l'autre, des éléments qui marquent les relations entre unités textuelles, qui guident leur perception. Depuis Halliday & Hasan (1976), on tend à regrouper ces marques sous le nom de **cohésion**. Les relations sémantiques entre les différentes unités donnent au texte une **texture** (Halliday & Hasan 1976:2) en créant des liens de cohésion (*cohesive ties*) réalisés à la surface du texte par des indices de cohésion. Plus les unités sont liées entre elles, plus la texture du texte est dense et inversement. Le texte (II.4) présente une texture particulièrement lâche où la continuité entre les différentes phrases est relativement difficile à (re)créer, ce qui implique une difficulté cognitive à construire une représentation mentale cohérente de ce qu'a voulu communiquer l'enfant, comme le commente Schnedecker (1997) à propos de cet exemple.

(II.4) Sortie à la dune

> *Nous sommes partis de l'école à bicyclette pour aller à la côte découvrir la dune. L'euphorbe est une plante toxique, elle fait gonfler la langue et étouffe. Les plantes les plus répandues sont l'oyat et le chiendent. Les plus vieilles dunes ont plus de cinq mille ans. La criste marine est une plante comestible. Madame l'inspectrice accompagnait M. Berger. Nous avons continué la visite avec eux. Puis nous sommes rentrés à l'école. [exemple issu de Schnedecker 1997]*

> « Il ne suffit pas, on le sait, d'aligner les unes après les autres des phrases pour obtenir un texte. […] Ce texte, rédigé par un élève, présente entre autres malformations, un manque de consistance certain au sens où il paraît décousu. Sa texture est distendue faute de liens suffisants pour serrer, nouer les fils constituant le discours. En effet, les objets dont il est ici question sont nombreux et disparates thématiquement parlant. Ils sont, qui plus est, simplement juxtaposés dans le texte, sans être véritablement ni repris ni développés. » (Schnedecker, 1997:5)

Le principal problème de ce texte réside dans la discontinuité des propos tenus, c'est à dire le manque de liens cohésifs entre les différentes unités pour former un tout. Il y a des liens cohésifs (dont certains sont explicités ci-dessous) mais ils sont insuffisants. D'après la SF, ces liens se construisent par procédés de référence, d'ellipse (et substitution), de conjonction et/ou de cohésion lexicale.

La **continuité référentielle** est indiquée par l'ensemble des expressions anaphoriques (pronoms, descriptions définies ou démonstratives co-référentielles, expressions autres impliquant un lien inférentiel) et peut concerner aussi bien les participants que les circonstances impliqués dans le monde du texte. Par exemple, le pronom « *elle* » permet de lier référentiellement les deux propositions de la phrase : *l'euphorbe est une plante toxique, elle fait gonfler la langue et étouffe*. De même, « *eux* » permet de relier les deux phrases : *Madame l'inspectrice accompagnait M. Berger. Nous avons continué la visite avec eux*.

L'**ellipse** correspond à l'omission d'un élément structurellement nécessaire. Ce phénomène n'est pas présent dans l'exemple II.4 et ne sera pas traité ici, l'ellipse restant difficile à repérer automatiquement[29].

La **conjonction** se réalise typiquement par l'utilisation de connecteurs ou d'adverbiaux à fonction de connexion (« linking adverbials » selon la grammaire anglaise de Biber *et al.* 1999). Ces expressions mettent en relation deux unités. Par exemple, *Puis* relie les deux dernières phrases de (II.4) par une relation temporelle de succession : *Nous avons continué la visite avec eux.* ***Puis*** *nous sommes rentrés à l'école*. Les relations en jeu à ce niveau peuvent être représentées par la RST (Mann & Thompson 1986).

Enfin, la **cohésion lexicale** relève, comme son nom l'indique, du lexique, c'est-à-dire du choix des mots. Deux catégories de réalisation sont alors envisageables : la répétition d'un mot, plus communément appelé redénomination (*cf.* Schnedecker 1997 pour le français) et l'utilisation d'un mot lexicalement voisin (synonyme/antonyme, hyperonyme/hyponyme, co-occurrent, etc.) Ainsi, une redénomination de *la dune* permet de lier *Nous sommes partis de l'école à bicyclettes pour aller à la côte découvrir la dune* au titre *Sortie à la dune* ; une relation d'hyperonymie relie *l'euphorbe est une plante toxique* à *Les plantes les plus répandues sont l'oyat et le chiendent* ; une relation de collocation entre *côte* et *dune* dans la première phrase.

Malgré une construction respectueuse des règles de cohésion citées par Halliday, le texte (II.4) ne semble pas cohérent. Ce sentiment résulte du fait que ce texte ne possède pas vraiment de liens de cohésion jouant à un niveau global. Les phrases deux à deux, entretiennent un lien de cohésion – à un niveau local – mais cette cohésion ne va pas au delà. Il n'y a pas de progression thématique*, i.e.* un sujet commun à propos duquel l'auteur apporte de façon progressive des informations (voir III.3.1). Le seul lien de cohésion de niveau global est donné par le titre du texte : « *sortie à la dune* ». Il concerne les circonstances lors desquelles l'auteur a fait l'expérience des propos tenus dans le texte. Mais ce lien est beaucoup trop ténu pour souder l'ensemble des propositions de cet extrait. Nous reviendrons plus longuement sur cette cohabitation nécessaire entre des continuités au niveau des circonstances et des continuités au niveau des référents. L'étude de cette cohabitation constitue en effet un aspect essentiel dans notre étude.

Les quatre façons de tisser un texte sont, on le voit d'après les illustrations, très basiques et couvrent des ensembles différents et variés d'indices. Elles peuvent d'ailleurs utiliser des connaissances tant langagières (l'utilisation des expressions anaphoriques) qu'encyclopédiques (le recours aux réseaux sémantiques du lexique). Les indices cités pour chacune sont considérés pour leur capacité à marquer une certaine continuité dans le discours, continuité relative à la composante idéationnelle (*cf.* III.1.2.b). Ce phénomène existe également au niveau d'une structuration plus rhétorique (*cf.* III.1.2.a), ce qu'illustre l'exemple II.5[30] où le *Mais* en initiale de paragraphes marque l'introduction d'un nouveau segment rhétorique.

(II.5) *C'est pour cette raison que l'entreprise française se contente actuellement de respecter les décisions de l'ONU. [...] Le gouvernement français n'a pas adopté pour l'instant de mécanisme de contrôle plus strict.*

29 Jacques (2004) propose des pistes pour un repérage automatique des termes complexes dits « réduits », c'est à dire pour lesquels il y a omission de la tête ou de l'expansion d'un syntagme nominal et pour lesquels on peut parler d'ellipse.

30 L'exemple plus complet est donné dans la partie précédente.

> ***Mais*** *les satellites commerciaux non-américains progressent eux aussi vers des résolutions plus fines, aux alentours de deux mètres. Le satellite israélien Eros s'approche de la résolution métrique avec une résolution de 1,8 mètres. Spot 5 [...]. Rocsat pour la Corée et Alos pour le Japon[...]. La compagnie russe Sovinform Spoutnik [...]. L'Inde [...]. L'intérêt de l'imagerie à deux mètres reste plus limitée pour des utilisateurs hostiles ou pour les médias. [GEOPO_2]*

Dans cet exemple, le *Mais* marque non pas une continuité mais une discontinuité. Il constitue ainsi un marqueur de segmentation. Les marqueurs de segmentation servent à avertir l'allocutaire qu'il n'y a pas continuité totale entre deux segments mais qu'il y a un changement soit dans la définition du « text world » (déplacement d'un référent à un autre, d'une temporalité à un autre, d'une localisation spatiale à une autre, d'un point de vue à un autre, d'un ensemble notionnel[31] à un autre, etc.) soit dans la structure rhétorique. Car le discours n'est pas que continuité. Le discours se construit par une séquentialité de périodes de continuité, délimités par des périodes de discontinuité (Goutsos 1996). Le chapitre III est entièrement consacré à notre définition de la segmentation textuelle et de la séquentialité.

Dans l'exemple II.6, les différents adverbiaux temporels (en gras) constituent des marqueurs de segmentation dans la mesure où ils délimitent des segments discontinus sur l'axe temporel, alors que les expressions co-référentielles référant à Béla Bartók (surlignées) signalent une continuité référentielle. En même temps, la progression temporelle mise en place par les adverbiaux renforce la texture de cet extrait (l'effet est d'autant plus fort que les localisations temporelles apparaissaient dans un ordre chronologique).

(II.6) *La première tournée aux États-Unis se déroule de décembre 1927 à février 1928.* ***Pendant deux mois,*** *Bartók parcourt le pays, faisant des conférences sur la musique populaire hongroise et sa place dans la musique savante, illustrée au piano par lui-même. Il donne des récitals de musique de chambre avec Jelly Arányi et József Szigeti, joue avec orchestre sous la direction de Willem Mengelberg, Fritz Reiner et Serge Koussevitsky.* ***En janvier 1929,*** *il séjourne en U.R.S.S. et Ø donne des récitals à Kharkov, Odessa, Leningrad et Moscou.* ***En 1931,*** *il termine son Concerto pour piano et orchestre no 2, Ø compose les Quarante-Quatre Duos pour violons sur des mélodies populaires et orchestre quelques-unes de ses pièces pour piano. Cette année est particulièrement bien remplie : Bartók a cinquante ans et Ø reçoit la Légion d'honneur des mains de Louis de Vienne, ministre plénipotentiaire, chef de la légation de France à Budapest. Au début de l'été, il est à Genève, à la session de la Commission de la coopération intellectuelle de la Société des Nations, avec Carel Capek, Gilbert Murray, Thomas Mann, Paul Valéry, etc.* ***L'année suivante,*** *il compose Six Chants sicules pour chœur d'hommes, puis il termine,* ***en 1934****, son Quatuor à cordes no 5.* ***À partir de cette année,*** *il est affecté à l'Académie hongroise des sciences pour la préparation de l'édition systématique de la musique populaire hongroise. [PEOPL_15]*

Nous avons là une portion de texte où cohabitent des indices de discontinuité au niveau de la référence temporelle et des indices de continuité au niveau de la référence aux participants. Dans une vision plus globale, les adverbiaux temporels participent également à la cohésion du discours, par la relation de cohésion lexicale temporelle qui renforce la texture du passage. Les organisations évoquées jouent à un niveau uniquement idéationnel. Mais un phénomène de continuation à un niveau textuel est également présent dans cet exemple.

La succession d'adverbiaux temporels antéposés organise ce texte selon un point de vue temporel. D'un point de vue textuel, on peut dire que les informations du paragraphe sont organisées selon une chaîne de référence au cours de laquelle une succession de localisations temporelles sont délimitées. Cela peut être schématisé par la figure II.8 où chaque nouvel item correspond à une discontinuité au niveau de la référence temporelle :

- *La première tournée aux États-Unis se déroule de décembre 1927 à février 1928. … Bartók … . Il … .*
- ***En*** *janvier 1929, il … .*
- ***En 1931,*** *il … . Cette année … : Bartók … . Au début de l'été, il … .*
- ***L'année suivante,*** *il … il … ,*
- ***en 1934,*** *… .*
- ***À partir de cette année,*** *il … .*

Figure II.8 : Continuité référentielle et discontinuité temporelle

domaines thématiques, etc., voir partie V.4.1.

Cet exemple nous présente donc deux modes organisationnels, deux stratégies textuelles de présentation des informations que nous appelons des **continuités texto-stratégiques – TSC**[32] selon la terminologie d'Enkvist (1981, 1989) et plus récemment de Virtanen (1992, 2004). Les phrases de la portion indexée par les adverbiaux temporels ont toutes la caractéristique commune de suivre une même TSC temporelle. Cette conception de l'organisation du discours par continuités texto-stratégiques permet de dépasser l'organisation locale d'un cadre de discours pour considérer l'encadrement du discours comme une sorte de 'mise en item' d'un segment de texte dans une structure énumérative plus globale. La partie III.1.2.c définit plus longuement cette idée de continuation texuelle.

II.3.2. Des procédés de connexion et des procédés d'indexation

Selon Charolles, les liens qui s'établissent entre deux unités peuvent être de deux ordres selon qu'ils se définissent vers l'avant ou vers l'arrière. Charolles fait ainsi la distinction entre des **relations de connexion** et des **relations d'indexation** (Charolles 1997). Les relations de connexion fonctionnent vers l'arrière en liant les propos à venir aux propos précédents ou le référent exprimé aux précédents référents. Ce type de relation est particulièrement observée au niveau de la construction des chaînes de référence.

À l'inverse, les relations d'indexation fonctionnent vers l'avant, correspondant à la mise en relation entre un élément exprimant un critère d'interprétation – un index – et une portion de texte. L'élément index est 'projeté' sur l'ensemble de la portion sur laquelle il étend sa référence – nous parlons alors de **portée sémantique** – et/ou pour laquelle il oriente l'interprétation – nous parlons alors de **portée cadrative**.

La distinction entre portée sémantique et portée cadrative rejoint celle définie entre composante idéationnelle et composante textuelle. A la surface du texte, portée sémantique et portée cadrative peuvent coïncider, tout comme la localisation du Thème et de l'entité sujet. En conséquence, il est relativement délicat de ne pas confondre l'identification des deux effets de portée. Cependant, ces deux phénomènes discursifs ne suivent pas la même définition.

Pour Charolles & Vigier 2005, la portée sémantique correspond à « l'influence à distance (*i.e.* au delà de sa phrase d'accueil) du trait sémantique spécifié par l'adverbial introducteur de cadre ». Cette capacité est liée à leur « usage fonctionnel (*i.e.* leur exploitation pour la répartition des informations textuelles au fur et à mesure du discours) ». Selon ces auteurs, la portée sémantique est une conséquence de la portée cadrative des adverbiaux antéposés.

Ainsi, dans l'exemple II.7, c'est son rôle dans la répartition des informations qui confère à l'adverbial *En 1931* une portée sémantique allant jusqu'à l'adverbial temporel suivant. Par contre, dans l'exemple suivant (repris de la partie précédente), le *Mais* à l'initiale de paragraphes étend une portée uniquement cadrative sur le reste du paragraphe. L'articulation qu'il explicite est purement argumentative et ne porte pas sur la composante idéationnelle.

(II.7) *C'est pour cette raison que l'entreprise française se contente actuellement de respecter les décisions de l'ONU. [...] Le gouvernement français n'a pas adopté pour l'instant de mécanisme de contrôle plus strict.*
***Mais** les satellites commerciaux non-américains progressent eux aussi vers des résolutions plus fines, aux alentours de deux mètres. Le satellite israélien Eros s'approche de la résolution métrique avec une résolution de 1,8 mètres. Spot 5 [...]. Rocsat pour la Corée et Alos pour le Japon[...]. La compagnie russe Sovinform Spoutnik [...]. L'Inde [...]. L'intérêt de l'imagerie à deux mètres reste plus limitée pour des utilisateurs hostiles ou pour les médias. [GEOPO_2]*

32 Notre traduction des *Text-Strategy continuities*. Pour la version française (continuités texto-stratégiques), nous gardons la siglaison anglaise : TSC.

Les titres de sections donnent une illustration parfaite de la relation d'indexation. En effet, ils permettent d'orienter l'interprétation de tout ce qui va suivre dans la section : soit en précisant ce que constitue cette section dans le déroulement du texte (on parle alors de titres fonctionnels ou formels tels que *introduction*, *partie théorique*, etc.) ; soit en exprimant des circonstances, des référents et/ou des procès qui entrent en jeu dans le sous-monde que toute la section relate (on parle alors de titres référentiels et de titres thématiques, selon leur implication dans la construction de la représentation mentale cf. Hồ-Đắc *et al.* 2004). Dans le cas des titres fonctionnels et sans doute dans le cas des titres thématiques, la portée est purement cadrative ; alors que dans le cas de titres référentiels, la portée est sémantique*, i.e.* les entités, procès et circonstances qui y sont exprimées sont actives tout au long de la section.

Dans le cas des titres, il est évident que leur fonction se définit naturellement au niveau de l'organisation globale du texte. En effet, on ne lit pas un titre comme un élément quelconque au sein d'une succession de phrases, mais comme un point particulier situé un peu en dehors de la linéarité du texte (*cf.* Hồ-Đắc *et al.* 2004).

La relation d'indexation peut également être caractéristique de la fonction organisationnelle de certains adverbiaux circonstanciels. En effet, c'est ce type de relation qui explique l'hypothèse de l'encadrement du discours (Charolles 1997, Charolles & Vigier 2005, Le Draoulec & Péry-Woodley 2003, 2005). Selon cette hypothèse, un adverbial (un « introducteur de cadre ») initie un cadre à l'intérieur duquel le critère qu'il exprime (spatial, temporel, notionnel, rhétorique, modal, textuel...) reste valide.

Dans ce cas, l'idée d'une fonction au niveau de l'organisation globale du texte est beaucoup moins évidente que pour les titres du fait que les adverbiaux sont intégrés au flot du texte sans s'en démarquer. Cependant, les introducteurs de cadres peuvent suivre directement un titre, voire se situer dans un titre (Laignelet 2003 propose une étude en corpus des influences entre titres et introducteurs de cadre).

Dans l'exemple (II.6), les trois localisations temporelles *En janvier 1929*, *En 1931* et *en 1934*, ont une fonction de connexion au niveau local, dans la succession des phrases. Ainsi, si l'on extrait de l'exemple (II.6) les deux phrases reprises ci-dessous, *En 1931* permet de lier dans une relation temporelle de succession le séjour en URSS à l'achèvement du *Concerto pour piano et orchestre n°2* :

> *En janvier 1929, il séjourne en U.R.S.S. et donne des récitals à Kharkov, Odessa, Leningrad et Moscou. En 1931, il termine son Concerto pour piano et orchestre n°2.*

À cette fonction locale s'ajoute une fonction d'indexation telle que le montrait la structure énumérative établie par la figure II.8 (p.49) et reprise ici :

- *La première tournée aux États-Unis se déroule de décembre 1927 à février 1928. … Bartók … . Il … .*
- ***En janvier 1929,** il … .*
- ***En 1931,** il … . Cette année … : Bartók … . Au début de l'été, il … .*
- ***L'année suivante,** il … il … ,*
- ***en 1934,** … .*
- ***À partir de cette année,** il … .*

L'interprétation des procès dans lesquels Bartók a été impliqué se construit en fonction de la localisation temporelle déterminée par l'adverbial correspondant. Chaque item peut être envisagé comme un sous-monde temporellement spécifique et constitutif du monde général dont la temporalité égale la période de vie de Bartók. Chaque sous-monde correspond à un cadre de discours tel que le définit Charolles (1997).

Nous avons remarqué précédemment que ces cadres sont en relation de cohésion lexicale et de ce fait participent à une même façon d'organiser le discours (une même TSC, *cf.* Virtanen 1992 et III.3.4). Cela nous permet

de passer à un niveau d'organisation encore plus global. L'exemple (II.8) nous montre une autre illustration de ce type de segmentation.

(II.8) *L'année 2002 a été marquée par une* ***série de scandales financiers*** *qui ont ébranlé la confiance de l'investisseur américain - autant dire du citoyen - dans l'intégrité et la transparence des marchés financiers.* ***L'enchaînement des faits****, tout d'abord. La faillite d'Enron était déclarée en décembre 2001, celle de Global Crossing en janvier 2002. Pour son rôle dans l'affaire Enron, le cabinet d'audit Arthur Andersen était mis en examen en mars.* ***En juin,*** *Enron reconnaissait avoir versé un total de 310 millions de dollars en espèces à ses dirigeants au cours de l'année 2001 et Worldcom corrigeait ses comptes de 3,8 milliards de dollars.* ***Le 21 juillet,*** *la faillite de Worldcom était déclarée.* ***Le 24,*** *la Securities and Exchange Commission (SEC) portait plainte contre les dirigeants d'Adelphia, accusés d'avoir dissimulé 2,3 milliards de dollars de dettes dans des sociétés non consolidées.* ***En août,*** *l'ancien Chief Executive Officer (CEO) de imclone était mis en examen pour délit d'initié.* ***En septembre,*** *c'était au tour du CEO et du Chief Financial Officer (CFO) de Tyco d'être mis en examen pour corruption : il leur était reproché d'avoir détourné 600 millions de dollars, dont 170 millions de prêts personnels accordés par la société.* ***Enfin le 5 novembre 2002,*** *Harvey L. Pitt, président de la SEC et champion du laisser-faire réglementaire, était contraint de démissionner.*
Tous ces scandales *se sont produits dans un contexte économique morose, très différent de l'euphorie des années 1990 : la " bulle Internet " a […][GEOPO_31]*

Dans cet exemple, toute la portion de texte surligné correspond à un segment organisé selon une TSC temporelle et de ce fait distinct du reste du texte. Cette TSC temporelle n'est pas uniquement construite grâce aux adverbiaux temporels en position initiale. Différents indices empaquettent cette TSC. Tout d'abord, le paragraphe est introduit par un référent temporel : « *l'année 2002* » qui englobe de sa portée sémantique toutes les propositions du paragraphe. Ainsi, chaque adverbial est une localisation temporelle à) l'intérieur de l'année 2002. Autre indice préparatoire à la présence d'une TSC, l'effet d'amorce (« *une série de scandales financiers* », « *l'enchaînement des faits* ») et de conclusion (« *Tous ces scandales* »), que seules nos connaissances lexicales nous permettent de percevoir, nous y revenons en III.1.2.b. Nous voyons bien dans cet exemple comment, de la portée locale d'adverbiaux temporels, nous arrivons à la création d'une structure organisée de plus en plus importante : d'abord par une série d'adverbiaux temporels (la TSC), puis par un effet d'amorce et de conclusion qui délimitent une sorte de structure énumérative, jusqu'à atteindre un niveau beaucoup plus 'global' marqué ici par la portée de la localisation temporelle (exprimée en fonction sujet grammatical).

La progression du niveau local des phrases à celui de segments plus grands est présentée plus largement au chapitre III . Les marqueurs discursifs retenus dans cette étude sont pour l'essentiel des marqueurs de segmentation qui permettent cette progression, notamment par leur capacité à délimiter un cadre de discours et à marquer une TSC.

II.3.3. Niveau local, niveau global

En distinguant un niveau local d'un niveau global, nous soutenons l'hypothèse qu'il y a dans le texte des indices signalant les relations entre les phrases successives (au niveau local) et des indices signalant des relations entre des parties du texte plus grandes (au niveau global). Ces dernières s'établissent en dehors de la linéarité du texte.

La frontière entre niveau local et niveau global n'est pas précise. Il s'agit plutôt d'un passage progressif : des liens au niveau local délimitent des segments qui eux-mêmes entretiennent des liens qui de fil en aiguille nous permettent d'arriver au segment le plus grand : le texte en entier.

Certaines stratégies textuelles permettent, par un phénomène de récursivité, de grimper du niveau local au niveau global ou de descendre du niveau global vers le niveau local, suivant que l'on se situe dans une analyse top-

down ou bottom-up. Ce principe de récursivité est propre au langage humain et se retrouve au niveau morphologique, syntaxique et discursif[33].

En dégageant les relations rhétoriques entre un noyau et ses satellites, selon une analyse RST, on peut arriver à schématiser le texte sous forme de segments enchâssés jusqu'à arriver à une relation propositionnelle entre deux segments représentant le texte dans sa globalité.

L'architecture textuelle (*i.e.* la mise en forme matérielle – MFM – du document) montre le même phénomène de récursivité : l'auteur compose un texte T. Il titre/nomme T titre. Il segmente T en x parties P_1, P_2,…, P_n. Il titre P_1 titre de niveau 1, P_2 titre de niveau 1, etc. Il segmente P_1 en x sous-parties sp_1, sp_2,…, sp_n. Etc… jusqu'à arriver aux paragraphes, le plus petit segment textuel délimité par sa MFM.

Le modèle de Grosz & Sidner (1986) propose plusieurs niveaux de segmentation, supposant alors une hiérarchie entre différents types de segments. Dans ce modèle, trois niveaux graduels permettant d'aller de l'énoncé au texte entier sont dégagés.

> "Our main thesis is that the structure of any discourse is a composite of three distinct but interacting components :
> - The structure of the actual sequence of utterances in the discourse ;
> - A structure of intentions ;
> - An attentional state." (Grosz & Sidner 1986:176)

Les différents niveaux de cette application sont articulés de telle façon que les énoncés dans un segment remplissent une fonction par rapport à ce segment, qui lui-même remplit une fonction par rapport au discours entier.

> "the utterances in a segment, like the words in a phrase, serve particular role with respect to that segment. In addition, the discourse segments, like the phrases, fulfil certain functions with respect to the overall discourse." (Grosz & Sidner 1986:176).

Les chaînes de référence, les cadres de discours et les paragraphes ne délimitent que des structures locales dans l'organisation du texte : leur existence est réduite au niveau des phrases et du paragraphe voire des paragraphes (exemples attestés de chaînes et de cadre dépassant le paragraphe ou étant à cheval sur deux paragraphes). Cependant, il existe des hyperthèmes et des macrothèmes, des adverbiaux spatiaux et temporels qui posent une unité de situation sur des sections entières (que dire des adverbiaux détachés en initiale de titre), on pourrait parler d'hyper-circonstant et macro-circonstant. Pour accéder à ce niveau global, l'idée ici défendue est de suivre les même processus de construction du texte que ceux utilisés au niveau local des chaînes et cadres : les processus de continuité et de rupture, mais appliqués à des unités, à des segments de plus en plus grands. Il s'agit de voir s'il existe des indices à la surface du texte permettant au lecteur de procéder au regroupement de plusieurs chaînes de référence, de plusieurs cadres ; tout comme il existe des titres qui permettent le regroupement de plusieurs paragraphes, ainsi que des titres de niveaux n qui permettent le regroupement de plusieurs sections de niveau *n-1*, et ainsi de suite jusqu'à arriver au regroupement final : le texte entier.

33 Ce principe de récursivité semble constituer une caractéristique fondamentale du langage humain (cf. Bouchard 2007).

Chapitre III
Segmentation textuelle et séquentialité

Sommaire

Le phénomène de segmentation peut correspondre à plusieurs définitions selon le niveau d'analyse et le point de vue adopté. Le fait de mettre en titre de ce chapitre *segmentation textuelle* et non *segmentation* tout court signale notre intention de concentrer notre champ d'étude sur le phénomène de segmentation tel qu'il existe dans les textes, c'est-à-dire indiqué à sa surface. Tout phénomène de segmentation implique nécessairement deux processus : un regroupement de plusieurs éléments autour d'un critère spécifique et en même temps un découpage du texte en portions plus ou moins grandes. Cette définition rejoint celle utilisée par Charolles pour décrire une part du rôle discursif des adverbiaux extra-prédicatifs par rapport à l'hypothèse de l'encadrement du discours :

> « Les expressions introductrices d'univers de discours […] servent à **répartir les contenus propositionnels dans des blocs homogènes relativement à un critère spécifié** par le contenu de l'introducteur » (Charolles 1997:24)

Dans notre conception, la segmentation textuelle est associée au phénomène de séquentialité (Goutsos 1996). Le texte est considéré comme le résultat de deux stratégies principales : des stratégies de continuité et des stratégies

de discontinuité. Les premières permettent de regrouper des éléments et les secondes permettent d'établir une transition entre des ensembles d'éléments. La mise en séquence de ces « zones de continuation » et « zones de transition » constitue la base de la construction textuelle (III.2).

Cette segmentation et cette séquentialité sont le fruit d'intentions auteuriales (d'où la notion de stratégies). Nous nous intéressons à la mise en texte de ces stratégies textuelles, *i.e.* aux techniques utilisées pour appliquer ces stratégies. Ces techniques laissent des traces (intentionnelles ou non) à la surface des textes, traces qui constituent ensuite des indices pour le lecteur dans la construction de son *text-world* (*cf.* II.1.1).

Nous nous distinguons des approches TAL de segmentation thématique tels que celles établies par Hearst (1994, 1997), Masson (1995) ou Ferret et Grau (*cf.* Ferret & Grau 1998, 2000, Ferret *et al.* 1998, Ferret 2002) qui envisagent la segmentation du seul point de vue de la cohésion lexicale. Cette conception de la segmentation ne s'appuie pas sur les autres techniques d'organisation du discours. Comme le souligne les études en systémique fonctionnelle (Halliday & Hasan 1976, Halliday 1985, Martin 2001), la cohésion lexicale n'est qu'*une* technique parmi d'autres (voir II.3.1). De plus, il s'agit d'une technique de continuation ; or il existe également des techniques de discontinuité (nous distinguons les techniques de déplacement et de rupture). Dans ce dernier cas, les indices laissés signalent les transitions entre segments et non la cohésion interne du segment.

III.1. La segmentation textuelle : regrouper * découper

III.1.1. Un signalement de la segmentation conceptuelle

La segmentation textuelle se distingue de la segmentation conceptuelle par le fait qu'elle existe par le texte grâce à la présence d'indices de séquentialité ; tandis que la segmentation conceptuelle existe en dehors du texte, au niveau des représentations mentales et processus mentaux qu'il implique.

En effet, la **segmentation conceptuelle** consiste par exemple à subdiviser une représentation mentale en sous-représentations, une entité en différentes parties à décrire, une procédure en différentes étapes à expliquer, une histoire en différents épisodes à narrer, un exposé en différents arguments. Ces segments correspondent à des appellations et définitions plurielles (unité de discours, segment discursif, etc.) Nous retiendrons la notion de **blocs d'information** que Heurley définit comme des « unités de traitement dans le cadre du processus de composition » (Heurley 1994:235).

Mais ces segments conceptuels sont plus ou moins signalés à la surface du texte par la mise en forme matérielle, l'utilisation d'expressions particulières, ou encore par d'autres caractéristiques linguistiques telles que le choix de positionner en initiale tel élément plutôt qu'un autre. C'est cette segmentation marquée à la surface du texte que nous qualifions de **segmentation textuelle**. Typiquement, le découpage du texte en sections est de l'ordre de la segmentation textuelle.

Que ce soit d'un point de vue conceptuel ou textuel, un segment se définit par deux processus combinés : un regroupement et un découpage que Heurley associe aux propriétés fonctionnelles et structurelles des blocs d'information :

> "Each text was composed of several subparts, each being organized around a single topic, called *information blocks*. Indeed, an informationally based segmentation procedure revealed that, overall,

> texts were composed of information blocks (i.e. frame, goal, instruction and results) that were characterized both by functional and structural properties." (Heurley 1997:190)

La preuve principale de l'existence de ces blocs informationnels est la présence, lors du processus d'écriture et de lecture, de longues pauses de part et d'autre des différents blocs. Ces pauses se retrouvent durant le processus de lecture. Cependant, la localisation de ces pauses peut différer, que ce soit entre locuteur et lecteur ou entre lecteurs différents. En fait, le problème de la mise en correspondance entre segmentation conceptuelle et segmentation textuelle consiste à repérer dans les textes les différentes traces et signaux de ces pauses. Cette correspondance peut par exemple se réaliser par le découpage en paragraphes, ce qu'étudie particulièrement Heurley (1994, 1997) ; mais ce n'est pas systématique.

> "on-line and off-line analysis showed that block boundaries did correspond more to breaks in the writing process than to physical break on the page of paper. Thus, only 41,5% of information blocks were marked by at least one paragraph marker." (Heurley 1997:190)

Les études psycho-linguistiques portant sur les raisons qui poussent un locuteur à changer de paragraphe montrent clairement qu'il n'y a pas de règles générales de ce découpage. De plus, la délimitation des blocs informationnels peut tout à fait correspondre à des indices textuels autres que le changement de paragraphes, ce que nous verrons dans la suite de cette thèse. La non-correspondance entre segmentation conceptuelle et segmentation textuelle n'est aujourd'hui ni prouvée ni réfutée. Notre hypothèse est que la segmentation conceptuelle se retrouve au niveau textuel.

La correspondance est certainement partielle, le degré de correspondance pouvant fortement varier selon les type de texte et genre discursif considérés. Il est en effet très probable que les auteurs de textes expositifs cherchent davantage à expliciter leur organisation que les auteurs de textes narratifs (qui peuvent d'ailleurs jouer à 'tromper' le lecteur).

De plus, le marquage de la segmentation textuelle n'est pas binaire : il n'y a pas de corrélation absolue entre une forme et une fonction. C'est l'influence conjointe de différents indices qui nous fait interpréter la délimitation d'un nouveau segment, et la constitution de ces configurations d'indices varie elle aussi selon les types de texte et genres discursifs considérés, voire selon le texte et les stratégies textuelles adoptées par l'auteur, à un niveau individuel[34].

Concernant les propriétés fonctionnelles relatives au critère de regroupement qui soude les constituants du segment entre eux, Heurley en mentionne deux selon que l'on se place d'un point de vue sémantique (un topique commun) ou d'un point de vue plus rhétorique (une même étape dans le déroulement de la procédure : présentation du cadre théorique, du but de la procédure, des instructions 1, 2, 3, etc.) Un parallélisme peut être établi avec les structures intentionnelle et attentionnelle de Grosz & Sidner (1986) : aux segments constitués d'éléments regroupés autour d'un topique commun se superposent des segments régis par des intentions discursives particulières, chacune contribuant à son niveau à la visée discursive générale du texte. La partie suivante présente plus en détail ces différents critères de regroupement.

34 Nous revenons sur la définition des indices de segmentation au chapitre V. C'est précisément cette problématique qui justifie notre méthodologie exploratoire en corpus (chapitre VII).

III.1.2. Différents modes de segmentation

Nous avons défini le phénomène de segmentation textuelle comme procédant à la fois à un **regroupement de plusieurs unités linguistiques** autour d'un point commun et au **découpage du texte**. Nous avons donc deux types de procédés mis en cause :

- Des **techniques de continuation** qui consistent à relier les unités entre elles autour d'un critère d'interprétation commun.
- Des **techniques de transition,** qui consistent à délimiter au sein des textes des segments, *i.e.* des ensembles d'unités successives (mots, propositions, phrases, paragraphes, sections, cadres de discours etc.)

La nature du critère d'interprétation utilisé pour lier les différentes unités entre elles définit trois types de continuations :

- des **continuations sur le plan rhétorique** : les unités ont en commun leur contribution à une même intention auteuriale, à un même « but local de discours » (voir infra).
- des **continuations sur le plan idéationnel** : les unités sont liées par leur relation à un même objet de discours, *i.e.* un même ensemble de référent(s) ou un même ensemble de circonstances. Le segment composé est alors indexé comme étant à propos du référent ou comme s'interprétant dans la ou les circonstance(s) exprimée. Dans les deux cas, on peut considérer que la thématique (au sens large) du segment concerne tant la circonstance exprimée que le référent.
- des **continuations sur le plan textuel** : les unités sont liées sur le plan textuel par un même mode de structuration. En suivant les idées de Virtanen et sa définition de la rupture (Virtanen 1992), un passage structuré par une TSC temporelle se distingue d'un passage structuré autour d'une TSC topicale.

Bien entendu, la composante textuelle 'sert' les continuations sur le plan rhétorique et idéationnel. Ainsi, une certaine structuration sert une certaine intention ou la 'mise au monde' de certains objets du discours.

Ces trois niveaux de structuration se retrouvent dans le **modèle de Grosz & Sidner (1986)**. Ce modèle considère que le texte est segmenté simultanément selon la structure intentionnelle et la structure attentionnelle, toutes deux existant au travers d'indices textuels. La structuration intentionnelle rassemble des énoncés par le fait qu'ils servent un même 'but discursif' dans l'édification du *text-world*. La structuration attentionnelle réunit des énoncés selon les entités, circonstances et procès qui sont exprimés et mis en focus. L'existence de ces regroupements – de ces segments – se réalise par un marquage discursif réalisé par ce que Grosz et Sidner appellent des *cue-phrases*[35].

> « Le segment de discours est déterminé par trois facteurs : le fait qu'il permet au locuteur d'accomplir **un but local de discours homogène**, contribuant, d'une façon ou d'une autre, à un but de discours plus global ; le fait qu'un « **espace de focus** » lui est associé, domaine qui réunit les entités, propriétés et relations sur lesquelles le locuteur veut focaliser l'attention de son allocutaire à ce moment là du discours ; et **le fait qu'il délimite une unité minimale de discours** signalée par des marqueurs linguistiques (expressions « signaux » - *cue phrases* -, choix des temps, des modes, de l'aspect, ainsi que certains types de marqueurs référentiels plutôt que d'autres). » Cornish (2000:9)

III.1.2.a) Continuations intentionnelles : autour d'une même visée rhétorique

La valeur intentionnelle d'un segment correspond à sa visée discursive locale, par exemple : « camper la scène dans le récit d'une histoire, raconter un épisode dans un récit, ou établir et justifier une prémisse au sein d'une argumentation » (Cornish 2000:9). La structure intentionnelle peut être représentée par le modèle de la **Rhetorical Structure Theory – RST** (Mann & Thompson 1986) qui modélise les relations rhétoriques entre les énoncés ou groupes d'énoncés grâce à des propositions relationnelles.

35 Notre conception des «indices de séquentialité » englobe la notion de *cue-phrases*.

La RST cherche à schématiser la structure rhétorique des textes en représentant les relations sémantico-pragmatiques qui unissent les différents énoncés ou groupes d'énoncés entre eux. Cette théorie rend compte de la cohérence textuelle en faisant ressortir les propositions implicites (propositions relationnelles) qui relient entre elles les propositions explicites. Aux segments dégagés[36] est associé le statut de noyau ou de satellite. Ainsi, les segments sont liés soit par une relation de subordination (un satellite est subordonné à un noyau) soit par une relation de coordination (une liaison entre deux noyaux). La relation de subordination correspond à la relation de « dominance » identifiée par Grosz & Sidner (1986) : un segment A est dominé par un segment B si A a un but discursif local qui contribue au but discursif de B. La relation de coordination se retrouve plus ou moins au niveau dans la relation de « précédence » (*satisfaction-precedence*) chez Grosz & Sidner : un segment A précède un segment B si le but discursif de A doit être satisfait avant celui de B.

Dans l'exemple III.1, le premier paragraphe s'oppose au second par une relation de contraste. Nous pouvons dire que, du point de vue de l'auteur, la compréhension du premier paragraphe est nécessaire à celle du second (relation de précédence). Les énoncés constitutifs de ces deux paragraphes peuvent également être regroupés en segments. Dans le premier paragraphe, les énoncés [1] et [2] préparent l'énoncé [3]. Dans le second paragraphe, les énoncés [5-7] constituent un segment servant à développer les propos tenus dans l'énoncé [4] (relation d'élaboration où [4] *domine* [5-7]). L'énoncé [8] propose une synthèse du segment [4-7], qui s'oppose finalement à l'énoncé [9]. Enfin le tout peut être mis en relation avec le titre [0] par une relation de préparation ou d'élaboration[37].

(III.1) **[0]*8.3. Baccalauréat, scolarité, société*** [titre niveau 2]
[1]*En 1950, 5% seulement des jeunes de chaque génération obtenaient le baccalauréat.* **[2]***La proportion a atteint 11% en 1960, 20% en 1970, 25% en 1980, 36% en 1988, 47,5% en 1991 ; elle a dépassé 50% en 1993.* **[3]***À ce rythme-là, les deux tiers des jeunes de chaque génération seront bacheliers en l'an 2000.*
[4]*Malgré cette progression rapide et générale, les écarts restent sensibles entre les académies et les départements.* **[5]***En 1990, la proportion de bacheliers par classe d'âge n'atteint pas 40% dans les académies de la grande couronne parisienne.* **[6]***Elle se situe entre 40 et 45% dans le Nord, en Lorraine et en Alsace, dans les académies de Toulouse, Limoges, Grenoble, Rennes ainsi qu'en Corse.* **[7]***En Île-de-France les écarts sont écrasants entre Paris (67% [...]) et la Seine-Saint-Denis (29%)...* **[8]***Dans l'ensemble, la moitié sud du pays continue à avoir de plus fortes proportions de bacheliers que la moitié nord, Bretagne et, à présent, Lorraine exceptées.* **[9]***Mais ces différences s'atténuent.* *[ATLAS_2]*

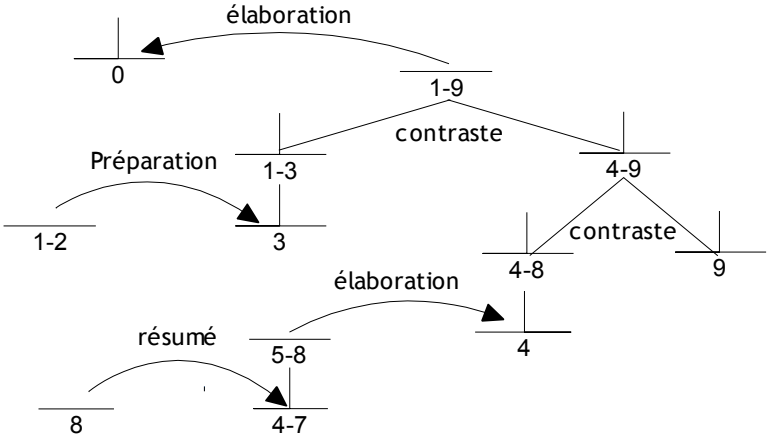


Figure III.1 : Exemple de représentation RST

36 La taille des segments identifiés va d'une proposition au texte entier.

37 La relation entre le titre et sa section peut être représentée par une relation de préparation. Alors le titre est un satellite préparant l'interprétation de la section, noyau. Si l'on représente la relation titre/section par une relation d'élaboration, les rôles sont inversés : la section est satellite du noyau-titre, ce qui signifie que la section élabore, développe ce qui est exprimé dans le titre. Ces deux relations sont tout à fait valables et relèvent davantage d'une différence de point de vue que d'une vérité *vs.* fausseté.

La représentation des propositions relationnelles identifiées dans l'exemple III.1 montre clairement le caractère implicite du marquage des relations dégagées. Les propositions relationnelles font plus souvent l'objet d'inférences que de marquage spécifique[38]. Cette forte propension des relations rhétoriques à être inférées rend leur annotation très délicate et nécessairement manuelle. Actuellement, certains travaux cherchent en corpus des indices textuels permettant l'établissement d'une caractérisation automatique des différentes propositions relationnelles[39]. La plupart de ces travaux se basent sur des listes de correspondances *a priori* entre une forme et une relation rhétorique pour repérer et caractériser une proposition relationnelle, et ainsi développer et vérifier les listes de départ. Nous reviendrons sur ces études dans la partie VI.3 consacrée aux corpus annotés discursivement.

Un tout autre type de segmentation rhétorique a été développé par Teufel (Teufel 1998, 1999 et Teufel & Moens 2002) : le zonage argumentatif (« argumentative *zoning* »). Ce modèle se restreint à l'étude d'un certain type de texte : les articles scientifiques, et a une visée applicative précise : le résumé automatique ou l'extraction d'information. Il ne cherche pas à modéliser la pensée humaine (la RST appartient à l'approche logique du discours, voir I.2.1) mais à proposer une représentation de la structure rhétorique générale des textes pour orienter un traitement automatique tel que l'extraction d'information ou le résumé automatique. Sept statuts rhétoriques sont définis afin de catégoriser toutes les phrases du texte :

statut rhétorique	***intention définitoire***
AIM	présentation de l'objet de l'article
TEXTUAL	explicitation de la structure de l'article ou de la section
OWN	description (neutre) de son propre travail relatif à l'objet d'étude : méthodologie, résultats, discussion
BACKGROUND	état de l'art des modèles et courants auxquels appartient l'étude
CONTRAST	Mise en valeur (comparaison ou contraste) de l'étude par rapport à d'autres travaux
BASIS	présentation des travaux sur lesquels se base cette étude, ou dans lesquels s'inscrit cette étude
OTHER	présentation (neutre) d'autres travaux

Tableau III.1 : Statuts Rhétoriques dans les articles scientifiques selon Teufel & Moens (2002)

Le texte peut ainsi être découpé en zones argumentatives, *i.e.* en ensembles de phrases adjacentes présentant le même statut rhétorique. À la différence de la RST, les travaux de Teufel ont pour but premier la détection d'indices textuels permettant le repérage automatique des différentes zones. La représentation de la sémantique des formes n'est absolument pas recherchée. Seule compte la visée applicative : pouvoir repérer les différentes zones argumentatives pour générer automatiquement un résumé des buts d'un article, des critiques qu'il porte, etc. mais également renvoyer la zone adéquate à une requête (nous revenons sur leur méthode de segmentation en VI.3).

III.1.2.b) Continuations idéationnelles : autour d'un même objet de discours

La valeur attentionnelle d'un segment porte sur les objets de discours sur lesquels le locuteur souhaite porter l'attention du lecteur. Une continuation idéationnelle est créée lorsque plusieurs unités successives concernent les mêmes entités, les mêmes circonstances et/ou les mêmes procès. Nous retrouvons approximativement l'idée des chaînes de référence.

> "Les chaînes sont constituées par des suites d'expressions coréférentielles […]. Seules peuvent appartenir (donner lieu à) une chaîne des expressions employées référentiellement, c'est à dire toutes et rien que les expression nominales (ou pronominales) permettant d'identifier un individu (un objet de

38 La notion de marquage implicite est abordée en V.1.

39 Par exemple, Vergez-Couret (2006, 2007) sur la relation d'élaboration, proposition relationnelle très utilisée.

> discours) quelle que soit sa forme d'existence (personne humaine, événement, entité abstraite)." (Charolles 1988:8)

Deux niveaux de représentation peuvent être distingués si l'on considère le **principe figure-fond**. Ce principe gestaltiste considère que notre perception fait une distinction entre la figure, qui se détache et possède un contour défini, et le fond au contour moins net duquel se détache la figure. En considérant que le langage est structuré par notre perception du monde, la distinction entre figure et fond constitue un des principes fondamentaux de l'organisation discursive des informations. Notre capacité à articuler les informations résulte essentiellement de notre perception : elle permet, d'une part, d'isoler les entités, de les identifier (figure) et, d'autre part, de les positionner dans l'espace et de leur attribuer une action et/ou des informations qui les spécifient (fond).

Au niveau de l'organisation discursive, le principe figure-fond permet, entre autres, d'expliquer la relation entre **les entités et les *settings*.** Nous avons présenté la construction d'un *text-world* comme nécessitant une étape de mise en place que nous avons comparée à la construction des fondations d'une habitation (II.1.1). Cette étape revient à poser le décor, à définir les propriétés du fond sur lequel se déroulent le ou les procès que le locuteur veut représenter. Ainsi, les procès et les entités qui y participent se trouvent au niveau 'figure', tandis que les *settings* se situent au niveau 'fond'.

Nous supposons qu'il y a des continuations idéationnelles tant au niveau des figures que du fond. Cela revient à considérer que plusieurs modes de segmentation peuvent cohabiter entre une segmentation au niveau de la 'figure' et une segmentation au niveau du 'fond'. L'exemple III.2 (repris de la partie précédente) affiche un regroupement au niveau du fond géré par les différentes localisations temporelles et un regroupement au niveau de la figure géré par des progressions thématiques autour de la notion d'écarts et de proportions de bacheliers. Nous considérons également que le titre prépare les notions générales : « *Baccalauréats, scolarité, société* » qui constitueront les figures principales de la section.

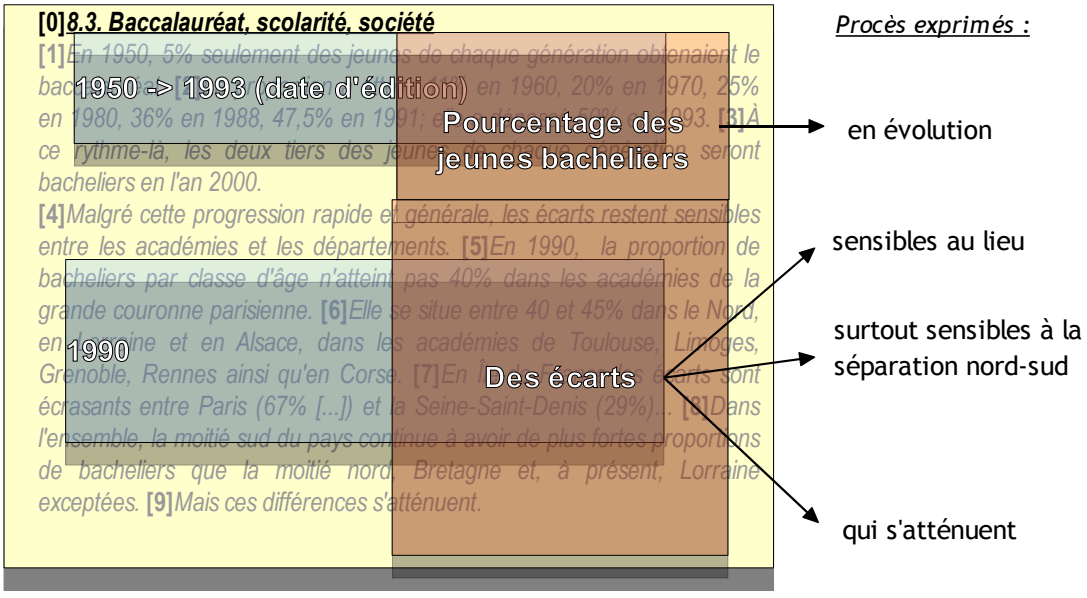


Figure III.2 : Représentation des différents niveaux de continuation idéationnelle

(III.2) **[0]*8.3. Baccalauréat, scolarité, société*** [titre niveau 2]
[1]*En 1950, 5% seulement des jeunes de chaque génération obtenaient le baccalauréat.* **[2]***La proportion a atteint 11%*

en 1960, 20% en 1970, 25% en 1980, 36% en 1988, 47,5% en 1991 ; elle a dépassé 50% en 1993. **[3]***À ce rythme-là, les deux tiers des jeunes de chaque génération seront bacheliers en l'an 2000.*
[4]*Malgré cette progression rapide et générale, les écarts restent sensibles entre les académies et les départements.* **[5]***En 1990, la proportion de bacheliers par classe d'âge n'atteint pas 40% dans les académies de la grande couronne parisienne.* **[6]***Elle se situe entre 40 et 45% dans le Nord, en Lorraine et en Alsace, dans les académies de Toulouse, Limoges, Grenoble, Rennes ainsi qu'en Corse.* **[7]***En Île-de-France les écarts sont écrasants entre Paris (67% [...]) et la Seine-Saint-Denis (29%)...* **[8]***Dans l'ensemble, la moitié sud du pays continue à avoir de plus fortes proportions de bacheliers que la moitié nord, Bretagne et, à présent, Lorraine exceptées.* **[9]***Mais ces différences s'atténuent. [ATLAS_2]*

Cette représentation associée à celle effectuée selon les relations rhétoriques illustre l'existence de plusieurs modes d'organisation du texte. Fait remarquable, tous ces modes organisationnels semblent prendre en compte le découpage en paragraphes et la relation entre le titre et sa section[40].

Contrairement à Grosz & Sidner, nous ne mettons pas de hiérarchie entre les continuations intentionnelles et les continuations idéationnelles. Grosz & Sidner suppose qu'un changement de but local discursif entraîne l'empilement d'un nouvel espace focus et de fait, une remise à zéro des entités accessibles. Chaque changement d'intention discursive entraîne automatiquement une rupture de la continuation idéationnelle.

Nous soutenons une certaine indépendance des différents niveaux. Walker (2000) propose de remplacer le modèle en « pile » de Grosz & Sidner par un modèle en « cache ». Dans ce modèle, les entités accessibles sont rangées dans un cache (et non une pile) qui correspond à la mémoire tampon (voir II.1.2). Cette mémoire temporaire conserve les informations reçues lors de l'interprétation d'un segment de discours. A chaque cycle de lecture[41], un tri est fait entre les entités à retenir dans le cache et les entités à ranger dans la mémoire principale. Ainsi, certaines entités anciennement activées peuvent être retenues dans le cache, comme par exemple, lorsqu'un but local subordonné arrive.

> « Le modèle en cache maintient la distinction de Grosz & Sidner entre la structure des états intentionnel et attentionnel. Cette distinction est cruciale, cependant le modèle en cache ne postule pas que l'état attentionnel est isomorphe à la structure intentionnelle. Par exemple, quand une intention est reconnue comme subordonnée par rapport à l'intention en cours, il se peut que des entités nouvelles doivent être créées dans le cache ou récupérées dans la mémoire principale, toutefois les entités anciennes restent dans le cache tant qu'elles ne sont pas déplacées » (Walker 2000:36)

Les entités retenues dans le cache sont actives, alors que les entités déplacées en mémoire principale sont plus ou moins accessibles selon la distance chronologique et les relations rhétoriques entre les segments.

Le modèle en cache permet de 'sauter' des énoncés, d'outrepasser la linéarité des textes, ce qui est nécessaire pour expliquer la réalité linguistique des chaînes de références. Nous verrons un exemple de 'sauts' en partie III.3.2 consacrée à la théorie du centrage (Grosz *et al.* 1995, Walker *et al.* 1998), dérivé algorithmique du modèle de Grosz & Sidner.

III.1.2.c) Continuations textuelles : autour d'une même texto-stratégie

Nous pouvons tout d'abord parler de continuation textuelle pour définir ce qui relie les énoncés d'une même unité textuelle. L'écrit possède des unités textuelles propres qui organisent le texte, qui lui confère une architecture textuelle (Virbel 1986, Luc *et al.* 2001). Cette architecture participe à la cohésion d'un texte en entraînant des attentes chez le lecteur sur lesquelles reposent parfois l'organisation du contenu du texte (Pascual 1991, Goutsos 1996, voir partie

40 Les parties V.2 et V.3.2 s'intéressent à ces deux unités textuelles qui créent de la segmentation à niveau essentiellement textuel.

41 Pour Grosz & Sidner, ainsi que pour Walker, les cycles de lectures correspondent à un segment de discours défini par un but discursif local, un marquage textuel et un ou plusieurs états attentionnels.

II.2.2). Ces attentes sont du type : les propos contenus dans un même paragraphe suivent un principe de « continuité par défaut » (Stark 1988, III.2.1), une section est homogène relativement à une thématique abordée et/ou à un but global de discours particulier (Hồ-Đắc *et al.* 2004 , V.2), les items d'une liste sont en relation de parallélisme (*cf.* Luc 2000). Pour l'exemple de la continuation textuelle des sections, le critère de regroupement se trouve fréquemment exprimé dans le titre de la section. Un titre de section comme *« Continuations textuelles : autour d'une même texto-stratégie »* indique qu'*a priori* toute la section 'parle' des continuations textuelles. De même, le titre *« introduction »* signale que toute la section sert d'introduction au reste du texte. Pour les paragraphes, si un élément détaché en initiale de paragraphes apparaît, il peut alors jouer le rôle d'un 'titre de paragraphe' (voir II.2.2). Sans élément détaché en initiale, aucun critère de regroupement n'est spécifié, ce qui n'enlève pas pour autant la continuation textuelle assignée par défaut à cette unité textuelle.

Mais l'idée de continuation textuelle ne se limite pas seulement au découpage typodispositionnel du texte. Les texto-stratégies qui servent de critère de regroupement aux continuations textuelles peuvent répondre à un mode organisationnel plus complexe. Nous nous appuyons ici sur la notion de « continuités texto-stratégiques » (*Text-strategic continuities* - TSC) telles que définies par Virtanen (1992) sur la base des travaux d'Enkvist (1981, 1989).

> "*Text-strategic continuity* (TSC) may thus be defined as a thematic or topical uniform text-structuring orientation chosen to attain, in view of the communicative goal, a maximally profitable text organization, for the benefit of the text receiver." (Virtanen 1992 : 85)"

Une continuation textuelle correspond à un regroupement autour d'un même mode d'orientation. La fonction d'orientation (Chafe 1976, Dik 1997, Fries 1995a,b,c, voir partie IV.2) correspond à un 'guidage' de l'interprétation des propos. Ce guidage peut se faire en réquisitionnant un certain type de connaissances d'arrière-plan – c'est le cas typique des adverbiaux circonstanciels qui posent les *setting*[42] d'un *text-world* – ou en marquant de manière explicite l'organisation rhétorique du contenu – c'est le cas typique des marqueurs d'intégration linéaire qui indexent chaque item d'énumération pour le situer dans le déroulement de la structure énumérative (*premièrement, ensuite, finalement*) et des méta-marqueurs discursifs (*pour conclure*, *dans la partie I.*, etc.) voir Teufel 1999 et Hernandez 2004). Cette fonction justifie grandement le positionnement d'éléments détachés en initiale (typiquement les adverbiaux), ce que nous verrons en partie IV.2.

Proposer l'idée de continuations textuelles autour d'un même mode d'orientation consiste à identifier des segments dont l'organisation interne s'articule autour d'une même dimension. Ainsi, l'exemple III.3) présente une TSC autour de la dimension temporelle (mise en relief par le surlignage gris). Cette TSC se réalise par une série d'adverbiaux temporels qui orientent l'interprétation du contenu selon un axe chronologique.

(III.3) *L'année 2002 a été marquée par une série de scandales financiers qui ont ébranlé la confiance de l'investisseur américain - autant dire du citoyen - dans l'intégrité et la transparence des marchés financiers. L'enchaînement des faits, tout d'abord. La faillite d'Enron était déclarée en décembre 2001, celle de Global Crossing en janvier 2002. Pour son rôle dans l'affaire Enron, le cabinet d'audit Arthur Andersen était mis en examen en mars.* ***En juin,*** *Enron reconnaissait avoir versé un total de 310 millions de dollars en espèces à ses dirigeants au cours de l'année 2001 et worldcom corrigeait ses comptes de 3,8 milliards de dollars.* ***Le 21 juillet,*** *la faillite de worldcom était déclarée.* ***Le 24,*** *la Securities and Exchange Commission (SEC) portait plainte contre les dirigeants d'Adelphia, accusés d'avoir dissimulé 2,3 milliards de dollars de dettes dans des sociétés non consolidées.* ***En août,*** *l'ancien Chief Executive Officer (CEO) de imclone était mis en examen pour délit d'initié.* ***En septembre,*** *c'était au tour du CEO et du Chief Financial Officer (CFO) de Tyco d'être mis en examen pour corruption : il leur était reproché d'avoir détourné 600 millions de dollars, dont 170 millions de prêts personnels accordés par la société.* ***Enfin le 5 novembre 2002,*** *Harvey L. Pitt, président de la SEC et champion du laisser-faire réglementaire, était contraint de démissionner.*
Tous ces scandales se sont produits dans un contexte économique morose, très différent de l'euphorie des années 1990 : la " bulle Internet " a […][GEOPO_31]

42 La définition des *setting* est donnée dans la partie II.1.3.

La TSC de cet exemple se distingue par son mode d'orientation temporelle, de même que le segment surligné de l'exemple III.4 se distingue du reste du texte par sa structure énumérative marquée par une simple mise en forme (sauts de ligne, indentation et puces).

(III.4) *2. Les difficultés actuelles de la stabilisation macroéconomique*
Tous les États du flanc sud de la Russie ont été confrontés aux mêmes phénomènes macroéconomiques.
Malgré les spécificités de chaque pays, la transition peut être caractérisée par plusieurs étapes :
- baisse très importante des agrégats de production entre 1990 et 1992 ;
- éclatement de la zone rouble à l'automne 1993 ;
- hyperinflation au cours des années 1993-1994 ;
- reprise de la croissance en 1995-1996 ;
- impact de la crise russe (1998-1999).
Sur la décennie, le PIB a baissé en moyenne d'un tiers [...] [GEOPO_29]

Nous avons choisi délibérément un exemple de structure énumérative associée à un déroulement temporel afin de souligner la ressemblance entre une TSC temporelle et une structure énumérative. En effet, l'indexation des énoncés, que ce soit par un adverbial circonstanciel ou par une puce, crée une cohésion non pas lexicale mais purement textuelle. Le rapprochement entre les TSC et les structures énumératives est encore plus forte dans cet exemple où la TSC est encadrée d'une amorce et d'une conclusion comme dans les structures énumératives les plus complètes. Selon Luc (2000:102), « la structure énumérative comprend une amorce [phrase introductrice précédant l'énumération], une énumération (*i.e.* ensemble d'items) et parfois une conclusion ». Nous trouvons tous ces éléments en (III.3). Les expressions « *une série de scandales financiers* » et « *l'enchaînement des faits* » préparent le lecteur à une TSC temporelle, comme le ferait une amorce standard telle que « la transition peut être caractérisée par plusieurs étapes » en (III.4). Enfin, le démonstratif résumant « *tous ces* scandales » permet une encapsulation de tous les faits relatés dans l'énumération (la structure énumérative de (III.3) n'a pas de conclusion).

La différence entre ces deux exemples se situe au niveau de l'importance discursive accordée à la dimension temporelle. Dans l'exemple III.3, les adverbiaux temporels jouent conjointement au niveau de la composante textuelle et idéationnelle ; alors qu'en (III.4), ils ne jouent qu'au niveau idéationnel, le regroupement selon une certaine organisation textuelle se faisant par des tirets, totalement dénués de valeur idéationnelle[43]. Les TSC du type de celles rencontrées en (III.3) créent des « zones d'orientation», caractérisées par une organisation par **orientation** (ou indexation) plutôt que par **connexion (**voir partie II.3.2).

L'organisation par connexion consiste à lier les énoncés les uns à la suite des autres par progressions thématiques (voir III.3.1). L'exemple III.1 repris ici en (III.5) montre justement une organisation par progressions thématiques autour des thèmes du pourcentage des jeunes bacheliers et des variations géographiques de ces pourcentages.

(III.5) **[0]*8.3. Baccalauréat, scolarité, société*** [titre niveau 2]
[1]*En 1950, 5% seulement des jeunes de chaque génération obtenaient le baccalauréat.* **[2]***La proportion a atteint 11% en 1960, 20% en 1970, 25% en 1980, 36% en 1988, 47,5% en 1991 ; elle a dépassé 50% en 1993.* **[3]***À ce rythme-là, les deux tiers des jeunes de chaque génération seront bacheliers en l'an 2000.*
[4]*Malgré cette progression rapide et générale, les écarts restent sensibles entre les académies et les départements.* **[5]***En 1990, la proportion de bacheliers par classe d'âge n'atteint pas 40% dans les académies de la grande couronne parisienne.* **[6]***Elle se situe entre 40 et 45% dans le Nord, en Lorraine et en Alsace, dans les académies de Toulouse, Limoges, Grenoble, Rennes ainsi qu'en Corse.* **[7]***En Île-de-France les écarts sont écrasants entre Paris (67% [...]) et la Seine-Saint-Denis (29%)...* **[8]***Dans l'ensemble, la moitié sud du pays continue à avoir de plus fortes proportions de bacheliers que la moitié nord, Bretagne et, à présent, Lorraine exceptées.* **[9]***Mais ces différences s'atténuent.* [ATLAS_2]

43 Ho-Dac *et al.* (2001) se penchent justement sur cette capacité des adverbiaux antéposés à jouer sur ces deux composantes discursives.

Qu'il s'agisse de continuations intentionnelles, idéationnelles ou textuelles, l'organisation discursive est un mélange – plus ou moins savant selon les auteurs – de ces trois niveaux de structuration. Les segments délimités selon les différentes continuations évoquées se succèdent, s'enchâssent ou se superposent, dans une certaine séquentialité du discours.

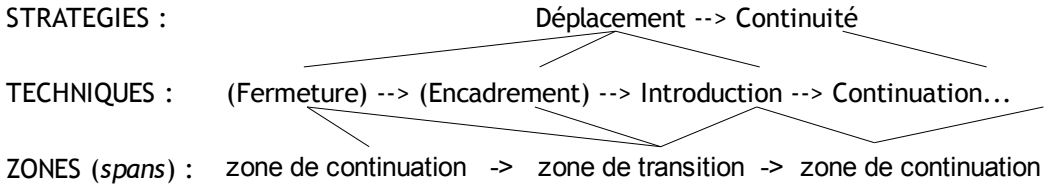
III.2. Continuité et discontinuité : la séquentialité du discours

Que ce soit au niveau des processus de production ou d'interprétation, le phénomène de segmentation textuelle répond à la question de la linéarité du texte. L'obligation de linéariser sa représentation mentale contraint le locuteur à placer les informations en séquence, à la suite les unes des autres. Or, la représentation que l'on souhaite communiquer ou reconstruire est généralement plus complexe qu'une simple succession d'informations qui s'ajoutent les unes aux autres. Pour pallier ce manque d'isomorphisme, le locuteur peut signaler à la surface du texte comment les informations sont organisées. Cela consiste à indiquer au lecteur comment les informations sont à intégrer les unes par rapport aux autres : si elles sont en relation de continuité ou de discontinuité.

> "An important task for the writer is to indicate discontinuity within the larger presupposed continuity of the text. In other words, the writer is faced with the tasks to manage the interaction through discourse in sequential terms and to segment discourse into chunks and indicate their boundaries, *i.e*. the discontinuity between one and another" (Goutsos 1996:504)

Goutsos (1996) propose un modèle de représentation des relations séquentielles structurant les textes expositifs[44]. Le modèle dégagé se schématise par la figure III.3.

La séquentialité du discours se construit à trois niveaux. À un niveau cognitif, des **stratégies** correspondent aux décisions du locuteur quant à la 'mise en texte' de ses propos. Pour satisfaire ces décisions, différentes **techniques** linguistiques sont disponibles. Ces techniques proposent différentes façons de signaler au lecteur si les informations entrantes sont en continuité des informations préalablement enregistrées ou en discontinuité de celles-ci. Il en résulte une organisation du texte selon des séquences de zones de continuation et de zones de transition.



() : optionnel, --> : suivi de

Figure III.3 : Modèle de séquentialité (Goutsos 1996)

Les trois techniques de déplacement distinguées dans ce modèle concernent respectivement : la fermeture d'une zone de continuation, la mise en place d'un nouveau critère d'orientation, l'ouverture d'une nouvelle zone de continuation. Ces trois techniques participent à la mise en place de relations de discontinuité.

La limite entre les relations de continuité et les relations de discontinuité n'est pas aussi nette que pourrait le faire croire le titre de cette section. Une grande partie des relations entre unités de discours est de l'ordre du « déplacement ». Par déplacement, nous entendons la progression d'une entité ou d'une circonstance à travers

44 Son étude se base sur l'observation de trois corpus de textes expositifs anglais (29 000 mots au total).

différents 'plans'. Ainsi, une continuité référentielle peut progresser à travers différentes localisations temporelles. À chaque changement temporel, nous considérons que la continuité référentielle se 'déplace', ce qui donne un effet de continuité discontinue, nous parlons de « déplacement » (III.2.4).

Les stratégies de discontinuités demandent généralement plus d'efforts, que ce soit au niveau de la production des textes ou de leur interprétation. Du point de vue de la production, le locuteur doit indiquer le plus clairement possible tout changement afin d'éviter que ne s'exerce la continuité par défaut.

> "The strategy of shift plays a central role in the management of sequentiality, since, as suggested above, the indication of discontinuity in text is more significant than the indication of continuity – continuity being the default case." (Goutsos 1996:507)

Du point de vue de l'interprétant, une discontinuité entraîne une remise en cause du sous-monde jusque là en construction, ce qui suppose alors la recherche de nouvelles fondations. Nous distinguons deux types de stratégies de discontinuités qui correspondent à deux degrés différents de discontinuité : les déplacements et les ruptures.

Les parties suivantes concernent davantage la séquentialité à un niveau idéationnel ; car celle-ci se réalise par un marquage plus explicite qu'au niveau de la structure rhétorique (V.1) et car celle-ci donne accès au contenu du texte. La structure rhétorique ne se construit pas par des techniques de continuation mais par des techniques de rupture qui consistent à articuler deux propositions, deux unités textuelles, etc. autour d'une certaine relation rhétorique (III.2.5).

III.2.1. De la continuité par défaut

La continuité entre deux éléments peut se faire de deux façons selon que le locuteur marque explicitement ou non qu'il y a continuité. En effet, contrairement à la discontinuité, la continuité s'établit par défaut. Ainsi, toute nouvelle information s'interprète *a priori* comme étant liée de façon cohérente aux précédentes.

> "For a continuation span to be established, there must be at least one utterance linked to the immediately previous one, either by an explicit signal or by default (i.e. because no signal of shift occurs)." (Goutsos 1996:515)

Nous supposons, en tant qu'allocutaire, qu'un texte est cohérent et que les unités qui le composent sont ordonnées. Et même s'il n'y a pas de cohérence au niveau de la production, le lecteur est tenté d'en créer (règle du jeu du « cadavre exquis »).

Il s'agit là d'un principe cognitif d'économie plutôt que d'un principe textuel (cf. le principe général de cohérence (Charolles 1983, 1994) ou de pertinence (Sperber & Wilson 1986)). En effet, le fait de continuer à ajouter des informations dans le même sous-monde[45] demande moins d'effort cognitif que de devoir construire un nouveau sous-monde pour y intégrer la nouvelle information. Gernsbacher fait référence à l'acte d'ajouter l'information au sous-monde en cours par le terme *mapping*.

> "According to the Structure Building Framework, comprehenders will map incoming information onto a mental structure when that incoming information coheres with the previous information. Mapping incoming information onto an existing structure or substructure takes less cognitive effort than shifting to initiate a new structure or substructure" (Gernsbacher & Robertson 2002:121)

Cette continuité par défaut peut concerner tous les critères de regroupement possibles. Ainsi le critère de regroupement en cours continue d'exister tant qu'aucune configuration d'indice ne signale à l'allocutaire qu'il y a un

45 La notion de sous-mondes a été présentée en partie II.1.4.

changement de topique, de circonstance (au sens de *setting*[46]), de modalité d'énonciation, d'intention, de point de vue, etc… Le terme de « portée » sert à qualifier cette continuation d'un critère d'interprétation[47].

L'exemple III.6 montre la continuité par défaut de critères spatiaux. Les adverbiaux extra-prédicatifs qui correspondent ici à des introducteurs de cadres spatiaux ont une portée qui dépasse leur phrase d'accueil. Cette portée s'achève par la présence d'une nouvelle référence spatiale, signalant un déplacement au niveau de la localisation spatiale (la portée des références spatiales est ici délimitée par le symbole ►). Ce déplacement est également marqué par l'expression « *au contraire* » succédant au deuxième adverbial spatial (*ce sont* ***au contraire*** ...).

(III.6) ***Au sud-est d'une ligne Nancy-Dijon-Poitiers-Bordeaux,*** *en 1990-91 plus du tiers des jeunes obtiennent le baccalauréat, à l'exception cependant de cinq départements, la Haute-Saône, l'Ain, la Dordogne, les Landes et les Charentes, que l'on retrouve habituellement dans le groupe des départements à scolarisation déficiente. La Haute-Garonne dépasse 45% ; les Alpes-Maritimes atteignent 41% ; le Rhône frôle 40% ; et de la Savoie à Montpellier, les taux oscillent le plus souvent entre 35 et 40%.►* ***Dans la moitié nord-ouest du pays,*** *ce sont au contraire les pourcentages supérieurs au tiers qui sont l'exception ; ils concernent Paris et sa banlieue occidentale, la Bretagne sauf le Morbihan. La proportion de bacheliers dans la classe d'âge tombe à moins de 30% dans la plupart des autres départements, et même à moins de 25% en banlieue nord et est de Paris, ainsi que dans la plupart des départements de la grande couronne, l'Eure et l'Eure-et-Loir, l'Yonne et l'Aube, l'Orne, l'Aisne, et plus loin les Ardennes, la Meuse.► [ATLAS_2]*

III.2.2. Des principes par défaut différents selon les genres discursifs

La continuité par défaut d'un critère spatial ou temporel tel que l'illustre l'exemple III.6 n'est pas générale. Un autre principe organisateur semble exister *par défaut* selon le genre discursif et les relations de discours qu'il implique. Le Draoulec & Péry-Woodley (2003, 2005) distinguent, pour le temps, deux principes organisateurs qu'elles nomment *progression in time vs. indexing through time* et que l'on retrouve dans la distinction qu'opère Virtanen (1992:98) entre les notions de *continuity vs. unity*.

Nous avons d'un coté la continuité par défaut (l'indexation chez Le Draoulec & Péry-Woodley et l'unité chez Virtanen) et de l'autre l'idée d'une évolution de la référence spatiale/temporelle selon la progression des événements dans le « text-world ».

> "I shall retain a terminological difference between unity and continuity of time, using the former more in the Aristotelian sense of an approximate fit between story-time and text-time, and the latter in the sense of a chain of references in the text to different points on a temporal frame." (Virtanen 1992:98)

La distinction entre unité et progression se retrouve au niveau de la localisation spatiale[48]. Alors que dans des textes descriptifs les références spatiales sont généralement fixes et ont une portée ; dans les textes procéduraux impliquant le mouvement, tels que les guides de voyage, les adverbiaux extra-prédicatifs spatiaux perdent leur capacité de portée pour ne servir que de point de départ au mouvement. L'exemple III.7 constitue un extrait d'un guide de voyage présentant ce type de progression spatiale.

(III.7) ***A Oslo,*** *prendre la direction vers Lillehammer.► Poursuivre la route vers Åndalsnes par la célèbre et impressionnante "Route des Trolls" (Trollstigen). Continuer vers Ålesund et redescendre vers Lom en traversant le Geirangerfjord.* ***De Lom à Kaupanger,*** *suivre la route qui traverse le Sognefjell et admirer la beauté des paysages du Jotunheimen.► Une*

46 Une définition de la notion de *setting* est données en partie II.1.3.

47 Nous empruntons le terme de « portée » à l'hypothèse de l'encadrement telle que définie par Charolles (1997) en modifiant quelque peu sa définition. En effet, selon l'hypothèse de l'encadrement du discours, la portée est le résultat d'une relation d'indexation et non la simple conséquence d'une absence d'indice de discontinuité (voir partie III.3.3).

48 Nous retenons le terme d'« unité » propre à Virtanen et le terme de « progression » propre à Le Draoulec et Péry-Woodley non pas pour occasionner une certaine confusion chez le lecteur, mais car nous réservons le terme d'« indexation » à un procédé de mise en relation entre une expression et un segment (II.3.2) et le terme de « continuité » au résultat d'un regroupement de plusieurs unités.

agréable croisière dans le Sognefjord vous conduira de Kaupanger à Gudvangen d'où vous reprendrez la route pour rejoindre Flåm. Emprunter la E 16, également appelée Eventyrveien et arrivant à Bergen, une multitude de curiosités vous attend. Prendre la Rv 7 et offrez-vous une "halte fraîcheur" un peu avant Nordheimsund et continuez la route en admirant les jolies rives de Hardangerfjord avec ses vergers de pommiers jusqu'à Bruravik. Un ferry vous conduira de Bruravik à Brimnes où vous continuerez par la Rv 7 pour rejoindre Gol puis retour par la merveilleuse vallée du Hallingdal pour arriver enfin à Oslo. [extrait du guide officiel de Norvège : http://www.visitnorway.com, rubrique « suggestion de circuits »]

On voit ainsi dans cet exemple qu'une fois effectué le trajet ayant pour point de départ *Oslo*, cette référence spatiale n'est plus valide. Concernant la référence spatiale exprimée par *De Lom à Kaupanger,* il est plus difficile de juger sans la carte associée au pays à partir de quand nous ne sommes plus entre Lom et Kaupanger. Mais nous sentons très vite que cette référence spatiale est supplantée par toutes les autres références réalisées par les nombreux adverbiaux intra-prédicatifs.

Dans ce type d'organisation textuelle, la référence spatiale ou temporelle se déplace par défaut avec l'enchaînement des événements – ici, des déplacements – et alors la continuité spatio-temporelle consiste en la succession (chrono)logique – dite « iconique » – de ces événements / déplacements.

Le Draoulec & Péry-Woodley (2003, 2005) étudient les adverbiaux temporels selon qu'ils sont utilisés dans des textes narratifs *vs.* non-narratifs. Dans ce travail, les auteures montrent que les textes narratifs présentent de nombreuses suites de propositions liées par la relation de narration. Dans ces suites, le phénomène d'encadrement temporel (regroupement de plusieurs propositions autour de la même référence temporelle exprimée par un adverbial en initiale) est dominé par celui de la succession propre à la narration, l'adverbial temporel jouant alors plus un rôle de jalon à partir duquel se déroulent les événements, comme c'est le cas pour *En 1933* et *En 1938* dans l'exemple III.8.

(III.8) ***En 1933,*** *il [Klaus Mann] fonda à Amsterdam la revue antinazie « Die Sammlung ». Il sillonna l'Europe pour mobiliser les intellectuels contre le fascisme, donna des conférences, écrivit des articles virulents contre le régime hitlérien, notamment dans le « Pariser Tageblatt », journal des Allemands antinazis en France, et collabora au cabaret satirique dirigé par soeur Erika, « Die Pfeffermühle » (le moulin à poivre).* ***En 1938,*** *il se rendit en Espagne pour faire des reportages sur la guerre civile ; il prit partie pour les Républicains dans ses articles très polémiques. [exemple issu de Le Draoulec & Péry-Woodley 2005].*

Dans les textes non narratifs, au contraire, l'adverbial pose un cadre à référence temporelle fixe dans lequel les faits relatés sont à interpréter.

(III.9) En 1950, 5% seulement des jeunes de chaque génération obtenaient le baccalauréat. La proportion a atteint 11% en 1960, 20% en 1970, 25% en 1980, 36% en 1988, 47,5% en 1991 ; elle a dépassé 50% en 1993. À ce rythme-là, les deux tiers des jeunes de chaque génération seront bacheliers en l'an 2000.
Malgré cette progression rapide et générale, les écarts restent sensibles entre les académies et les départements. ***En 1990,*** *la proportion de bacheliers par classe d'âge n'atteint pas 40% dans les académies de la grande couronne parisienne. Elle se situe entre 40 et 45% dans le Nord, en Lorraine et en Alsace, dans les académies de Toulouse, Limoges, Grenoble, Rennes ainsi qu'en Corse. En Île-de-France les écarts sont écrasants entre Paris (67% de bacheliers dans la classe d'âge) et la Seine-Saint-Denis (29%)... Dans l'ensemble, la moitié sud du pays continue à avoir de plus fortes proportions de bacheliers que la moitié nord, Bretagne et, à présent, Lorraine exceptées.► Mais ces différences s'atténuent. [ATLAS_2]*

Ainsi, dans ce dernier exemple, *En 1990* fixe une référence temporelle qui reste valable pour sa phrase d'accueil et les deux phrases suivantes. On peut effectivement affirmer que l'interprétation de la phrase commençant par « *Dans l'ensemble, ...* » se fait toujours relativement au critère temporel : « *En 1990* ».

Il semble donc que, dans des 'suites narratives', les adverbiaux en position initiale perdent leur capacité d'indexer de larges portions de texte, la progression des événements dans le temps propre à la relation de narration se faisant

par défaut. La distinction entre indexation et progression est particulièrement bien illustrée par l'exemple III.10. Ici, la répétition de la référence à l'année 1830 est nécessaire afin d'indexer toute la portion de texte. Cette répétition empêche l'effet de progression provoqué par les relations de narration établies, par défaut, entre les différents événements décrits.

(III.10) [49]***En 1830,*** *La Mode publie El Verdugo, dans sa livraison du 30 janvier : c'est la première oeuvre « Honoré de Balzac ».* ***En 1830,*** *encore signées « Balzac », paraissent les Scènes de la vie privée, six nouvelles dont le thème est l'échec, toujours semblable, toujours varié, de la « vie privée » : ce sont La Vendetta, Les Dangers de l'inconduite (qui deviendra Gobseck), Le Bal de Sceaux, Gloire et malheur (qui deviendra La Maison du chat-qui-pelote), La Femme vertueuse (qui deviendra Une double famille), La Paix du ménage.* ***Dès cette année 1830*** *se trouvent inventés le mot, l'usage, le principe des Scènes. Les Scènes de la vie privée se gonfleront de beaucoup d'autres nouvelles et de maint roman. La section comprendra vingt-sept titres, dont Modeste Mignon, écrit en 1844****. Dans la même année 1830,*** *paraissent plusieurs contes ou nouvelles qui ne ressortissent nullement au genre des Scènes :* ► *ces œuvres, par exemple L'Élixir de longue vie* et *Sarrasine, sont encore indépendantes, elles restent en attente de rubrique, jusqu'au jour où l'idée de la rubrique appropriée sera née. [PEOPL_7]*

En utilisant des mécanismes de co-référence temporelle, l'auteur marque l'unité temporelle de l'extrait. Il indique explicitement que la validité de la référence temporelle à l'année 1830 continue.

III.2.3. Continuité marquée et continuité référentielle

Le fait de marquer une continuité (et non de laisser faire la continuité par défaut) est nécessaire dans deux cas principaux :

- les continuités référentielles, *i.e.* les continuités référentielles autour d'une même entité, et
- les unités temporelles ou spatiales qui vont à l'encontre d'une progression par défaut de la dimension temporelle ou spatiale (comme l'a illustré l'exemple III.10)

Du point de vue de la composante idéationnelle, la phrase se construit autour d'un ou plusieurs procès mettant en jeu une ou plusieurs entité(s). Il est donc nécessaire (et obligatoire dans le cas des langues à sujet obligatoire comme le français) de préciser à chaque fois l'entité ou les entités impliquée(s) par ces procès. Ainsi, si l'entité en jeu dans une phrase est identique à une entité déjà en jeu précédemment, une continuité référentielle se met en place sous la forme de ce que l'on appelle communément une chaîne de référence, telle que définie par Corblin.

> « On appelle chaîne de référence une suite d'expressions d'un texte entre lesquelles l'interprétation établit une identité de référence. » (Corblin 1985:27)

De nombreuses techniques participent à la construction des chaînes de référence et de nombreuses formes permettent d'établir une identité référentielle. La technique la plus fréquente pour marquer une continuité referentielle est l'utilisation d'expressions anaphoriques qui sont des expressions « dont l'interprétation référentielle dépend d'éléments déjà saillants ou manifestes » (Kleiber 1994:22)

Cette définition fait apparaître deux traits caractéristiques de l'anaphore :

- l'expression anaphorique est incomplète, puisqu'il lui manque de l'information pour pouvoir référer à une entité spécifique par elle même ;
- l'occurrence d'une expression anaphorique implique l'existence d'éléments d'information donnée ou de référents accessibles (dans le contexte textuel ou mémoriel de l'anaphorique).

Ces deux traits complémentaires soulignent la fonction de connexion que jouent les expressions référentielles en *obligeant* un raccord au discours précédent du fait de leur incomplétude sémantico-pragmatique. Une des formes principales des expressions anaphoriques est la forme pronominale. Les pronoms constituent des expressions

49 Extrait du portrait de Balzac issu du sous-corpus PEOPL.

remarquables par leur absence de contenu représentationnel, absence qui leur confère un rôle essentiellement instructionnel. En effectuant une pronominalisation, le locuteur indique à son destinataire que le référent dont il parle est toujours le même. Les pronoms constituent ainsi des marqueurs privilégiés pour les stratégies de continuité[50] et plus particulièrement pour des stratégies de continuité topicale (Virtanen 1992).

> "text continuity is related to successive reference to the same entity through pronouns. It has already been mentioned that continuation spans usually contain pronominalization, whereas transition involves re-nominalization. The use of pronominal reference creates cohesive chains (Halliday and Hasan, 1976:15), which extend the technique of continuation for more than two adjacent sentences." (Goutsos, 1996:517)

Schnedecker (2005) démontre de façon tout à fait pertinente l'influence du genre discursif sur le marquage de la continuité référentielle. Son étude se base sur des portraits journalistiques de personnalités[51]. Ces textes présentent la caractéristique d'être mono-référentiels, c'est-à-dire qu'ils sont à propos d'un même référent du début à la fin, référent introduit par le titre de l'article. Il est donc évident que la gestion des chaînes de référence se fait différemment que dans des textes pluri-référentiels. Notre étude en corpus se base justement sur trois sous-corpus dissociant des textes se situant équitablement sur un axe allant d'un pôle mono-référentiel à un pôle pluri-référentiel.

D'après les théories sur l'accessibilité des référents, la technique la plus économique pour signaler que l'on continue de parler à propos d'un même référent est la pronominalisation. Cependant, d'autres techniques sont utilisées, et cela même si l'on a affaire à des textes mono-référentiels. Dans ses données, Schnedecker comptabilise une occurrence de nom propre pour deux pronoms personnels, et remarque que, dans près de la moitié des textes, les syntagmes nominaux – SN – à tête lexicale sont aussi nombreux que les pronoms.

> « Compte tenu du fait que les textes sont centrés tout leur long sur un même référent et que celui-ci reste infailliblement accessible, la présence d'un marqueur réputé de basse accessibilité comme le Np est, en effet, plus qu'étonnante. Le raisonnement serait le même du reste si l'on tenait compte également des marqueurs dits de moyenne accessibilité [SN définis et démonstratifs] » (Schnedecker 2005:95)

Ces autres techniques, en plus de connecter une expression à une référence exprimée précédemment, participent également à la segmentation du discours en signalant un déplacement.

III.2.4. Des stratégies de déplacement : de la continuité discontinue

L'exemple III.10 présente différentes techniques de reprise réalisées sous trois formes différentes faisant toutes référence à l'année 1830 :

- Une redénomination qui correspond à une reprise à l'identique
  En 1830,
- Une reclassification qui correspond à la reformulation d'un référent
  Dès cette année 1830,
- Une reprise avec insistance sur la similitude de référence par l'adjectif *même*
  Dans la même année 1830,

Ces différentes techniques sont considérées ici comme autant d'indices permettant l'unité temporelle de ce passage. Cependant, comme le remarque Goutsos, ces expressions apparaissent dans des portions de textes où se déroule une transition, un déplacement. En effet, le fait d'avoir ces reprises ponctue l'extrait et, tout en lui donnant une

50 La partie V.4.3.a définit les pronoms et leur sens instructionnel, les pronoms étant considérés comme des marqueurs discursifs.

51 Ces articles sont issus du journal Le Monde (numéros parus entre mai et juillet 2003).

même indexation temporelle, lui confère une segmentation textuelle qui correspond finalement à une organisation conceptuelle des oeuvres écrites par Balzac cette année là.

De telles formes, tout en maintenant un référent activé, peuvent signaler ou plutôt accompagner une discontinuité. Ainsi, un changement de sous-monde peut nécessiter la reformulation du référent activé, afin de réancrer celui-ci dans les nouvelles définitions de ce sous-monde, de cette sous-structure.

Le recours à cette technique de co-référence 'redondante' signale ou plutôt accompagne un déplacement dans la description du référent. Schnedecker (2005), qui s'intéresse à la construction des chaînes de référence dans des portraits met en évidence deux types bien dissociés de chaînes : des chaînes homogènes construites par des alternances entre nom propre et pronominalisation, et des chaînes hétérogènes construites par des pronoms, parfois des noms propres mais surtout par des descriptions définies ou démonstratives.

Nous nous trouvons donc face à des expressions qui en même temps qu'elles continuent l'expression d'un référent, signalent que la façon de traiter de ce référent change, que le cadre dans lequel il est à considérer s'est déplacé. Ces reformulations peuvent correspondre à un changement de *setting* (*i.e.* changement de point de vue, de situation spatiale et/ou temporelle, de cadre notionnel) mais également à une articulation rhétorique ou un changement de visée discursive. Les changements de *setting* sont généralement exprimés par un introducteur de cadre. Cependant, le passage d'une localisation spécifique à une autre, notamment à la localisation impliquée par la situation d'énonciation, n'est pas toujours explicité par un circonstant extra-prédicatif. Les articulations rhétoriques sont elles aussi plus ou moins explicitées par la présence d'un connecteur ou d'un adverbial textuel (voir V.3.4). Il est en effet fréquent de ne pas en avoir un marquage explicite (voir V.1).

Virtanen appelle ces déplacements des "étapes dans la continuité du texte" : *"intra-continuity steps in the sense of new stages in the TSC"* (Virtanen 1992:89). Les déplacements peuvent être assimilés à des glissements. Ce dernier terme souligne le caractère relativement impalpable de ce type de discontinuité. Il est très fréquent de 'sentir' un déplacement sans pour autant 'lire' un marquage explicite à la surface du texte. Les déplacements, surtout lorsque le critère de regroupement est rhétorique, semblent souvent inférés. C'est dans ce sens que, lors de l'élaboration de la RST, Mann et Thompson refusent de corréler des formes aux relations rhétoriques dégagées. Ils récusent d'ailleurs l'idée même d'un signalement de telles relations.

> "Our point is that it is the implicit relations which are important, with the conjunctions acting occasionally to constraint the range of possible relational propositions which can arise at a given point in a text." (Mann & Thompson 1986:71)

Nous sommes tout à fait d'accord avec l'idée qu'il n'existe pas de corrélations 'absolues' entre une forme et une relation de discours, qu'elle soit rhétorique ou idéationnelle. Cependant, cela ne récuse pas l'idée d'un marquage au sens large de telles relations. Le chapitre V présente justement notre définition de cette idée d'un marquage large ainsi que la notion d'«indices de séquentialité ».

III.2.5. Des stratégies de rupture

Les stratégies de rupture consistent en une discontinuité non empreinte de continuité. Elles consistent à signaler que les informations entrantes sont à intégrer dans une structure différente de celle jusque là en cours. Leur marquage semble davantage explicite – ou du moins nécessaire – que celui des déplacements.

Ces stratégies peuvent correspondre par exemple à un changement radical des entités, circonstances et procès en jeu dans le discours, au passage à un autre niveau d'énonciation (description / commentaire), au changement de mode organisationnel. Le rôle des titres semble essentiel dans ce type de stratégie[52].

Notre distinction entre déplacement et rupture rejoint dans une approche plus généraliste celle faite par la théorie du centrage (Grosz *et al.* 1995, Walker *et al.* 1998) entre les déplacements en douceur et des déplacement brutaux. La théorie du centrage définit quatre degrés de continuité selon la présence, la fonction, la position et la nature de l'élément qui fait lien entre deux phrases. Nous présentons ces quatre degrés plus loin, dans la partie III.3.2.

Un autre type de rupture est défini par Virtanen qui distingue elle aussi des phénomènes de déplacement (*shift*) et de rupture (*break*). Pour elle, la rupture se situe à un niveau purement textuel : il y a rupture dès lors que la TSC utilisée par le locuteur change.

> "[a break is] a change in the type of TSC, e.g. from a temporal text strategy to, say, a locative one. A break in one TSC is often accompanied by breaks in the other cooccuring continuities. Breaks in the TSCs may also signal boundaries between textual units, for instance, in the sense of introducing a new text type in multitype texts." (Virtanen 1992:90)

La notion de rupture étant associée à un changement de stratégie textuelle, le passage d'une séquence narrative à une séquence descriptive peut également être qualifiée de rupture. Si l'on considère la théorie compositionnelle des séquences soutenue par Adam (1999), un texte est un ensemble de séquences, chaque séquence pouvant répondre à un « fait de régularités » tel que narration, description, argumentation, explication, dialogue. Bien entendu, la notion de séquence chez Adam ne correspond par à notre définition qui repose sur des relations de continuité et discontinuité, tel que le conçoit Goutsos (1996).

Finalement, nous retrouvons les trois plans de continuation au niveau de la rupture. Un segment peut être en rupture sur le plan rhétorique, sur le plan idéationnel et sur le plan textuel. Ces trois natures de rupture s'illustrent facilement par le changement de section. Dans notre thèse, le passage de la partie VII.3 à la partie VII.4 est un bon exemple de rupture rhétorique puisque l'on passe d'une étape dans la présentation des analyses effectuées à la mise en place d'un petit didacticiel pour comprendre la présentation des résultats (nous ne savons d'ailleurs pas quelle représentation rhétorique serait adaptée pour indiquer les propositions relationnelles entre ces deux sections). Le passage d'un chapitre à l'autre est très souvent accompagné d'une rupture idéationnelle. Enfin, le chapitre V peut présenter une certaine rupture textuelle : sa première section plus argumentative défend notre définition du marquage discursif, à l'opposée des sections suivantes, plus descriptives, qui énumèrent les différents indices de séquentialité en position initiale.

III.3. Des modèles pour représenter la séquentialité du discours

Nous pouvons regrouper en deux catégories les modèles linguistiques qui nous permettent de représenter la séquentialité du discours. La premère catégorie cherche à représenter les relations de continuité, la seconde s'intéresse davantage aux relations de discontinuité. Alors que la première catégorie de modèles s'intéresse essentiellement aux relations de connexion qui s'établissent de phrase en phrase, la deuxième a pour objet d'étude le découpage du texte en blocs dépassant la taille de la phrase et la relation d'indexation. Nous avons donc d'une part

52 La partie V.2 précise la fonction discursive des titres de sections.

des modèles qui nous permettent de construire des segments par regroupement et d'autre part, des modèles qui nous permettent de construire des segments par délimitation.

Pour modéliser les continuités référentielles, deux modèles sont présentés : le modèle de progressions thématiques (Daneš 1974) et la théorie du centrage (Grosz *et al.* 1995). Ces modèles n'ont pas été établis sur la langue française mais sont assez abstraits pour s'y appliquer facilement. Ce n'est pas qu'il n'existe pas de modèle français sur les continuités référentielles mais qu'il n'existe pas de modèle opérationnel. La plupart des travaux français s'intéressent aux expressions co-référentielles – *i.e.* aux formes morpho-syntaxiques qui permettent la continuité – plus qu'aux parcours référentiels.

Au niveau de la délimitation des zones de discontinuités, il n'y a pas à proprement parler de modèle, si ce n'est celui de Goutsos, déjà défini. Cependant, notre réflexion s'appuie sur deux hypothèses : l'encadrement du discours (Charolles 1997) et les continuités texto-stratégiques (Virtanen 1992).

III.3.1. Les Progressions Thématiques – TP (Daneš 1974)

Le modèle établi par Daneš propose différents schémas de Progression Thématique – TP[53] – qui permettent de représenter le cheminement de la référence de propositions en propositions. Ce modèle envisage la phrase comme un message constitué d'un Thème (point de départ de ce qui est dit, voir partie IV.3) et d'un Rhème (ce qui est dit à propos du Thème). Il permet d'expliquer, d'un point de vue fonctionnel, la structuration de l'information dans le discours.

> "Our basic assumption is that text connexity is represented, *inter alia*, by thematic progression (TP). By this term we mean the choice and ordering of utterance themes, their mutual concatenation and hierarchy, as well as their relationship to the hyperthemes of the superior text units (such as paragraph, chapter, ...), to the whole text, and to the situation. Thematic progression might be viewed as the skeleton of the plot." (Daneš 1974:114)

Bien que Daneš ne revendique aucune restriction à un genre discursif, son modèle reste un modèle élaboré sur des exemples écrits[54], plus particulièrement sur des textes écrits expositifs. Les travaux qui s'inscrivent dans ce modèle s'appliquent d'ailleurs exclusivement à ce genre de texte. Nous retiendrons particulièrement les travaux de Combettes & Tomassone (1988) qui exposent quelques aspects linguistiques du texte informatif, Francis (1989) qui réalise une étude comparée sur la sélection thématique dans différents registres de l'écrit, Downing (2001) qui compare la structuration d'écrits professionnels classés selon que leur rhétorique va du divertissement à la persuasion, et Gil (2001) qui élabore un système d'annotation des TP basé sur des textes issus du *National Geographic*.

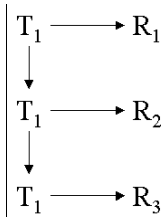
Trois types de TP sont généralement distingués[55] : des TP constantes, des TP linéaires et des TP dérivées (voir figure III.4).

53 Nous conservons l'abréviation anglaise d'origine (Thematic Progressions : TP).

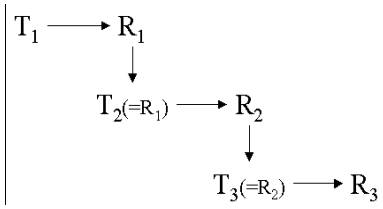
54 Grobet (2002:54) note la difficulté de l'appliquer à un corpus oral.

55 Enkvist propose un modèle similaire intitulé *Theme dynamics* (Enkvist 1973) dans lequel il conçoit également une progression de Rhème en Rhème (*rheme iteration*) et une progression de Thème en Rhème (*rheme regression*). Dubois (1987) préconise la réduction à deux types de TP (constante ou linéaire), chacune proposant une variante dérivée.

- Pour les TP constantes, ce qui se trouve en Thème reprend le Thème précédent.



- Pour les TP linéaires, ce qui se trouve en Thème reprend (en partie) le Rhème du message précédent.



- Pour les TP dérivées, le Thème dérive d'un hyper-Thème.

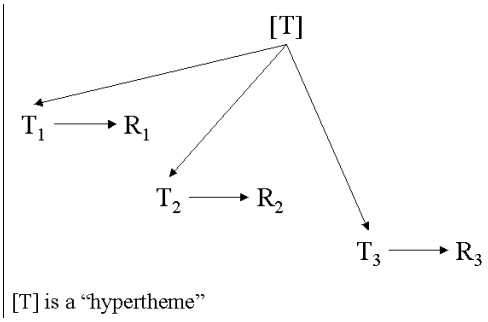


Figure III.4: Les trois principales progressions thématiques selon Daneš

Les analyses utilisant ce modèle portent parfois sur la caractérisation des genres textuels. Ainsi, Combettes & Tomassone (1988) soutiennent plus l'idée d'une caractérisation des textes informatifs par les TP constantes. Francis (1989) remarque que les progressions thématiques utilisées dans des textes expositifs suivent des schémas très hétérogènes et qui ne sont pas toujours assimilables à un schéma défini par Daneš (propos rapportés dans Fries 1995c).

Notre intérêt pour cette modélisation des relations de continuité se situe dans l'importance accordée à la position initiale (liée ici à la fonction de Thème[56]) dans le marquage des stratégies de continuité et dans la prise en compte de différents niveaux de TP grâce aux TP à Thème dérivé. Cependant, la détection automatique des TP reste improbable. Il faudrait faire appel à des ressources lexicales et sémantiques trop importantes afin de repérer les relations d'hyperonymie, d'hyponymie, de synonymie, de proximité lexicale... Car le modèle de Daneš est un modèle interprétatif et non un modèle TAL, dont le seul but est d'aider à la représentation de la structure du discours.

Nous utilisons cependant cette idée des progressions thématiques qui permet de parler de la continuité thématique sans entrer dans les concepts de chaînes de référence ou de chaînes topicales, trop précises pour notre étude. Il est important de noter que ce qui nous intéresse dans l'idée de progression thématique (et des chaînes de référence), ce n'est pas la recherche de l'antécédent, mais :

- le fait qu'elles marquent un processus de continuité, de reclassification (déplacement) ou de rupture ;

56 Nous utilisons la forme « Thème » avec majuscule afin de distinguer notre acception de la fonction Thème selon la terminologie de la Systémique Fonctionnelle (présentée dans la partie V.3) de l'acception courante liée à la notion de thématique telle que utilisée dans les calculs de segmentation thématique.

- Leur cohabitation avec les cadres de discours (voir Hồ-Đắc & Laignelet 2005).

L'idée de modélisation de la progression thématique se retrouve dans la théorie du centrage qui vise la caractérisation automatique du type de continuité/discontinuité reliant deux énoncés successifs.

III.3.2. Théorie du Centrage (Grosz et al. 1995, Walker et al. 1998)

La théorie du centrage (Grosz *et al.* 1995, Walker *et al. 1998*) est une théorie à visée informatique axée sur le traitement cognitif du discours née dans le cadre des recherches en Intelligence Artificielle des années 70. Cette théorie tente d'expliquer la cohérence textuelle en dégageant une structure intentionnelle et attentionnelle et un modèle de continuité thématique. Elle se présente comme « une théorie qui établit des liens entre des focus d'attention, des choix d'expressions référentielles et des relations de cohérence à l'intérieur d'un même segment de discours » (Grosz *et al.* 1995:204 traduit dans Walker 2000:52). Nous avons présenté les notions de structures intentionnelle et attentionnelle en III.1.2 et III.1.2.a. Nous nous intéressons ici aux degrés de transition entre énoncés.

La théorie du centrage, désormais le Centrage, modélise le discours autour de la notion de "**centre d'attention**" qui correspond aux entités présentes dans le « (sub) text world » en cours : les centres sont des entités sémantiques qui font partie du modèle de discours de chaque énoncé dans un segment de discours. Comme le suggère la terminologie adoptée, ces centres se situent au niveau de la structure attentionnelle. L'algorithme développé par le Centrage assigne à chaque énoncé une liste de centres anticipateurs (Ca), un centre préféré (Cp) et un centre rétroactif (Cr).

Chaque entité sémantique d'un énoncé est considérée comme un centre anticipateur (Ca), c'est-à-dire une entité pouvant faire l'objet d'une continuité dans l'énoncé suivant. Ces Ca sont classés selon leur probabilité à être continués. Cette probabilité n'est pas universelle et dépend, pour plusieurs langues dont le français, de la fonction syntaxique de l'entité et de sa position dans la phrase. Ainsi, une entité remplissant la fonction de sujet constitue l'entité qui a le plus de chance de constituer le centre rétroactif (Cr) de l'énoncé suivant[57]. Elle acquiert ainsi le statut de centre préféré Cp.

Selon la relation entre le Cp de l'énoncé (i) et le Cr de l'énoncé (i+1), le Centrage distingue quatre degrés de transition entre énoncés, ce que schématise le tableau suivant :

	$Cr(E_i)=Cr(E_{i-1})$ ou $Cr(E_{i-1})=[?]$	$Cr(E_i)\neq Cr(E_{i-1})$
$Cr(E_i)=Cp(E_i)$	continuation	Déplacement en douceur
$Cr(E_i)\neq Cp(E_i)$	rétention	Déplacement brutal

Tableau III.2 : Les quatre degrés de transition pour la Théorie du Centrage (Walker et al. 1998:6)

Nous avons donc des transitions qui répondent à chaque fois à deux contraintes : l'une portant sur le lien entre le Cr et le Cp, l'autre portant sur le lien entre le Cr actuel et le Cr précédent. La grande faiblesse de cette théorie est qu'elle limite le discours précédent à la phrase précédente. Or, il est fréquent d'observer une continuité référentielle 'sauter' une phrase, comme dans l'exemple III.1.

En partant de l'exemple III.1, nous avons indiqué après l'extrait les Ca, le Cr de chaque phrase ainsi que le type de transition. Les différents Ca sont classés par ordre de préférence. Ainsi, le premier Ca équivaut au Cp de l'énoncé.

57 L'échelle des fonctions syntaxiques les plus propices à être Cr sont les suivantes : Sujet > Objet indirect animé > Objet direct > Objet indirect impliqué > Objet oblique.

(III.11) ***[1]*** *Malgré cette progression rapide et générale [de la proportion de bacheliers], les écarts restent sensibles entre les académies et les départements.***[2]** *En 1990, la proportion de bacheliers par classe d'âge n'atteint pas 40% dans les académies de la grande couronne parisienne.***[3]** *Elle se situe entre 40 et 45% dans le Nord, en Lorraine et en Alsace, dans les académies de Toulouse, Limoges, Grenoble, Rennes ainsi qu'en Corse.***[4]** *En Île-de-France les écarts sont écrasants entre Paris (...) et la Seine-Saint-Denis (...)...* **[5]** *Dans l'ensemble, la moitié sud du pays continue à avoir de plus fortes proportions de bacheliers que la moitié nord, Bretagne et, à présent, Lorraine exceptées.***[6]** *Mais ces différences s'atténuent. [ATLAS_2]*

(1) *Ca = les écarts > les académies > les départements > cette progression... > la proportion des bacheliers ; Cr = ? (cette progression correspond à une description résumante du paragraphe précédent, voir l'exemple entier p.68)*

(2) *la proportion de bacheliers... > les académies de la grande couronne parisienne > la grande couronne parisienne > 1990 ; Cr = la proportion des bacheliers*

(3) *Elle > le Nord > la Lorraine > l'Alsace > les académies de Toulouse > les académies de Limoges > les académies de Grenoble > les académies de Rennes > la Corse ; Cr = Elle = la proportion des bacheliers*

(4) *les écarts > Paris > la Seine-Saint-Denis > l'île de France ; Cr = ? (les écarts correspond au Cp de la phrase (1))*

(5) *la moitié sud du pays > proportions de bacheliers > la moitié nord > la Bretagne > la Lorraine ; Cr = ? (la proportion de bacheliers correspond au Cp et au Cr des phrases (2) et (3))*

(6) *ces différences ; Cr = ? (ces différences correspond à une description résumante de tous les paragraphe)*

Comme le montre l'analyse de cet exemple, l'attribution d'un Cr directement lié à la phrase précédente n'est pas toujours possible. Des phénomènes comme l'encapsulation (utilisation de descriptions résumantes) ou l'élaboration qui entraîne une continuité 'sautant' un ou plusieurs énoncés ne rentrent pas dans les définitions de la théorie du centrage à ses premières heures. Walker (2000) propose d'intégrer le modèle en cache pour pouvoir adapter le Centrage à la structure globale du discours (voir la fin de la partie III.1.2.b). Nous pouvons alors identifier entre les énoncés (1) et (4) un déplacement en douceur 'global' et entre les énoncés (3) et (5) une rétention 'globale'. Cependant, les phénomènes d'encapsulation restent en marge des transitions identifiées – ainsi qu'aux études sur les continuités référentielles[58].

III.3.3. L'encadrement du discours

III.3.3.a) Définition

L'hypothèse linguistique de l'encadrement du discours (Charolles 1997) identifie des segments textuels particuliers, dits cadres de discours, blocs textuels homogènes du point de vue sémantique : certaines expressions, en général des adverbiaux, de par leur position spécifique à l'initiale (détachée) de la proposition, ont cette capacité d'étendre un critère d'interprétation au-delà de la phrase dans laquelle ils apparaissent. Les expressions présentant cette fonction discursive sont appelées introducteurs de cadres – IC.

Selon la nature du critère exprimé par l'IC, différents types de cadres peuvent être définis. À la suite de Charolles (1997), nous distinguons des univers de discours qui regroupent des propositions autour d'un même critère idéationnel (une circonstance, un thème) et des espaces de discours rassemblant des propositions appartenant au même argument énonciatif, à la même « étape » dans le déroulement du discours. À ces deux types de cadres, nous pouvons ajouter des « cadres modalisés » qui rassemblent des propositions sur lesquelles le locuteur porte un certain regard.

Les adverbiaux textuels qui introduisent des espaces de discours font l'objet de la partie V.3.3 et V.3.4, les adverbiaux circonstanciels pouvant jouer le rôle d'introducteur d'univers sont vus en partie V.4.1 et les adverbiaux modalisateurs caractéristiques des cadres modalisés sont présentés en V.5.1.

58 Quelques travaux s'y intéressent particulièrement : Conte 1996, Legallois 2006, Álvarez-de-Mon y Rego 2001. Ces constructions pourraient être fortement utiles pour une tâche de résumé automatique, mais c'est là un autre travail de thèse.

III.3.3.b) Portée cadrative et portée sémantique

Les IC sont associées à un phénomène de portée sémantique et/ou cadrative (Charolles & Vigier 2005) selon la relation entre le critère exprimé et les propositions contenues dans le cadre. Par sa portée sémantique, le critère exprimé par l'IC persiste sur plusieurs propositions. Par sa portée cadrative, l'IC délimite des blocs qui structurent le texte en procédant à un découpage du texte similaire au découpage en sections. L'idée d'une portée sémantique rejoint celle d'une portée véric onditionnelle. Ainsi, dans l'exemple III.12, l'avant-dernière phrase est sous la portée sémantique de l'IC « Concernant l'hôtellerie homologuée ». On peut affirmer que c'est seulement dans le domaine de l'hôtellerie homologuée que la proposition est vraie. On peut également concevoir que le thème de cette phrase est : la clientèle de l'hôtellerie homologuée.

(III.12) ***En hôtellerie de plein air,*** *les régions Bretagne et Pays-de-la-Loire comptent le plus de nuitées, plus du quart sont étrangères sauf en Ile-de-France où la clientèle est essentiellement étrangère, les enquêtes montrent que les français préfèrent se loger chez un ami ou dans la famille.*
Concernant l'hôtellerie homologuée, *l'Île-de-France avec Paris concentre le plus grand nombre de nuitées (52 millions), les autres régions comptent entre 2 et 6 millions de nuitées. La clientèle est plus internationale dans le Nord-Pas-de-Calais qu'en Bretagne et Pays-de-la-Loire.*
En général, *on peut distinguer trois types de régions [...].[ATLAS_1]*

Considérer les IC pour leur portée cadrative correspond à un intérêt qui ne se situe plus au niveau de la composante idéationnelle (représentation du contenu) mais au niveau de la composante textuelle (représentation de sa structuration). Du point de vue réalisation (la délimitation des fins de portée), portée sémantique et portée cadrative peuvent se confondre. C'est d'ailleurs généralement le cas. En conséquence, il est délicat de ne pas confondre l'identification des deux effets de portée. Cependant, ces deux phénomènes discursifs ne suivent pas la même définition.

Pour Charolles & Vigier (2005), la portée sémantique correspond à « l'influence à distance (i.e. au delà de sa phrase d'accueil) du trait sémantique spécifié par l'adverbial introducteur de cadre ». Cette capacité est liée à leur « usage fonctionnel, *i.e.* leur exploitation pour la répartition des informations textuelles au fur et à mesure du discours ». Selon ces auteurs, la portée sémantique est une conséquence de la portée cadrative des adverbiaux antéposés. Les exemples de la partie précédente ont d'ailleurs assez clairement montré que la portée sémantique pouvait être conséquence d'autres phénomènes tels que le détachement par construction clivée ou les processus de continuité par défaut. De plus, la portée sémantique n'étant qu'une conséquence du pouvoir cadratif des IC, la borne finale de l'une ne correspond pas toujours à la fin d'un cadre. Ainsi, toujours selon Charolles & Vigier (2005), on peut s'attendre à ce que certains segments appartenant à un même cadre ne soient pas à interpréter selon les conditions exprimées par l'IC.

Crompton (2006) qui cherche à comparer la portée des adverbiaux circonstanciels en position initiale vs. non-initale définit ces deux effets de portée en discernant deux hypothèses : (i) les adverbiaux sont placés en position initiale pour pouvoir étendre leur influence au-delà de leur phrase d'accueil ; et (ii) les adverbiaux sont placés en position initiale pour délimiter des segments de discours. Il apparaît clairement une relation entre (i) et l'effet de portée sémantique ; et (ii) et l'effet de portée cadrative. Cependant, à l'inverse de Charolles & Vigier (2005), Crompton suggère que (ii) découle de (i) :

> "I assume that Meaning (ii) is derivative from Meaning (ii) with the common factor of being under the scope of the same adverbial helping constitute discourse spans. As to investigate Meaning (ii) independently would require a fairly robust model of discourse structure, and I do not feel that such a model exists yet." (Crompton 2006:249)

De notre point de vue, (i) peut tout à fait découler de (ii). Pour prouver cette hypothèse, nous allons nous appuyer sur le fonctionnement de la structure du document et notamment sur le découpage d'un texte en sections. Nous espérons montrer que c'est la position initiale qui confère un statut particulier aux éléments qui s'y trouvent, statut relatif à la composante textuelle plus qu'idéationnelle[59].

Le découpage en sections constitue une illustration relativement claire de la distinction entre portée cadrative et portée sémantique. La portée cadrative correspond au fait que toutes les propositions contenues dans une même section sont référencées sous un même "index" exprimé par le titre. Notre appréhension et finalement notre interprétation de l'ensemble du segment seront guidés par cet index. En ce sens, les titres orientent notre interprétation des sections plus qu'ils n'expriment un élément du procès (qu'il s'agisse d'un titre purement textuel comme "introduction" ou porteur d'un trait sémantique comme "La place des cadres dans l'organisation du discours"). Ce phénomène d'indexation permet notamment de pouvoir renvoyer au contenu propositionnel de toute une section par l'expression de son titre (par exemple en notant : "comme nous l'avons présenté dans la partie implémentation").

Par contre, du point de vue de la portée sémantique des titres, on ne peut pas affirmer que toutes les propositions contenues dans une section établissent un lien sémantique avec le ou les référents exprimés dans le titre. Hồ-Đắc *et al.* (2004) ont d'ailleurs montrer que cette portée sémantique du titre peut varier d'une portée quasi-nulle à une portée équivalente à la toute la section. Cette variabilité de la fonction discursive des titres peut s'expliquer par de nombreux facteurs de nature très diverse tels que le genre textuel, la forme du titre, son niveau dans la titraille, etc. (voir Hồ-Đắc *et al.* 2004 et Jacques & Rebeyrolle 2006).

Nous pouvons retrouver certaines de ces caractéristiques au niveau de l'encadrement. Tout d'abord, le pouvoir cadratif des adverbiaux antéposés semble être fortement tributaire du style de texte considéré, *i.e.* du genre et de la visée discursive qui le caractérise (Le Draoulec & Péry-Woodley 2005) ou simplement de sa forme (longueur, découpage ou non en sections, etc.) Crompton (2006) montre que les adverbiaux en position initiale n'ont pas plus – voire moins – de portée sémantique que ceux en position intraprédicative. Cette étude en corpus va précisément à l'encontre de l'hypothèse de l'encadrement du discours. Cependant, il faut relativiser ces résultats. En effet, Crompton étudie uniquement la portée sémantique des adverbiaux, et cela dans des textes relativement courts (500 mots). Or il paraît moins utile de recourir à une organisation en cadres lorsque l'on construit un texte de moins d'une page[60] (voir I.4).

On peut également remarquer que le rôle cadratif prend toute son ampleur lorsque l'élément introducteur est considéré en tant que membre d'une structure plus globale : le titre dans une titraille, un cadre dans une organisation en cadres (une continuité texto-stratégique, voir partie suivante). Ce rapprochement entre fonctionnement discursif de l'encadrement et du découpage en sections n'est pas nouveau (voir II.2.2). Il semble en effet nécessaire de distinguer le rôle d'un cadre isolé, de celui d'un cadre imbriqué dans une série.

L'exemple suivant montre une organisation en cadres selon l'axe temporel qui répartit le contenu de cette section en quatre périodes (on voit bien ici le rôle organisationnel du titre).

(III.13) ***b. Quatre périodes*** [titre niveau 3]
Entre 1949 et 1970, *la production pétrolière américaine (brut et "condensats") est multipliée par deux ; dans le même*

59 Nous ne discréditons pas pour autant la portée sémantique qui relève, de notre point de vue, d'un tout autre fonctionnement. Nous y revenons dans notre conclusion.

60 Les tests psycholinguistiques mis en place pour mesurer l'hypothèse de l'encadrement dans les processus de compréhension se sont précisément heurtés à cette difficulté. En effet, les méthodes psycholinguiguistiques ne permettent pas de travailler avec des textes longs (mais sur des textes composés d'un maximum de quatre phrases simples). De fait, les résultats obtenus n'ont pas apporté à l'heure actuelle de résultats probants

temps, la part de la demande couverte par le pétrole importé passe de 10% à 23%. La croissance de la production intérieure [...]
***A partir de 1970,** toutes les formes d'investissement susceptibles d'augmenter les réserves de pétrole connaissent, aux États-Unis, des coûts fortement croissants. Les évènements de 1973 introduisent de nouveaux paramètres, en particulier réglementaires. L'explosion des prix du brut aurait dû favoriser un relatif redressement de la production intérieure et une baisse de la demande, donc une décroissance des importations. Mais les dispositions législatives prises pour soulager les raffineurs face à l'augmentation de leurs coûts d'approvisionnement (entitlements system) fonctionnent comme une subvention aux importations. [...]*
***Entre 1978 et 1985,** deux effets se conjuguent pour précipiter une chute des importations (Figure 10, Figure 11) :*
- La compétitivité marginale de la production intérieure se redresse. [...]
- L'ajustement de la demande, longtemps entravé par la réglementation des prix, s'effectue brutalement (-2 Mb/j entre 1979 et 1983).
En conséquence, les importations chutent sur cette période, tant en valeur absolue (-3,8 Mb/j) que relative (-16 points de part de marché).
***De 1985 à aujourd'hui,** la part du pétrole importé dans la couverture de la demande ne cesse d'augmenter. La production américaine baisse au rythme de 2% par an en moyenne. Cette baisse ralentit après 1990, [...]. [GEOPO_14]*

Dans cet exemple, les cadres permettent de procéder à une sorte de découpage en sections. Le découpage nécessaire à un niveau idéationnel est très local, ce qui peut expliquer le recours à un découpage en cadres plutôt qu'en sections titrées. L'usage de titres à un niveau si local aurait sans doute dérangé la lecture. Il semble en effet évident que ce qui détermine l'homogénéité des cadres est davantage de l'ordre de la composante textuelle que de la composante idéationnelle. Le recours à une organisation en cadres permet de répartir les contenus en quatre segments. Pour que cette segmentation soit effective, la position des IC est primordiale. En effet, tout comme un titre dénué de mise en forme matérielle perd de son pouvoir structurant, un IC en position intraprédicative perd de son pouvoir cadratif. Les effets de portée sémantique, eux, peuvent persister par un phénomène de continuité référentielle implicite.

III.3.4. Les TSC – Text-Strategic Continuities

Nous avons déjà cité la notion de TSC et le travail de Virtanen (1992, 2004) qui constitue l'une des sources importantes de notre travail. Nous nous inspirons également de la conception de Virtanen concernant la position initiale (chapitre IV).

Le travail de Virtanen a pour point de départ une question : qu'est-ce qui gère le positionnement des adverbiaux circonstanciels dans la phrase ? Se basant elle-même sur les travaux d'Enkvist – que nous présentons également dans le chapitre suivant –, elle aboutit à un modèle d'organisation du discours qui joue tant à un niveau local que global. Comme Enkvist, Virtanen conçoit le texte comme le résultat d'un équilibre entre des forces différentes, *i.e.* des modes organisationnels différents. Les TSC sont la correspondance textuelle des différentes stratégies textuelles que met en oeuvre le locuteur pour construire son texte. Une des stratégies basiques consiste à connecter les phrases les unes aux autres par une progression thématique. Dans un texte où les thèmes sont des topiques (tels que les textes expositifs), cette stratégie consiste à établir des continuités texto-stratégiques orientées topique. Dans un texte où les thèmes sont des acteurs (tels que les textes narratifs), il s'agit de TSC orientées participant. Ces stratégies basiques peuvent 'cohabiter' avec d'autres modes organisationnels, *i.e.* d'autres TSC. Le temps et l'espace peuvent être un moyen d'organiser son texte. Ainsi, dans l'exemple suivant, l'auteur a choisi d'organiser ses propos autour d'une chronologie temporelle, il a opté pour une TSC temporelle. Cette TSC temporelle s'associe à une TSC orientée topique, le topique principal concernant la part de pétrole importé par les États-Unis.

(III.13) ***b. Quatre périodes*** [titre niveau 3]
***Entre 1949 et 1970,** la production pétrolière américaine (brut et "condensats") est multipliée par deux ; dans le même temps, la part de la demande couverte par le pétrole importé passe de 10% à 23%. La croissance de la production*

intérieure [...]
***A partir de 1970,** toutes les formes d'investissement susceptibles d'augmenter les réserves de pétrole connaissent, aux États-Unis, des coûts fortement croissants. Les évènements de 1973 introduisent de nouveaux paramètres, en particulier réglementaires. L'explosion des prix du brut aurait dû favoriser un relatif redressement de la production intérieure et une baisse de la demande, donc une décroissance des importations. Mais les dispositions législatives prises pour soulager les raffineurs face à l'augmentation de leurs coûts d'approvisionnement (entitlements system) fonctionnent comme une subvention aux importations. [...]*
***Entre 1978 et 1985,** deux effets se conjuguent pour précipiter une chute des importations (Figure 10, Figure 11) :*
- La compétitivité marginale de la production intérieure se redresse. [...]
- L'ajustement de la demande, longtemps entravé par la réglementation des prix, s'effectue brutalement (-2 Mb/j entre 1979 et 1983).
En conséquence, les importations chutent sur cette période, tant en valeur absolue (-3,8 Mb/j) que relative (-16 points de part de marché).
***De 1985 à aujourd'hui,** la part du pétrole importé dans la couverture de la demande ne cesse d'augmenter. La production américaine baisse au rythme de 2% par an en moyenne. Cette baisse ralentit après 1990, [...]. [GEOPO_14]*

Les TSC temporelles tout comme les TSC spatiales se construisent essentiellement par séquences d'adverbiaux détachés en initiale qui, vraisemblablement, introduisent des cadres. Ainsi, une TSC temporelle correspond à une succession de cadres temporels. Nous nous trouvons ici au coeur de la séquentialité du discours. Ce type de TSC assure à la fois une cohésion interne au cadre, un déplacement marqué par les adverbiaux et une continuation textuelle entre cadres.

> "Sentence-initial adverbials of time and place in narrative and descriptive texts tend to form chains of text-strategic markers which have two basic functions in the discourse. They help to create coherence and at the same time they signal text segmentation." (Virtanen 2004:82)

Cette conception de l'organisation des textes se situe à un niveau très global. Il s'agit de considérer le texte comme un objet continu, relativement à une même stratégie de communication : "*Text-strategic continuity* (TSC) may thus be defined as a thematic or topical[61] uniform text-structuring orientation chosen to attain, in view of the communicative goal, a maximally profitable text organization, for the benefit of the text receiver." (Virtanen 1992:85)

Nous sommes relativement loin des considérations de modélisation des articulations rhétoriques ou des continuations idéationnelles. Le modèle (ou plutôt la théorie) de Virtanen se situe pleinement au niveau de la composante textuelle. La cohérence d'un texte n'existe pas seulement au niveau de ce qu'il raconte ou de l'intention qui a poussé le locuteur à produire ce texte, elle se situe également au niveau de son mode de construction.

> " … it is the text that is continuous, not necessarily the referential frame. Time, place or a member of a group of participants may change but the chain of references to a common temporal, spatial or participant-oriented frame still has the effect of forming continuity in the text and discourse" (Virtanen 1992:89)

Les textes suivent généralement plusieurs stratégies qui se succèdent ou co-occurrent selon des relations hiérarchisées. Virtanen distingue des TSC locales et des TSC globales. Les TSC locales, observées à un niveau intraparagraphique, sont généralement dominées par des TSC plus globales, allant de l'organisation du paragraphe à celle de la section ou du texte entier.

> "The scope of a text-strategic continuity may vary from a sentence or two to the entire text. […] *global* continuities frame the text by steering it on a higher level than *local* strategies, which are found, for instance, within a paragraph or across a few sentences. […]
> The span of a TSC may vary from a global textual chain covering any length of text to arches whose scope is limited to just few sentences, or even clauses. One or more local strategies can thus be embedded into the main or global strategy." (Virtanen 1992:121)

61 in the sense of Givòn's topic continuity (Givòn 1983)

La TSC temporelle illustrée par l'exemple III.13 est une TSC globale marquée par des adverbiaux temporels positionnés en initiale de paragraphes, tandis que la TSC locale est une stratégie intra-paragraphique, telle qu'observée dans l'exemple III.14. Cet extrait se situe dans un portrait dont l'organisation globale se construit autour d'une TSC orientée selon la personne de Bartók. Cette TSC orientée participant 'domine' la TSC temporelle locale observée ici.

(III.14) *La première tournée aux États-Unis se déroule de décembre 1927 à février 1928.* ***Pendant deux mois,*** *Bartók parcourt le pays, faisant des conférences sur la musique populaire hongroise et sa place dans la musique savante, illustrée au piano par lui-même. Il donne des récitals de musique de chambre avec Jelly Arányi et József Szigeti, joue avec orchestre sous la direction de Willem Mengelberg, Fritz Reiner et Serge Koussevitsky.* ***En janvier 1929,*** *il séjourne en U.R.S.S. et Ø donne des récitals à Kharkov, Odessa, Leningrad et Moscou.* ***En 1931,*** *il termine son Concerto pour piano et orchestre no 2, Ø compose les Quarante-Quatre Duos pour violons sur des mélodies populaires et orchestre quelques-unes de ses pièces pour piano. Cette année est particulièrement bien remplie : Bartók a cinquante ans et Ø reçoit la Légion d'honneur des mains de Louis de Vienne, ministre plénipotentiaire, chef de la légation de France à Budapest.* ***Au début de l'été,*** *il est à Genève, à la session de la Commission de la coopération intellectuelle de la Société des Nations, avec Carel Capek, Gilbert Murray, Thomas Mann, Paul Valéry, etc.* ***L'année suivante,*** *il compose Six Chants sicules pour chœur d'hommes, puis il termine,* ***en 1934****, son Quatuor à cordes no 5.* ***À partir de cette année,*** *il est affecté à l'Académie hongroise des sciences pour la préparation de l'édition systématique de la musique populaire hongroise. [PEOPL_15]*

Il semble assez naturel que des portraits soient structurés à la fois autour de TSC orientées participants et de TSC temporelles. Nos analyses confirmeront cette intuition langagière. Notre thèse pose au centre de sa méthodologie la variation de mode de construction entre différents types de texte. La notion de variation textuelle est depuis longtemps prise en compte dans l'étude du lexique et particulièrement avec les travaux de Biber. Nous soutenons l'hypothèse que l'organisation des textes (*i.e.* des différentes stratégies utilisées par son auteur) varie selon le type de texte et principalement selon son domaine d'appartenance et sa visée discursive (voir VI.2.3). Notre point d'accès pour observer l'organisation de différents types de texte se situe en position initiale. Et comme Virtanen, nous donnons un poids différent aux positions initiales de phrases, de paragraphes et de sections.

Partie 2.

Problématique

« Figure »

Chapitre IV
Une position stratégique : l'initiale

Sommaire

Cette thèse propose d'étudier le rôle de la position initiale dans l'organisation du discours. Le choix d'aborder l'organisation du discours par ce biais se justifie essentiellement par l'impact cognitif de cette position qui donne aux éléments qui s'y trouvent un double statut : celui de point de départ et celui de point d'ancrage. Nous utilisons délibérément deux termes vagues pour définir notre intérêt pour la position initiale. Ces termes présentent l'avantage de pouvoir être utilisés ici dans le même sens que dans la langue courante. Le point de départ est l'endroit à partir duquel se déroule, se développe, se construit quelque chose et le point d'ancrage correspond à l'endroit où se situe l'accroche, le lien avec d'autres éléments. Au niveau de la construction du texte : la position initiale est un pivot puisqu'elle permet dans un 'regard avant' de commencer un texte, un paragraphe, une phrase ; et dans un 'regard arrière' de relier ce qui va se dire à ce qui s'est dit. Elle est de ce fait une position privilégiée pour les indices de cohésion et les marques de séquentialité.

> "Cohesive ties can occur anywhere in a sentence. However, as several studies have shown, the initial part of a sentence lends itself easily to cohesive functions." (Enkvist 1976:65, Virtanen 1992:185)

La position initiale est définie en Systémique Fonctionnelle par la fonction sémantico-pragmatique de Thème[62] :

> "The Theme is a function in a clause as message. It is what the message is concerned with : the point of departure for what the speaker is going to say" (Halliday 1985:38), et "the peg on which the message on which the message is hung" (Halliday 1970:161).

62 Nous utilisons la forme « Thème » avec majuscule afin de distinguer notre acception de la fonction Thème selon la terminologie de la Systémique Fonctionnelle (présentée dans la partie V.3) de l'acception courante liée à la notion de thématique telle que utilisée dans les calculs de segmentation thématique.

La position Thème s'associe à la fonction d'orientation qui consiste à spécifier les ensembles de connaissances à activer pour comprendre les propos tenus.

> "[elements in initial position] limit the domain of applicability of the main predication to a certain restricted domain […] set[ting] the spatial, temporal or individual framework within which the predication holds" (Chafe 1976:53)

Les éléments initiaux ancrent l'unité qu'ils inaugurent dans le discours précédemment interprété tout en indexant, en orientant, en guidant l'interprétation de la suite du discours ; et ceci que ce soit au niveau local (les phrases) ou global (les paragraphes, les sections, le texte entier). Cette fonction d'orientation que peuvent acquérir les éléments positionnés en initiale (les éléments Thème) est une conséquence directe du caractère linéaire des textes. Car plus on avance dans l'unité, plus les possibilités d'interprétation se réduisent comme l'explique Bolinger cité par Enkvist (1989:175).

> "Let us consider what happens when elements – call them words, for convenience – are laid end-to-end to form a phrase. Before the speaker begins, the possibilities of what he will communicate are practically infinite, or, if his utterance is bound within a discourse, they are at least enormously large. When the first word appears, the possibilities are vastly reduced, but that first word has, in communicative value for the hearer, its fullest possible semantic range. The second word follows, narrowing the range, the third comes to narrow it still further, and finally the end is reached at which point the sentence presumably focuses on an event." (Bolinger 1952/1972:32)

Selon la conception traditionnelle de la structure d'information (Bolinger 1952, *cf.* Östman & Virtanen 1999), le Thème est le point d'entrée du message. Il offre un grand potentiel d'incertitude*, i.e.* les informations qui vont suivre ce ou ces éléments initiaux sont imprévisibles. Plus on avance dans le discours, plus cette incertitude est réduite. On peut ainsi représenter l'évolution de l'information potentielle par la figure IV.1 où l'on voit le Thème – en tant que point d'entrée – ouvrir sur une multitude de possibilités et de choix pour la suite du discours, et le Rhème réduire progressivement cette multiplicité.

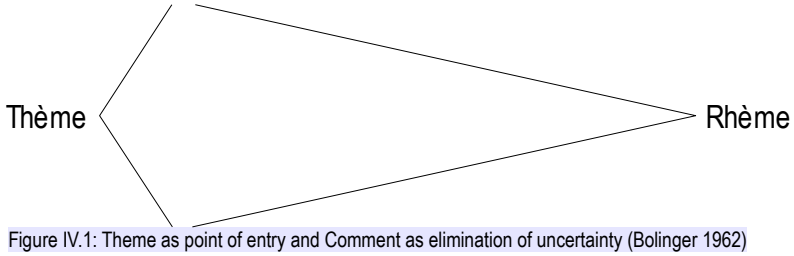


Figure IV.1: Theme as point of entry and Comment as elimination of uncertainty (Bolinger 1962)

> "The entry point chosen opens up a particular direction to continue in – from left to right in the traditional Bolingerian manner of graphically representing Theme-Rheme progression. The entry point opens up a multitude of possibilities and choices at the next stage of text processing. [...] Potential uncertainty – to the left – is gradually eliminated as the utterance moves 'rightwards' in accordance with the principle of End-weight in English." (Östman & Virtanen 1999:97,101)

Dans cette vision, chaque élément apporte un critère d'interprétation qui s'ajoute aux autres. Il est donc logique que plus on se trouve vers la position initiale, plus l'élément s'y trouvant indexe l'ensemble de l'unité, étant le premier mentionné et donc prenant sous sa portée les autres éléments à venir. Ainsi, dans l'unité paragraphe constituant l'exemple IV.1, le premier élément qui définit la localisation spatiale de l'événement réduit les possibilités d'interprétation à la ville de Rouen, puis le deuxième permet une réduction à l'année 1640, puis le troisième à la personne de Pascal. Ces éléments en position Thème constituent les fondements du sous-monde construit ici. La suite

du paragraphe précise quels sont les procès que l'auteur raconte (parmi tous les procès possibles impliquant les circonstances *Rouen* et *1640* et le participant *Pascal*).

(IV.1) À Rouen, en 1640, Pascal envisagea de construire une machine effectuant les quatre opérations arithmétiques élémentaires. Son objectif était de faciliter les pénibles opérations comptables dont son père avait la charge.[PEOPL_4]

Les éléments Thèmes ont donc cette capacité à orienter la suite du discours. Celle-ci leur permet de marquer des unités telles que les cadres de discours ou les continuités texto-stratégiques (voir III.3.3 et III.3.4). Du point de vue de la production d'un texte, la question est essentiellement de savoir pourquoi l'auteur a choisi de mettre tel élément en point de départ. Du point de vue interprétation, il s'agit d'étudier le statut attribué aux éléments en position initiale, l'influence des éléments Thèmes sur le reste du texte.

IV.1. Positionner les informations dans le discours

> "Since language is a linear phenomenon and experience need not be, the linguistic coding of experience often involves a fairly strict (re-)organisation of information units. Relations which may be of a hierarchic nature have to be represented by means of a linear structure. The order of elements in such linear structure is not arbitrary ; on the contrary, word order is crucial both in the coding and in the interpretation of a message." (Hasselgård 1996:63)

La nécessité d'ordonner les informations contenues dans le texte a déjà été mentionnée plusieurs fois dans cette thèse, notamment au sujet de l'obligation textuelle de linéariser/délinéariser et du phénomène de séquentialité (chapitre III). Cet ordonnancement des informations répond à plusieurs choix de répartition comme celui de la répartition entre l'information donnée *vs.* nouvelle, entre le référent de base *vs.* son développement (topique/propos), entre l'information même et le contexte pertinent pour cette information (figure/fond).

La position initiale correspond par nature au point de départ de l'unité dont elle est l'initiale. Pour comprendre ce qui peut raisonnablement commencer une unité, nous nous appuyons essentiellement sur les travaux de Nils Erik Enkvist (et de Tuija Virtanen qui poursuit ces travaux) qui a largement consacré ses études à l'ordonnancement des mots, notamment le positionnement à l'initiale des adverbiaux. Ces travaux se situent entre la linguistique descriptive et la linguistique cognitive. La position initiale répond effectivement à une conception très cognitive : ce ne sont pas des règles syntaxiques qui influencent le choix du 'bon' point de départ, mais des considérations cognitives sur l'importance du rôle de tel ou tel élément dans la construction du *text-world*. Par défaut, selon un principe cognitif humain basique, la position initiale est saillante, ce que défend Talmy Givón.

> "Givón explains the fact that in most languages the first position of the clause has a special function by a psychological feature of humans: it attracts more attention from the addressee. Givón states that the string-initial position invites the hearer to pay more attention, and thus to store and retrieve the information more efficiently." (Givón 1988:276 cité dans Boland 2006:17)

Selon Enkvist dans un article intitulé *A parametic view of word-order* (1985), il s'agit de trouver un compromis entre ces différentes répartitions qui peuvent être assimilées lors du processus de production à des forces allant dans des directions plus ou moins différentes. 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.

> "The word order of a sentence was the outcome of a conspiracy or a struggle between the different forces that affect the linearization of discourse. [...] A text is a resultant of various forces, some of which conspire towards the same end and some of which are antagonist to one another" (Enkvist 1985:321-322)

> "According to this principle [the CIF principle] initial position is often preferred for elements which convey information crucial for the understanding of the macrostructure of the discourse. Such elements can be markers of (a changed) setting" (Hasselgård 1996:62)

Toujours selon Enkvist, deux principes (ou forces) peuvent expliquer le choix du point de départ : soit on positionne en initiale l'information donnée, soit on positionne en initiale l'information la plus importante pour l'interprétation (Enkvist parle d'information cruciale). D'un côté, nous trouvons la vision phrastique de la Functional Sentence Perspective (FSP – Firbas 1959, 1964, 1992) qui développe sa théorie autour de la distinction information donnée/nouveau, et pour laquelle la structure phrastique non marquée consiste à placer l'information donnée (qui est la plus prévisible et donc la moins lourde à traiter) avant l'information nouvelle (qui est la moins prévisible et donc la plus lourde à traiter). De l'autre, nous trouvons une conception plus discursive qui consiste à associer à la position initiale l'information jugée la plus importante pour la bonne interprétation du texte. Enkvist (1989) parle alors du principe de l'information cruciale en premier (**CIF** : *Crucial Information first*) qu'il distingue du principe de l'information donnée en premier (**OIF** : *Old Information First*) :

Ces deux principes ne se situent pas seulement au niveau des processus de production. Le lecteur s'attend à ce que les éléments en position initiale n'y soient pas par hasard. Selon Givón, les éléments en position initiale sont lus et enregistrés comme des éléments qui 'doivent continuer' dans le discours (les thèmes ou topiques de discours) ou qui sont importants et 'inattendus'.

> "The element in string-initial position is either one which is able to persist in the discourse (theme or topic), or one which is highly unpredictable from the preceding discourse" (Givón 1988:259)

Gernsbacher (1990) accorde un rôle tout à fait particulier aux éléments exprimés en premiers. Les éléments initiaux ont un avantage (*the Advantage of first mention*) qui se traduit par une saillance plus forte que les éléments non initiaux[63]. Les éléments en initial revêtent une importance capitale en tant que fondations pour la construction d'une nouvelle structure ; en parallèle, la position initiale est particulièrement bien adaptée à contenir l'information donnée, celle là même qui permet d'intégrer les informations nouvelles à la structure en cours de construction. Il y a bien deux fonctions discursives à associer à la position initiale : les processus de déplacement par construction de nouvelles fondations (« *laying a foundation* ») et les processus de continuation (« *mapping subsequent information onto that foundation* »).

> "If first mention is selected in order to signal importance, then the function is accomplished because – by virtue of being first mentioned – initial information gets represented at the core or foundation of the structure.[…] On the other hand, if first mention is selected in order to signal giveness, then the function is also accomplished because – by virtue of being first mentioned – initial information organizes the representation of subsequent information. That is, subsequent information is mapped onto the developing structure vis a vis the initial information." (Gernsbacher 1990:47)

Dans la même conception que Gernsbacher, nous pensons que le principe CIF est également appliqué lorsque c'est l'information donnée qui est mise en premier. En effet, ce qui est généralement le plus important (cognitivement parlant), c'est de relier les énoncés entre eux. L'une des stratégies les plus économiques pour cela est de placer en premier l'information donnée, ou en d'autres termes, l'information commune aux énoncés précédents. Dans ces cas, l'information cruciale correspond à l'information donnée. Dans d'autres cas, ce qui est important, c'est d'indiquer qu'il y a une discontinuité. Nous avons alors en position initiale des éléments fondateurs (des *world-builders* cf.) Dans ce cas

63 Gernsbacher (1990) appuie son principe de l'A*dvantage of first mention* sur des expériences psycholinguistiques réalisée pour mesurer l'accessibilité des éléments exprimés en initiale *vs.* ceux exprimés en position non-initiale.

là, le CIF permet d'expliquer la position initiale de nombreux adverbiaux spatiaux et temporels (Enkvist 1985, Virtanen 1992 et 2004, Hasselgård 1996).

> "Besides being the unmarked locus of given information (e.g. Quirk *et al.* 1985: 1361), sentence-initial position is a place where discourse relations can be marked explicitly, being the obligatory position for conjunctions and the most frequent position of conjunct adverbials (Biber *et al.* 1999: 772)" (Hasselgård 2004b:65)

Nous retrouvons dans cette double fonction la dualité de la position initiale : la position initiale est tournée conjointement vers l'arrière (ce qui répond au principe OIF, à la fonction de connexion, aux processus de continuation) et vers l'avant (ce qui répond au principe CIF, à la fonction d'orientation, aux processus de déplacement).

IV.2. La fonction d'orientation

Notre conception de la fonction d'orientation est largement inspirée des travaux de Wallace Chafe (1976, 1994, voir aussi l'article de Allwood 1996 qui offre un regard général sur les travaux de Chafe) qui cherchent à mettre en évidence les liens entre les phénomènes linguistiques et la nature de notre conscience et de notre esprit.

> "I have a special interest in understanding how both the flow and the displacement of conscious experience affect the shape of language, and conversely, how language can help us better understanding these basic aspects of our mental lives." (Chafe 1994:4)

Selon Chafe, l'orientation constitue une propriété immuable de notre conscience ("*Consciousness has a need for Orientation*"). Elle consiste à orienter le *moi* dans le monde. La fonction d'orientation constitue une sorte de complément au principe gestaltique du Figure-Fond qui explique que la figure (le *moi*, le topique) n'existe que par son inscription sur un fond (voir III.1.2.b). Orienter un objet, c'est non seulement le situer dans son contexte mais surtout l'associer aux axes dans lesquels il s'inscrit. Comme l'illustre Chafe (1994:83), les textes ne sont pas composés de phrases indépendantes 'flottant' à sa surface comme une multitude bulles déconnectés.

> "Clauses do not express a random collection of independent events or states, floating in the air like so many disconnected bubbles. Rather, each has a point of departure, a referent from which it moves on to provide its own contribution." (Chafe 1994:83)

Les phrases – et dans notre thèse, les paragraphes, les sections, le texte dans son entier – ont un point de départ. Cet élément initial est le point d'entrée de l'unité, celui qui oriente son contenu en précisant le temps, l'espace, le domaine d'activité, le point de vue, etc. en rapport desquels les propos tenus sont pertinents. On le voit, la fonction d'orientation s'associe parfaitement à notre notion de *setting* (II.1.3).

> "Starting points can manifest lexical overlap with the preceding text and they typically convey given or inferable information. They help suggest common ground for what follows. They can indicate the setting, a source, or a restriction of some kind, for what follows, or a perspective that the reader is expected to adopt in processing the text. Such elements are not necessarily spatio-temporal though they can be abstractions from these. The fact that they are placed initially in the sentence invites the reader to interpret them as background for what follows." (Virtanen 2004:91)

Pour Virtanen (2004), le lecteur s'attend à trouver en position initiale soit des indices de connexion (l'information donnée qui permet la construction d'une continuation idéationnelle), soit des indices d'orientation qui vont guider la construction de son *text-world*. Ces deux fonctions peuvent théoriquement et très schématiquement être associées aux deux types d'éléments que l'on peut trouver en position initiale : des éléments détachés extra-prédicatifs et des éléments intégrés (la partie IV.4.1 est précisément sur cette distinction entre Thème topical et Thème scénique). C'est

pourquoi on retrouve la fonction d'orientation à un niveau plus linguistique pour expliquer le rôle des constituants extra-prédicatifs en position initiale.

En Grammaire Fonctionnelle (Dik 1997), trois fonctions d'orientation sont distinguées selon la nature de leur influence sur le reste de la phrase (la grammaire de Dik reste à un niveau phrastique) :

- Les conditions véridictionnelles
- Les *settings* (situation discursive : temporelle / spatiale / autre)
- Les *Themes*

Ces trois sous-fonctions se rejoignent dans le sens où elles situent le monde à prendre en compte pour interpréter l'information nouvelle. Les **conditions véridictionnelles** permettent de créer un monde dans lequel l'information est à considérer comme vraie. On y retrouve la plupart du temps des subordonnées de condition qui indiquent que si la condition est vérifiée alors l'information qui suit sera vraie, sinon, cette information ne la sera pas.

If John is rich, then he can help us out. (Dik 1997:396)

D'après les exemples donnés par Dik, il semble que la conception des conditions véridictionnelles en Grammaire Fonctionnelle soit très proche de celle des mondes hypothétiques :

> "Conditional clauses may be used by S [the speaker] to 'create' a world or 'mental model' within which that which is expressed in the apodosis clause is claimed to be relevant or true." (Dik 1997:396).

Il est généralement difficile d'étendre la portée des conditions véridictionnelles au delà de la phrase. Cependant, nous observons certains cas où des conditions véridictionnelles découpent des segments de texte, comme en (IV.2)[64].

(IV.2) Les incertitudes sont donc très importantes. [...] Les prévisions de prix varient mais n'anticipent pas d'augmentation significative, en termes réels, sur les vingt prochaines années. ***Si les États du Moyen-Orient n'effectuaient pas les investissements requis*** *(rappelons que les capacités de production dans le Golfe n'ont pas augmenté depuis 30 ans), le prix du pétrole pourrait être nettement plus élevé, la demande plus faible, et la production "hors Golfe" plus soutenue. Les importations américaines seraient alors plus faibles que ne l'anticipent les modèles et nettement moins concentrées sur le Moyen-Orient.* ***A l'inverse, si le processus concurrentiel s'enclenchait entre producteurs à coûts de production très bas****, le prix s'effondrerait, stimulant la demande et déprimant la production "hors Golfe". Les importations américaines seraient encore plus fortes qu'escompté, ainsi que la part du Moyen-Orient dans les approvisionnements extérieurs.[GEOPO_14]*

Cet exemple montre bien comment les conditions de Dik correspondent à la construction de mondes hypothétiques construits temporairement pour élaborer une argumentation.

Les **Settings**, expressions initiales référentielles, ne réduisent pas spécialement le *text-world*, mais précisent les circonstances dans lesquelles les faits relatés se sont déroulés (à quel moment, pendant combien de temps, dans quel lieu, avec qui...).

In the beginning of spring, John felt awful. [Dik 1997:397]

La différence entre conditions véridictionnelles et Settings est floue. D'ailleurs, Charolles définit la portée des introducteurs d'univers de discours comme étant une portée véridictionnelle. Mais nous pensons que la distinction entre ces deux types d'orientation est nécessaire. D'un côté, nous avons des expressions qui réduisent 'matériellement' le monde, de l'autre, nous avons des expressions qui canalisent certaines connaissances d'arrière-plan afin d'orienter la compréhension. Cette différence apparaît plus clairement à la lecture d'exemples. L'exemple IV.3 montre un univers de discours introduit par un adverbial temporel qui pose un Setting : l'année 1993 (un Setting global est posé par le titre de ce texte : *l'Ouest de la France*).

(IV.3) ***En 1993****, un tiers des électeurs ne se déplacent pas non plus. Il faut y ajouter un nombre important de blancs et nuls (1 400 000). L'abstention devient une dimension forte et durable de la vie politique française. Elle est encore plus*

64 Cet exemple est probablement le seul de notre corpus.

significative lors de ces législatives où l'électorat désigne son assemblée cinq années après le début du septennat. Le doute et la perplexité très présents depuis plusieurs années, manifestes lors des régionales de 1992, marquent encore le vote de 1993. [ATLAS_3]

Alors qu'on peut dire sans difficulté que dans (IV.2) les propos sont irrecevables en dehors des conditions explicitées, il est plus délicat de l'affirmer pour (IV.3). Le débat n'est pas de savoir si c'est l'emploi de telle construction syntaxique ou de tel élément détaché qui est marque la différence entre conditions et *settings*. Le fait est que l'on ne peut pas dire qu'en dehors de l'année 1993, le fait qu'un tiers des électeurs ne se déplacent pas est faux, de même pour les propos des phrases suivantes. Cette portée relativement insaisissable des *settings* leur permet d'orienter des pans de discours entiers. La partie III.3.3.b a justement précisé cette distinction entre d'une part la portée sémantique d'un introducteur de cadre, et d'autre part sa portée cadrative.

Les *settings* correspondent à des expressions temporelles, locatives, des spécifications de domaine, de source d'énonciation, etc. Ces expressions sont aussi appelées des "scene-setting topics" par Lambrecht (1994:118) ou encore "chinese-style topics" par Chafe qui les définit comme des éléments qui établissent un cadre spatial, temporel ou spécifique à l'intérieur duquel la prédication principale se déroule ("*a spatial, temporal or individual framework within which the main predication holds"*, Chafe 1976:53). Le fait que dans la terminologie de Lambrecht comme dans celle de Chafe les Settings soient associées à la notion de topique souligne le lien qui existe entre la position initiale et l'expression topique. Nous reviendrons sur cette confusion dans la partie IV.3 consacrée à la notion de Thème. Nous pouvons d'ores et déjà noter que le paragraphe de l'exemple n'est pas construit à propos de l'année 1993. Comme le souligne Charolles (2003:36) « les SP [introducteurs de cadre, comme ici *en 1993*] sont instrumentalisés pour la répartition des informations et l'organisation du discours, ils ne constituent pas, à proprement parler, le topique du texte, ils servent simplement, en coulisse, à ordonner les contenus communiqués [...] ».

L'hypothèse de l'encadrement du discours s'appuie totalement sur cette fonction d'orientation par Setting, notamment dans la définition des univers de discours (voir III.3.3). Parmi les différents types de cadres, Charolles (1997) distingue les univers de discours qui regroupent des phrases dont les propos se situent dans les mêmes Settings et les cadres thématiques qui regroupent des phrases portant sur le même thème, thème explicité par l'introducteur de cadre. Cette définition rejoint la fonction d'orientation par Thème selon le modèle de Dik. Pour la grammaire fonctionnelle, **le Thème** indique au lecteur comment ancrer l'information dans le discours en cours, en spécifiant parmi un ensemble d'entités potentielles celle à propos de laquelle la proposition associée apporte une information pertinente[65].

As for Paris, the Eiffel Tower is really spectacular. (Dik 1997:390)

Les constructions avec Thème sont assez proches des « *As-for constructions* » de Lambrecht en structure informationnelle de l'énoncé. Ces constructions utilisent, en anglais, la locution prépositionnelle *As for* pour introduire le thème de la phrase, tout en permettant à la proposition d'être à propos d'une autre entité. En français, Porhiel (2001, 2003) propose plusieurs études sur les cadres thématiques, notamment dans le projet REGAL qui vise la création d'une plateforme pour la gestion des ressources textuelles[66]. Les introducteurs de cadre thématique sont particulièrement intéressants pour repérer le thème d'un segment de texte, puisque leur fonction même est d'expliciter le thème d'un passage. Cependant, il faut garder à l'esprit la grande rareté de telles constructions.

65 Nous rappelons que cette fonction n'est associée qu'à des éléments détachés en initiale. Les sujets grammaticaux (qui réalisent généralement le Thème du point de vue de la Systémique Fonctionnelle ou des différents modèles de structure informationnelle) ne remplissent pas, pour Dik, la fonction de Thème.

66 REGAL : Répartition et gestion d'applications à large échelle (http://www.inria.fr/recherche/equipes/regal.fr.html)

La fonction d'Orientation est plus cognitive que linguistique. Il n'y a pas, à proprement parler, de fonction linguistique d'Orientation. En revenant sur les traces de Chafe, il semble que la fonction d'orientation est une fonction cognitive de base qui a une certaine influence sur la forme de nos productions langagières. De notre point de vue, la fonction d'Orientation n'est pas associée à un statut informationnel particulier. Elle n'est pas non plus associée à un contenu sémantique particulier. La fonction d'Orientation est un phénomène psychologique qui pousse le lecteur à interpréter des informations en les situant dans un contexte, en les décrivant relativement à un fond, un arrière-plan de connaissance. Ce même phénomène invite le locuteur à peindre le fond avant de décrire les faits qu'il veut y faire figurer. La fonction d'Orientation, du point de vue de la linéarité du texte, est évidemment la fonction de la position initiale.

IV.3. La notion de Thème

La position initiale correspond en SF à la notion de Thème. Bien que nous ayons déjà parlé de Thème au sens de Dik, pour présenter la fonction d'Orientation, nous ne conservons pas la définition dikienne du Thème et nous situons complètement dans le modèle de SF.

Notre présentation de la notion de Thème fait grandement référence au travail de María Ángeles Gómez-González (2001) et à l'ouvrage de Geoff Thompson (2004). Le travail de Gómez-González offre une étude descriptive très complète (basée sur corpus) des réalisations linguistiques du Thème. L'ouvrage de Thompson constitue une introduction à la SF.

Trois domaines d'acceptions de la notion de Thème sont à distinguer : une acception d'ordre sémantique (le Thème en tant que contenu), une autre plutôt informationnelle (le Thème en tant que statut informationnel) et une dernière syntaxique (le Thème en tant que position).

> "three dominant interpretations of communicative categories […]:
> **Semantic**, suggesting that Theme/Topic establishes a relationship of aboutness expressing "what a message is about.
> **Informational**, rendering Theme/Topic as given information ; and
> **Syntactic**, assuming that Theme/Topic constitutes a special point of departure that is associated with initial position, *i.e.* "from which the speaker proceeds" (Gómez-González 2001:9).

Ces trois niveaux de définition sont souvent confondus. Ainsi, la position Thème est souvent associée au statut de topique (ce à propos de quoi l'auteur parle, Lambrecht 1994), à l'information donnée *vs.* l'information nouvelle où l'information donnée correspond, selon la Functionnal Sentence Perspective (Firbas 1992) aux éléments ayant le plus faible dynamisme communicatif. Information topique, information donnée, information de faible dynamisme communicatif, il s'agit là plus des conséquences du statut cognitif de la position Thème que d'éléments pour sa définition.

IV.3.1. Définition

Dans notre étude comme dans la plupart des travaux en SF, le Thème correspond à la position initiale. Cette position, nous l'avons déjà dit, est importante d'un point de vue cognitivo-discursif pour son rôle dans l'organisation du discours : (i) au niveau de la production, le Thème est le point de départ du message tel que l'a choisi le locuteur ; (ii) au niveau de la compréhension, le Thème est ce qui va attirer l'attention et situer puis orienter le message dans le *text-world*.

Thompson (2004) montre particulièrement bien la nature du Thème en procédant à une substitution des premiers constituants de la phrase (nous l'avons traduite en français) :

(IV.4) *Pendant des siècles, les canaris jaunes ont été utilisés pour 'tester' l'air dans les mines.*

Cette phrase constitue la première phrase d'un article de journal consacré à une exposition sur l'histoire de l'industrie. Le fait qu'elle commence par un adverbial temporel est cohérent avec la thématique de l'article (l'histoire de l'industrie). Face aux phrases (IV.4[1]) et (IV.4[2]), l'impression est différente, et la mise en Thème des *canaris jaunes* ou des *mineurs*, peut être lue comme indiquant que l'article est à propos des canaris ou des mineurs, alors que cette phrase d'ouverture ne fait référence aux canaris et à l'exploitation minière qu'à titre d'exemple.

(IV.4[1]) *Les canaris jaunes ont été utilisés pour 'tester' l'air dans les mines pendant des siècles.*
(IV.4[2]) *Les mineurs de fond ont utilisé des canaris jaunes pour 'tester' l'air pendant des siècles.*
(IV.4[3]) *En exploitation minière, les canaris jaunes ont été utilisés pour 'tester' l'air pendant des siècles.*
(IV.4[4]) *Pour 'tester' l'air dans les mines, les canaris jaunes ont été utilisés pendant des siècles.*

La phrase (IV.4[3]), tout comme la phrase (IV.4[4]) réduisent considérablement le point d'entrée de l'article. Dans (IV.4[3]), on passe de l'histoire de l'industrie en général (thème général) à celle de l'exploitation minière, et dans (IV.4[4]), on réduit encore plus le champ de l'article, le lecteur, à la lecture de cette première phrase, s'attendant plus à lire un article technique qu'un article portant sur une exposition historique.

Pour chacun de ces exemples, nous avons commencé le message par un point différent, *i.e.* nous avons choisi un Thème différent. Cette différence de point de départ change le sens des phrases qui, pourtant, relatent un même fait. En quoi consiste cette différence de sens ? La première suggestion consiste à soutenir la définition informationnelle du Thème : le Thème est le topique de la phrase. Si l'on commence la phrase par *les canaris*, on s'attend à ce que la phrase soit à propos des canaris. De même pour *les mineurs* en (IV.4[2]). La différence entre (IV.4[1]) et (IV.4[2]) se situe donc au niveau du topique. Mais si l'on prend la phrase d'origine (IV.4), il semble qu'elle soit également au sujet des canaris. Pourquoi ? Car, au final, ce qu'on associe au statut de topique n'est pas le premier élément mais plutôt l'élément sujet.

> "The original version [(IV.4)] also seems intuitively to be 'about' yellow canaries, since that is the subject of the clause. In other words, this way of expressing the meaning of Theme [Theme is 'what the clause is about']makes it hard to distinguish it from Subject. That is why it is better to keep to the idea of Theme as the 'point of departure of the message'." (Thompson 2004:143).

Donc, ce qui diffère entre les différentes versions de (IV.4), ce n'est pas le topique sélectionné, mais le point de départ sélectionné. Nous en arrivons à la fonction d'Orientation. Le Thème, défini comme l'élément ou les éléments réalisés en premier, fonctionne non seulement au niveau idéationnel (en exprimant une circonstance ou un participant), mais également – et surtout – au niveau textuel, comme le schématise la figure IV.2 ci-dessous.

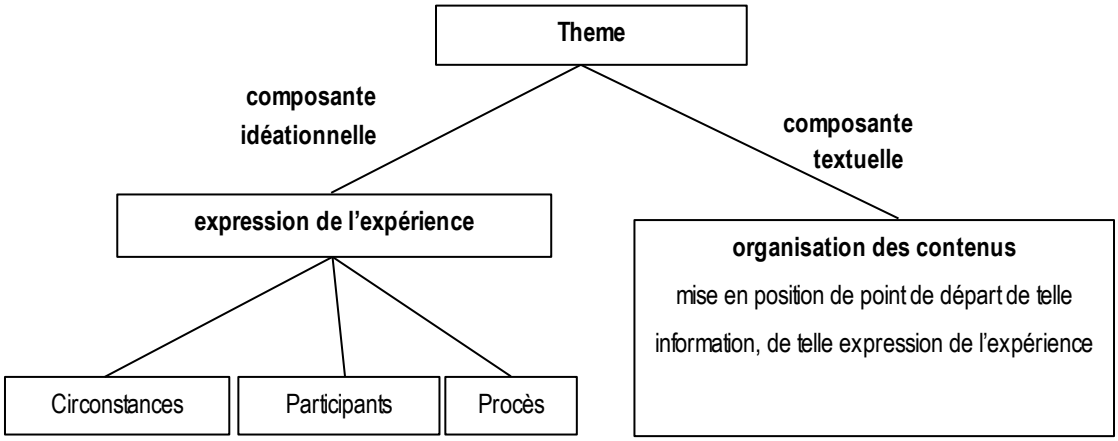


Figure IV.2 : Le Thème dans les composantes idéationnelle et textuelle

Parmi tous les ordres de constituants possibles, le lecteur va choisir telle organisation, car elle correspond à sa stratégie textuelle. Si l'auteur veut orienter son texte vers un style historique tout en gardant un point d'entrée large, il préférera commencer son texte par un adverbial temporel (un adverbial circonstanciel) comme dans la phrase originale (IV.4) qui constitue la première phrase d'un article sur l'histoire de l'industrie. L'adverbial suggère le début d'une TSC temporelle (l'organisation selon une TSC temporelle sera alors confirmée ou infirmée par la suite). De plus, la présence de cet adverbial 'pousse' en deuxième position le sujet grammatical. Ceci permet de ne pas attribuer au sujet (*les canaris*) un statut topical trop important tel que le statut de topique de discours (le sujet général – global – du texte). Les points d'entrée des versions construites ((IV.4$^{1...4}$) ne permettent ni d'initier une TSC temporelle, ni de rester 'large' et libre.

Bien que la définition du Thème en SF se situe à un niveau purement phrastique, les concepts de point de départ et d'orientation (définis généralement au niveau phrastique) s'adaptent relativement bien au niveau supra-phrastique. Ainsi, la définition de la fonction du Thème selon Fries (1995b:58) : "The Theme of a clause complex provides a framework within which the Rheme of the clause complex can be interpreted" se transforme aisément, dans notre thèse, en : "**The Theme (*i.e.* the first elements) of a discourse span provides a framework within which the rest of this span can be interpreted**"

La fonction d'orientation est rarement associée directement à la fonction Thème dans les études en SF, et pour cause, la fonction d'Orientation prend toute son ampleur au niveau de l'organisation discursive et le Thème en SF reste un concept grammatical, une fonction se réalisant au niveau phrastique. Malgré cette différence de niveau d'analyse, le lien entre orientation et Thème se retrouve en filigrane dans les cinq fonctions majeures du Theme selon la SF : (a) fournir un contexte à la prédication, (b) préciser une information nécessaire (cruciale) pour l'interprétation du message, (c) 'remplir' la position initiale afin de permettre aux informations focus d'être situées en fin de phrase, (d) contribuer à la séquentialité du discours en indiquant une continuité ou une discontinuité et (e) guider le lecteur dans sa façon d'interpréter les propos à venir.

> "Broadly speaking, the 'enabling' potential of SFG [Systemic Functionnal Grammar] Theme is substantiated in five major functional tasks :
> a. to provide a framework for the interpretation of Rheme ;
> b. to add information which is required for the interpretation of the message ;
> c. (acting negatively) to co-related with the principles of EF [end focus] and EW [end weight], helping to build up the *discourse prominence* of (an) item(s): placing an item late in a clause (complex) endows it with the status of new information (in unmarked cases) ;
> d. to contribute to the *Topic continuity* or *discontinuity* of discourse, by either developing or cancelling an assumption which has been established in the previous context ; and as a corollary
> e. to act as an *orientator* to both the message conveyed by the clause and to the addressee's expectations as to how to understand what is about to come." (Gómez-González 2001:97)

Selon la nature du Thème c'est-à-dire sa structure même, l'une de ces cinq fonctions est plus ou moins mise en avant. Ainsi, la mise en place d'un contexte, fonction (a), est vraisemblablement plus portée par des adverbiaux circonstanciels en position Thème. La fonction (b) rejoint le principe de l'information cruciale en premier et peut être réalisée par la précision d'un *setting*, mais également d'une modalité d'énonciation ou d'un certain mode organisationnel (en amorçant, par exemple, une énumération). Elle peut également se réaliser par la forme du sujet grammatical, notamment par l'utilisation de redénominations ou de reclassifications (voir le chapitre suivant). La fonction (c) touchant à la répartition équilibrée du message ne peut pas être réalisée par des constructions spéciales aussi appelées phrases à Thèmes spécifiques (voir infra). Le rôle du Thème dans le marquage de la continuité ou de la

discontinuité topicale va plutôt se jouer par des Thèmes réalisés par des sujets grammaticaux. Enfin, la dernière fonction ne peut *a priori* pas se réaliser par un pronom anaphorique en sujet grammatical.

Ces associations formes/fonctions paraissent sans doute très caricaturales. Ce sont cependant des corrélations de ce type qui sont testées dans cette thèse (voir le chapitre VII). Les parties suivantes présentent en détail ces différentes formes de Thème. Mais avant de procéder à ces présentations, il faut délimiter formellement ce que l'on considère comme constituant la position Thème.

IV.3.2. Jusqu'où étendre la position Thème ?

Il ne semble pas y avoir d'accord en SF pour délimiter la position Thème. Certains affirment que le Thème correspond simplement au premier élément de la phrase, d'autres précisent qu'il couvre toute la position initiale de la phrase jusqu'au premier élément idéationnel. D'autres, dont nous faisons partie, considèrent le Thème comme englobant toute la position préverbale. Dans cette définition, le dernier élément Thème correspond au sujet grammatical : le thème topical.

Selon notre identification le Thème de la phrase (IV.5) correspond à deux éléments : *Pendant des siècles,* et *les canaris jaunes*, le premier étant un Thème idéationnel marqué et le second un Thème idéationnel non marqué, ce qui correspond dans notre terminologie au Thème topical.

(IV.5) *Pendant des siècles, les canaris jaunes ont été utilisés pour 'tester' l'air dans les mines.*

Cette façon de délimiter n'est pas commune à la plupart des systémiciens qui préfèrent considérer que le Thème de (IV.5) couvre tout ce qui se situe entre le début de la phrase et le premier élément idéationnel : *Pendant des siècles.* Pourquoi envisager la possibilité d'avoir deux éléments idéationnels Thème ? Tout simplement, car nous considérons qu'il peut y avoir deux stratégies organisationnelles conjointes : par exemple une TSC temporel et une TSC topicale, la première fonctionnant au niveau de la construction du Fond par une technique d'encadrement, la seconde au niveau de la Figure par des techniques de progressions thématiques par exemple (voir III.1.2.b). Nous envisageons la possibilité de combinaisons entre des Thèmes scéniques et des Thèmes topicaux (voir IV.4.1)

La distinction entre fond et figure est pour nous de même importance que la différence entre les composantes interpersonnelle, textuelle et idéationnelle. Or la SF considère qu'il peut y avoir simultanément plusieurs Thèmes relatifs aux différentes composantes du discours. On parle alors de Thèmes multiples.

> "Lexical elements such as conjunctive and modal Adjunct, which express primarily textual and interpersonal meaning have the function of 'placing' the content, of signalling how it fits coherently with the content around it. They therefore naturally tend to gravitate towards the beginning of the clause, which is the structural slot (the Theme) where 'fitting-in work' is done." (Thompson 2004:158)

L'exemple IV.6 qui présente le début d'un paragraphe nous montre un Thème simple (*La mondialisation*) qui fait office de Thème topical pour la première phrase (et peut-être du reste du paragraphe ... voir IV.5). La phrase suivante présente un Thème multiple composé de quatre éléments : un Thème textuel (*Ainsi*), suivi d'un Thème idéationnel marqué (*en l'espace de 10 ans*), suivi d'un deuxième Thème textuel (*par exemple*) et enfin, nous avons le Thème topical sujet (*l'espace de référence des entreprises françaises*). La dernière phrase comporte deux Thèmes idéationnels : un marqué et un non marqué.

(IV.6) ***La mondialisation*** *intervient donc comme un processus d'élargissement de l'espace de référence dans lequel les acteurs sociaux s'insèrent.* ***Ainsi, en l'espace de 10 ans, par exemple, l'espace de référence des entreprises françaises*** *s'est déplacé de l'Europe vers le monde.* ***En dix ans, le fait de s'européaniser*** *s'est trouvé dépassé par la nécessité de se mondialiser. [GEOPO_25]*

La mondialisation	*intervient donc comme un processus d'élargissement de l'espace de ...*
non marqué	Rhème
Idéationnel	
Thème	

Ainsi,	*en l'espace de 10 ans,*	*par exemple,*	*l'espace de référence des ...*	*s'est déplacé ...*
text.	marqué	textuel	Non marqué	Rhème
	idéationnel		idéationnel	
Thème				

En dix ans,	*le fait de s'européaniser*	*s'est trouvé dépassé par la nécessité de se mondialiser.*
marqué	non marqué	Rhème
Thème idéationnel		

Nous voyons bien dans cet exemple comment la fonction d'orientation peut jouer à la fois au niveau de plusieurs composantes. *Ainsi,* et *par exemple* jouent à un niveau purement organisationnel. Ils indiquent seulement comment intégrer les informations arrivantes aux informations déjà enregistrées, quel statut leur donné dans le flot du discours. Les Thèmes idéationnels n'ont pas pour fonction première la gestion du signalement de la séquentialité du discours (nous verrons dans le chapitre suivant que de nombreux Thèmes idéationnels portent, dans le choix de leur forme, un sens instructionnel). Ils expriment les 'bouts' d'expériences concernés par la phrase : les circonstances, les entités, les participants, les actions, etc. En SF traditionnelle, il ne peut y avoir qu'un seul Thème idéationnel qui peut être non marqué s'il correspond au sujet grammatical (on parle alors de Thème topical) ou marqué s'il est extra-prédicatif. Ainsi, l'analyse traditionnelle de la seconde phrase de l'exemple est :

En dix ans,	*le fait de s'européaniser s'est trouvé dépassé par la nécessité de se mondialiser.*
Thème idéationnel marqué	Rhème

L'adverbial temporel 'prend la place' du Thème topical. Une telle analyse va à l'encontre de toutes les modalisations existantes autour de la continuité topicale. Selon la théorie du centrage, les centres préférés sont les sujets grammaticaux (III.3.2). Selon le modèle des progressions thématiques, la phrase 1 et 2 sont reliées par une relation par Thèmes dérivés qui n'est envisageable que si *l'espace de référence des entreprises françaises* a le statut de Thème.

De plus, la prise en compte de toute la zone délimitée à droite par le sujet grammatical permet de considérer *par exemple* comme un Thème textuel (une analyse traditionnelle s'arrêterait au premier élément idéationnel rencontré, à savoir *en l'espace de 10 ans*). La zone Thème étant ainsi délimitée à toute la zone allant du début de la phrase au sujet grammatical (inclus), différents types d'éléments Thèmes sont à distinguer selon leur mode de construction :

- les Thèmes topicaux – **ThTop** – qui correspondent aux sujets grammaticaux référentiels, *i.e.* qui expriment un élément idéationnel (une circonstance, une entité ou un procès) ;
- les Thèmes spécifiques – **ThSpe** – qui correspondent aux phrases relevant d'une construction spéciale où le sujet grammatical est sémantiquement vide ou ne réfère pas à un élément idéationnel ;
- les Thèmes marqués – **INIT** – qui présentent la caractéristique d'être détachés de – ou non intégrés à – la proposition principale. Ces Thèmes peuvent exprimer un élément interpersonnel, textuel ou idéationnel (nous parlons alors de Thème scénique).

IV.4. De l'ordre à l'initiale : les Thèmes multiples

Selon notre délimitation de la zone Thème, il est assez fréquent d'observer plusieurs éléments en position initiale. Selon Halliday, le Thème, tout comme la phrase répond à une structure qui s'inscrit elle aussi dans les différentes composantes fonctionnelles.

> "The internal structure of a multiple theme is based on the functional principle […] that the clause is the product of three simultaneous semantic processes. It is at one and the same time a representation of experience, an interactive exchange, and a message. […] **There is always an ideational element in the Theme.** There may be, but are not necessarily, interpersonal and/or textual elements as well. The typical overall sequence of these elements is : textual^interpersonal^ideational. […] the ideational element is always the final one." (Halliday 1985:53-54)

Donc selon Halliday, il y a un ordre bien précis des éléments Thèmes. Cependant, comme nous l'avons dans l'exemple IV.7, il peut y avoir un Thème textuel entre deux Thèmes idéationnels : un Thème scénique et un Thème topical. On peut également trouver un Thème interpersonnel après un Thème textuel (IV.8) ou après un Thème scénique (IV.9).

(IV.7) *Ainsi, en l'espace de 10 ans, par exemple, l'espace de référence des entreprises françaises s'est déplacé de l'Europe vers le monde.* *[GEOPO_25]*

(IV.8) *En second lieu, toutefois, le risque politique principal ne vient sans doute pas des 'aléas de la conjoncture' ou des changements de gouvernement ; ...* *[GEOPO_30]*

(IV.9) *En 69, sans doute, il est élu questeur, ce qui lui assure un siège au Sénat, et exerce cette fonction en Espagne.* *[PEOPL_6]*

Halliday souligne qu'il y a obligatoirement un Thème idéationnel. Cependant, en considérant les Thèmes spécifiques, on peut se trouver face à des phrases sans Thème idéationnel :

(IV.10) *C'est dans cette perspective aussi que doit être apprécié le déploiement de certains systèmes militaires.* *[GEOPO_28]*

On peut d'ailleurs avoir des Thèmes interpersonnels, textuels et scéniques devant des constructions à Thème spécifique. La SF qui considère que la position Thème s'arrête au premier élément idéationnel n'envisage pas de telles configurations qui, alors, poseraient le problème de considérer ou non qu'il y a un Thème spécifique, notamment dans celles où l'on a un Thème scénique orientant une construction spéciale.

(IV.11) *Depuis 1987, avec l'essor des 2e cycles, c'est le corps enseignant du secondaire qui a connu les changements les plus profonds.* *[ATLAS_2]*

Dans une telle phrase, nous considérons qu'il y a un Thème spécifique précédé d'un Thème scénique (la SF traditionnelle analyserait ici un Thème idéationnel simple : *Depuis 1987*).

Dans une tout autre perspective, Virtanen (1992) suggère la confrontation de deux facteurs dans l'ordonnancement des éléments en position initiale : d'une part un facteur lié à la syntaxe de la phrase, d'autre part, un facteur lié à la cohésion du discours, et plus précisément à la gestion des stratégies textuelles. Elle distingue ainsi des adverbiaux impliqués dans une TSC des autres adverbiaux, en remarquant que les premiers sont placés avant les derniers.

> "the first of them [homosemantic adverbials clustering at the beginning of the sentence] typically belongs to the global chain of text strategically important adverbials, while the other(s) convey additional and often more specific information. Similarly, heterosemantic adverbials clustering at the beginning of the sentence tend to follow the order of placing the text-strategically important adverbial first, before other adverbials" (Virtanen 1992:255)

Virtanen applique ici le principe de l'information cruciale en premier (CIF) qui consiste à positionner en premier ce qui relève de l'organisation discursive, puis de l'organisation phrastique. Ce principe peut également expliquer pourquoi les Thèmes textuels et interpersonnels apparaissent préférentiellement avant les Thèmes topicaux. Ce principe explique également pourquoi on peut vraisemblablement avoir un Thème scénique avant tout autre type de Thème, si ce Thème scénique joue au niveau du TSC. Cette répartition est confirmée par le travail d'Hasselgård (1996) qui observe que dans la plupart des cas de cluster spatio-temporel en initial, le premier adverbial est un adverbial de phrase.

IV.4.1. Thème topical et Thème scénique

Le Thème topical est un élément central en SF. Il constitue l'élément essentiel à tout message (excepté dans les cas de Thèmes spécifiques, voir la partie suivante). Il correspond à l'élément idéationnel qui constitue le topique de la phrase. Pour Halliday comme pour la plupart des systémiciens, le Thème topical peut tout à fait être un circonstant. Il suffit que celui-ci soit le premier élément idéationnel de la phrase.

> In principle, an ideational element is anything representing a process, a participant in process (person, thing, institution, etc.) Or a circumstance attendant on that process (time, place, manner, etc.) […] The ideational element within the Theme, then, is some entity functioning as Subject, Complement or Circumstantial Adjunct ; we shall refer to this as the TOPICAL THEME, since it correspond fairly well to the element identified as 'topic' in topic-comment analysis." (Halliday 1985:54)

Cependant, la définition du Thème topical par Halliday confond deux concepts : la position initiale et le statut topical. Or, si l'on considère les adverbiaux circonstanciels (qui nous intéressent particulièrement dans cette thèse), cette confusion devient inapplicable. En effet, comme le montrent Virtanen (1992) ou Charolles (2003), les adverbiaux extra-prédicatifs n'ont pas de statut informationnel défini.

> "Sentence-initial adverbials of time and place are thematic but they only occasionally convey textually given information. Instead, they typically convey information that has been characterized as 'inferable' by Prince (1981). While the kind of information that these strategy-marking adverbials represent can usually be situated in the middle part of a scale reaching from given to new information, it is important to note that they can have any status on such a scale." (Virtanen 2004:84)

Les adverbiaux circonstanciels, tout comme les adverbiaux modalisateurs ou textuels ont plus une fonction discursive qu'une fonction phrastique. Or, la définition du Thème en SF reste au niveau du message*, i.e.* au niveau de la phrase vue comme porteuse d'une représentation mentale, d'une intention et d'une organisation interne. Il n'y pas de fonction spécifique à associer aux éléments détachés en initiale si l'on reste dans les limites de la phrase, et surtout pas la fonction de topique de phrase, généralement remplie par l'élément en fonction sujet.

> « Le fait que les adverbiaux antéposés puissent servir de topique pour la proposition en tête de laquelle ils sont détachés ne va pas de soi dans la mesure où celle-ci a en principe déjà un topique non marqué (généralement son sujet). » (Charolles 2003:29)

Les adverbiaux circonstanciels posent le décor, construisent le « fond » pour la suite des propos (et pas seulement pour le procès en cours de description). Ils constituent ce que nous appelons des Thèmes scéniques. Ils appartiennent à des processus qui ne peuvent s'envisager que d'un point de vue discursif.

Thompson (2004:173-174) remarque, dans une note finale sur l'identification du Thème, que certains travaux en SF (il ne cite pas lesquels, mais nous en faisons partie) poussent les limites de la zone Thème jusqu'au premier Thème non marqué (*i.e.* Thème sujet). Dans cette approche, le Thème marqué précédant le Thème non marqué est appelé

« *Contextual Frame* » ou « *Orienting Theme* ». Cette approche correspond entièrement à notre conception de la zone Thème : une zone où l'on peut avoir des Thèmes discursifs, participant à des stratégies textuelles ne s'établissant pas forcément par processus de connexion de phrase en phrase (voir II.3.2) et des Thèmes locaux, participant à la construction du message au niveau de la phrase et à des stratégies textuelles établies par processus de connexion.

> "the marked Theme is seen as a 'Contextual Frame' or Orienting Theme', which typically has the function of changing the textual framework in some way ; the Subject Theme, on the other hand, typically serves to maintain the topic of the text." (Thompson 2004:173)

La notion de 'Orienting Theme' n'est pas vraiment présente dans les travaux en SF. Thompson remarque qu'il y a une réelle confrontation entre d'une part une définition fonctionnelle du Thème et une identification formelle de la zone Thème. Nous nous trouvons au coeur de cette recherche de corrélation entre forme et fonction (voir le chapitre suivant et la partie VI.1.4). Notre position est de considérer une position initiale 'large', afin de considérer la totalité des éléments pouvant participer à une stratégie textuelle et d'observer ainsi les comportements de tous les Thèmes possibles. Cette position semble être la plus judicieuse, comme le souligne Thompson (2004:174) :

> "It is probably only through large-scale analysis of Theme in many texts and many registers that we can hope to pin down more accurately what Theme does." (Thompson 2004:174)

IV.4.2. Les Thèmes spécifiques – ThSpe

Les constructions à thème spécifique, encore appelées constructions syntaxiquement marquées ou constructions spéciales, sont des structures syntaxiques qui offrent une structure informationnelle différente de la structure canonique *topic-comment.* Il s'agit soit d'inverser la structure, soit de mettre en fonction thème un élément vide au niveau idéationnel. Dans sa terminologie, Halliday fait la distinction entre des thèmes dits « topicaux » et des thèmes dits «spécifiques ». Nous garderons cette terminologie en utilisant l'abréviation ThTop/ThSpe. De fait, les ThTop correspondent aux sujets grammaticaux des phrases ne correspondant pas à une construction spéciale, et les ThSpe correspondent à la dénomination de la construction spéciale (et non pas à la nature du sujet grammatical de cette construction). Les phrases suivantes présentent toutes un Thème spécifique – ThSpe :

(IV.12) *C'est là le rôle essentiel du tout nouveau Secrétariat à la Protection du Territoire.* *[GEOPO_1]*

(IV.13) *Mais s'il n'y a pas de recette magique pour le succès, il y en a bien une pour l'échec : la fermeture des frontières.* *[GEOPO_16]*

(IV.14) *Mais il est encore difficile de savoir dans quelle mesure ces pratiques ont permis une décentralisation et une plus large contribution à la prise de décision, ...* *[GEOPO_16]*

(IV.15) *Parmi les versions de l'âge romantique, on peut encore mentionner l'Étudiant de Salamanque (1840), poème narratif d'Espronceda, lequel...* *[PEOPL_1]*

(IV.16) *Sont donc élus en 1973 de nouveaux députés dont certains jouent encore fin des années 1990 un rôle important dans la vie politique locale et nationale : ...* *[ATLAS_3]*

(IV.17) *Le joyeux trompeur, c'est Juan Tenorio, jeune seigneur qui se divertit à abuser des femmes en leur faisant croire qu'il les épousera, ainsi qu'à berner (burlar) leurs maris, leurs fiancés, leurs amis, qui sont parfois les siens.* *[PEOPL_1]*

Nous identifions six types de ThSpe : les constructions clivées et pseudo-clivées (IV.12), les constructions présentationnelles (IV.13), les constructions impersonnelles (IV.14), les phrases en *on...* (IV.15), inversions du sujet (IV.16) et les dislocations (IV.17). La catégorie des phrases en *on* n'est généralement pas identifiée comme étant un ThSpe (Gómez-González (2001) et Thompson (2004) n'y font pas référence). Ces constructions sont caractérisées par le fait que leur Thème topical réfère au locuteur même, à une communauté de pensée, à la pensée générale. Ces constructions se rapprochent des constructions passives et constructions impersonnelles qui permettent de mettre le sujet réel en arrière-plan. Tous comme les impersonnelles, les constructions en *on...* peuvent également mettre en

avant une modalisation d'énonciation. Tous ces rapprochements entre les constructions en *On* et les impersonnelles ou les passifs justifient le fait de classer ces constructions dans la catégorie des ThSpe.

L'utilisation des constructions spéciales produit un effet de rupture de la progression thématique. Cette rupture répond à plusieurs configurations. Il peut y avoir, comme le remarque Enkvist, une absence de candidat à la fonction de Thème. Alors une construction de type présentationnel permet d'introduire un nouveau référent.

> "The general rule is that new discourse referents are introduced into a text in the rheme ; if there is no chance of introducing the rheme by hanging it onto a theme expressing old information, special constructions (existential structures, etc.) must be used. Once a discourse referent has been introduced into the text, it counts as old information" (Enkvist 1978)

Mais cet usage est relativement rare. En fait, il est rare d'introduire abruptement un nouveau référent. La plupart du temps, les constructions spéciales permettent soit de thématiser un élément de la composante interpersonnelle (constructions impersonnelles), soit d'assurer la localisation correcte du focus d'attention (clivées, dislocations), soit de thématiser un élément du procès autre que son sujet – un attribut ou un complément (sujets inversés).

> "our findings indicate that [these constructions] tend to cluster around relational processes, which, rather than simply presenting reality, give subjective synoptic views of it, (a) thematising interpersonal meaning, (b) guaranteeing the right identification of a given referent by cancelling or reinforcing a previous presupposition made about it, and (c) thematising an Attribute of a rhematic referent." (Gómez-González 2001:250)

Les six types de ThSpe sont définis formellement en (VII.2.2.c) et leur contribution au marquage de la séquentialité est présentée dans le chapitre suivant. Nous avons classé ces différents types de ThSpe selon la composante dans laquelle ils s'inscrivent le plus. Les clivées, les présentationnelles, les inversions et les dislocations permettent une organisation différente de l'information en mettant le focus sur un élément de l'expérience (composante idéationnelle), alors que les impersonnelles mettent en position initiale une modalité d'énonciation (composante interpersonnelle). Concernant les constructions en *on...* nous les situons uniquement dans la composante textuelle.

IV.5. Une initiale de plusieurs niveaux

Dans notre étude de la composition de la position initiale, nous distinguons trois niveaux d'organisation associés à trois positions textuelles – PosTxt : l'initiale de sections, l'initiale de paragraphes (initiale de sections exclue) et l'initiale de phrases (initiale de paragraphes exclue). Nous supposons que le rôle de la position initiale diffère selon le niveau considéré. À titre d'hypothèse, nous associons les stratégies de rupture à la position initiale de sections – S1, les stratégies de continuités à la position initiale de phrases intraparagraphiques – P2, et les stratégies de déplacement à la position initiale de paragraphes (hors premier paragraphe de section) – P1. Ainsi, nous espérons que notre étude en corpus présentera des résultats du type :

- les marqueurs de continuité topicale prototypiques que sont les pronoms personnels de troisième personne (PRO3) apparaissent significativement plus en P2 qu'en P1 et S1.
- Les marqueurs de continuité référentielle marquée (car déplacement) que sont les redénominations apparaissent significativement plus en P1
- Les marqueurs de rupture apparaissent significativement plus en S1.

Cependant, notre étude en corpus va plus loin. Nous avons généralement évité de faire des associations faite entre des formes et des fonctions discursives ou des stratégies. Or le premier point cité ci-dessus est justement une association entre une forme et une fonction. Ce premier point est exposé dans un but principalement illustratif (bien que nous pensions assurément à la validité de cette association). En fait, notre étude aborde le problème dans l'autre sens. En partant de l'association entre les trois niveaux d'organisation et les trois types de stratégies de (dis)continuité,

notre exploration en corpus fait émerger des associations entre formes et fonctions. Elle n'est pas là pour valider des corrélats linguistiques, mais pour en mettre au jour tout en restant dans une recherche 'ouverte' (ce terme est expliqué dans la partie VI.1.4.L'EOD en corpus : une recherche ouverte).

Par exemple, en (IV.4) nous avons un Thème interpersonnel : *Malheureusement* par lequel le locuteur exprime au lecteur sa position vis-à-vis des faits relatés, et un Thème topical : *certaines difficultés structurelles et conjoncturelles*.

(IV.18) ***Malheureusement, certaines difficultés structurelles et conjoncturelles*** *sont apparues et cette étude officielle n'a pas encore vu le jour. Le groupe de réflexion chargé de l'étude était une sous-commission particulière du Policy Coordinating Committee sur l'espace (PCC-space) créé par le NSC. Dans les faits, les efforts du NSC en matière spatiale n'ont pas été suffisants. L'autorité au sein des sous-groupes n'était pas clairement attribuée au NSC. Ed Bolton, Director for Space au NSC sous l'autorité de Frank Miller, n'était pas de rang suffisant pour imposer des compromis aux différentes agences réunies dans le PCC-Space. Surtout, les événements du 11 septembre ont axé les priorités du gouvernement sur l'action et non sur la réflexion.[GEOPO_2]*

Nous avons retranscrit ici tout un paragraphe afin de montrer comment on peut avoir des Thèmes de paragraphes, quelle que soit la composante dans laquelle joue l'élément Thème (tout le paragraphe porte un regard triste sur les difficultés qu'a rencontrées l'étude que décrit cet article[67]). Notre conception d'une initiale à plusieurs niveaux fait écho au rapprochement que nous avons déjà souligné entre le fonctionnement discursif des titres et le fonctionnement discursif des introducteurs de cadres et des adverbiaux constitutifs d'une TSC. La position initiale de paragraphes – tout comme la position initiale de sections – comporte des éléments qui vont orienter la suite du paragraphe ou de la section. Il est alors tout à fait justifié de parler de Thème de paragraphe, de Thème de section, etc. L'exemple de Thompson (2004) s'appuie précisément sur cette idée d'un Thème dépassant la taille de la phrase pour expliquer le choix pour tel ou tel élément en position Thème de la première phrase d'un article sur l'histoire de l'industrie (voir IV.3.1).

Virtanen (1992) et Gernsbacher & Robertson (2002) défendent elles aussi l'idée d'une initiale de plusieurs niveaux. Ainsi, Virtanen implique les adverbiaux en initiale de paragraphes dans la construction de TSC globales (lorsqu'il y a plusieurs paragraphes successifs introduits par un adverbial), alors que les adverbiaux en position initiale de phrases intraparagraphiques sont associés à des TSC locales. De leur côté, Gernsbacher & Robertson (2002) étendent l'idée de l'avantage de la première mention (« Advantage of First Mention », voir IV.1) aux différentes unités textuelles. Ainsi, elles considèrent qu'au niveau phrastique, le 'thème' est souvent associée à la position initiale ; au niveau paragraphique, le thème ou topique de paragraphe sera préférentiellement exprimé en début de celui-ci ; enfin, au niveau des sections, le principe de la première mention prédit une accessibilité plus forte des éléments figurant dans sa première partie (Gernsbacher & Robertson 2002:120).

°, °, °, °

Nous avons dans ce chapitre posé notre conception de la position initiale, vue en tant que position stratégique dans l'organisation du discours. La position initiale correspond au point de départ de toute unité textuelle : de la phrase au texte. C'est en cette position que vont se trouver les éléments qui orientent le reste de l'unité : soit en posant les fondations du *text-world* représenté, soit en signalant comment intégrer les informations contenues dans l'unité avec le reste du discours. La position initiale est un attribut qui confère aux éléments qui s'y trouve la fonction d'orientation.

67 Cet article relate un rapport qui avait pour objectif initial de présenter les résultats d'une étude menée par le National Security Council (NSC) sur le contrôle de la diffusion de l'imagerie commerciale par le gouvernement américain. Lancée au printemps 2001, cette étude devait s'achever en début d'année 2002, mais les attentats de 2001 ont changé la donne puisque l'administration Bush a alors misé « *sur l'action et non sur la réflexion* ».

Quel que soit l'élément commençant l'unité textuelle, le lecteur lui donnera un poids particulier dans son interprétation. Cependant, selon la nature de l'élément qui s'y trouve, différentes instructions sont reçues. Le chapitre suivant dresse une liste des différents éléments que l'on peut trouver en position initiale. Il permet de poser nos connaissances sur le sens instructionnel de ces différents éléments qui, par leur positionnement à l'initiale, constituent des indices de séquentialité.

Chapitre V
Indices de séquentialité en position initiale

Sommaire

« *Discourse markers* » (Schiffrin 1987), « *cue phrases* » (Grosz & Sidner 1986), « marqueurs de segmentation » (Bestgen & Vonk 1995, 2000 Bestgen & Costermans 1997), « *textual organizers* » (Schneuwly 1997), « *text cues* » (Gaddy *et al.* 2001), « *text-signaling devices* » (Lorch 1989) ...[68], les dénominations pour référer aux éléments qui indiquent l'organisation du discours sont nombreuses et souvent associées à des familles de travaux dont certaines références sont données entre parenthèses.

68 Romera (2004) établit une longue liste de toutes les principales acceptions recensées autour de cette notion (voir Romera 2004:6) : *discourse markers*, *discourse operators*, *discourse particles*, *cue phrases*, *connectives phrases*, *pragmatic markers*, *pragmatics particles*, *punctors*, *discourse connectives*, etc.

La notion de marqueur est ancrée dans des travaux appartenant à des domaines bien différents, allant de la linguistique cognitive à la sémantique formelle. D'un point de vue cognitif, un intérêt particulier est porté sur les indices auxquels se fie le lecteur afin de construire sa représentation mentale et sur les traces laissées par le locuteur, signes de l'organisation de sa représentation mentale. D'un point de vue sémantique, les différentes études cherchent principalement à mettre en corrélation des relations propositionnelles[69] et des formes lexicales. Récemment, une nouvelle vague de travaux en EOD dans laquelle nous nous situons se développe, ayant pour objectif commun la découverte des réalisations textuelles de phénomènes discursifs, sans poser comme principe l'idée d'une corrélation fixe et binaire entre une forme et une fonction. Ces travaux se situent plus dans la lignée de la linguistique cognitive à la différence près qu'ils ne se basent pas sur des expérimentations psycholinguistiques mais sur des productions linguistiques issues de situations de communication réelles.

Nous n'avons pas utilisé le terme de marqueur discursif dans le titre de ce chapitre car la définition la plus commune des marqueurs discursifs ne correspond pas à notre définition (V.1.1). En parcourant les différentes études sur l'organisation discursive telle que nous l'entendons (Charolles 1988-1997, Gernsbacher 1990, Goutsos 1996, Péry-Woodley 2000-2001, Virtanen 1992 *cf.* chapitre II et chapitre III), toutes parlent d'indices, mais peu d'entre elles parlent de « marqueurs discursifs ».

Trois points éloignent notre définition des indices de l'organisation discursive de celle des marqueurs discursifs. Premièrement, la notion de marqueur discursif est communément réservée à l'analyse de la connexion entre deux propositions plus qu'à l'étude d'organisations discursives plus globales. Deuxièmement, le terme « marqueur » suppose une corrélation entre **une** fonction et **une** forme. Or, il semble plus réaliste de miser sur une configuration de formes, comme le défendent Schneuwly (1997), Péry-Woodley (2005) et nous-même. Troisièmement, la 'catégorie' des marqueurs discursifs, très liée à celle des connecteurs, regroupe des éléments définis essentiellement par une absence de contenu représentationnel. Or, de nombreuses expressions jouant au niveau de la composante idéationnelle participent également à la composante textuelle en délimitant des segments textuels. Comme nous l'avons esquissé dans le chapitre III, les introducteurs de cadre et les différentes formes que peut prendre le sujet grammatical portent un sens instructionnel en même temps qu'elles apportent des informations constitutives de la représentation mentale. Ces éléments jouant conjointement sur plusieurs niveaux nous intéressent plus particulièrement.

Avant de présenter les différents indices – et notamment leur sens instructionnel – que nous avons rencontrés au cours de notre exploration en corpus des éléments en position initiale, la première partie de ce chapitre expose plus précisément notre conception du signalement de l'organisation discursive à la surface du texte.

V.1. Signalement de l'organisation discursive à la surface du texte

V.1.1. Point de vue sémantique : les « marqueurs discursifs »

Les marqueurs discursifs sont traditionnellement définis par la caractéristique commune de ne pas avoir de contenu représentationnel (c'est-à-dire de ne pas contribuer à la composante idéationnelle). Leur sens est uniquement instructionnel, consistant à indiquer au lecteur comment relier les propos à venir au discours précédent. Ils constituent

69 Le terme « relations propositionnelles » est utilisé par la RST là où le modèle de la SDRT utilise le terme « relation de discours » (cf. III.1.2.b).

une marque explicite des relations propositionnelles entre segments de discours ou simplement des limites d'un segment de discours.

Les premiers marqueurs discursifs considérés dans la littérature sont des expressions propres à l'oral du type *Ben, j'veux dire, tu vois, euh, etc.* Ces expressions sont caractérisées par le fait qu'elles n'entrent ni dans la construction syntaxique des propositions ni dans la construction sémantique du procès relaté dans la proposition. Du fait de leur situation marginale, ces expressions se sont vues associer un rôle discursif. Petit à petit, la notion s'est développée dans le monde de l'écrit. Il s'agissait alors de trouver des équivalents à ces expressions. Ainsi, Schiffrin (1987) remarque que des expressions comme *justement, franchement, en fait (*ou en anglais *of course, in fact, in any case* et en allemand *etwa, jedenfalls, immerhin*[70]) peuvent fonctionner tant à l'oral qu'à l'écrit. En élargissant davantage la notion des marqueurs discursifs, Britton (1996) note que les marqueurs discursifs jouent un rôle essentiel dans les écrits narratifs, où ils servent, entre autres, à indiquer les lieux de transition ou à insister sur des événements importants. Siepmann (2005) note cependant que beaucoup de linguistes réservent encore la notion de marqueur discursif à l'analyse du discours oral (Siepmann 2005:38).

Désormais, et en dehors des partisans de l'oral, la notion s'est élargie à toutes les expressions porteuses d'un sens instructionnel. Nous sommes donc passée de l'étude des éléments marginaux qui étaient de fait candidats à un rôle non syntaxique à l'étude de tous les éléments capables d'expliciter une relation entre segments de discours.

Pour Romera (2004), les marqueurs discursifs sont des expressions qui apparaissent à la frontière d'unités de discours et établissent une relation de cohérence entre ces unités (cf. Romera 2004:72). La 'catégorie' des connecteurs devient alors la catégorie la plus étudiée des marqueurs discursifs. Ainsi, de nombreux travaux en sémantique cherchent à identifier le sens instructionnel des différentes formes de connecteurs ou à déterminer la liste de connecteurs susceptibles d'expliciter telle ou telle relation de discours (voir les travaux en SDRT – Asher & Vieu (2005), Danlos (2004), Knott (1996), etc., les projets de corpus annotés discursivement – Penn Discourse Treebank (Miltsakaki *et al.* 2004), RST corpus (Carlson *et al.* 2003), etc.)

Mais les marqueurs discursifs ne se réduisent pas à une catégorie, si l'on peut considérer les connecteurs comme une catégorie[71]. Ce qui fait d'un élément (une conjonction, un adverbe, un adverbial, etc.) un marqueur discursif, c'est sa fonction dans le discours et non une quelconque nature « absolue ». De ce fait, il est vraisemblablement difficile de trouver un terrain d'entente et de délimiter ce sur quoi il faut travailler, ce qu'il faut étudier. Certains traits formels semblent cependant pouvoir être utilisés en tant qu'indices pertinents d'accès à une fonction discursive, mais comme le précise Romera, ce ne sont que des indices dont l'usage est étroitement lié à l'objectif de l'étude ; et non des traits caractéristiques d'une fonction de niveau discursif :

> "Formal features do not constitute reliable criteria to support an independent definition of Discourse Markers or a characterization of their status as a discourse category. All the properties used to define the category of Discourse Markers come from analysis of partial sets of expressions which vary greatly depending on the researcher's perspective." (Romera 2004:7)

La seule définition valable des marqueurs discursifs est fonctionnelle ; et là encore, les marqueurs peuvent être porteurs de sens instructionnels variés, pouvant concerner :

- le statut informationnel et/ou l'importance discursive de tel élément ou telle unité à un niveau local ou global ;

70 Siepmann (2005) propose une analyse comparative des marqueurs discursifs en français, anglais et allemand.

71 La définition d'une catégorie « connecteur » reste encore à faire. Dans aucune grammaire il y a une liste finie des connecteurs. Nous y revenons en I.3.4.

- le type de relation en jeu entre deux unités, ou entre deux segments (continuité, discontinuité ou subordination) ;
- le rôle de tel ou tel segment dans l'articulation de portions de texte ou du texte dans son entier.

De plus, les marqueurs discursifs n'ont pas forcément un sens purement instructionnel. Ils peuvent également fonctionner sur le plan idéationnel (voir Siepmann 2005:40-41). Ce nouvel élargissement a fait entrer dans la catégorie des marqueurs discursifs de multiples expressions linguistiques comme, par exemple, les introducteurs de cadre de discours qui, au niveau textuel, participent à la segmentation du discours en cadres par leur portée cadrative et, au niveau idéationnel, posent un critère sémantique valable sur l'ensemble des informations contenues dans le cadre – c'est leur portée sémantique (voir Hồ-Đắc *et al.* 2001 et III.3.3.b).

Siepmann (2005) définit la notion de marqueur discursif par trois caractéristiques : une grande variété formelle, une fonction discursive de signalement des relations de cohérence et un impact cognitif sur les processus de compréhension.

> "The term 'discourse marker' can be applied to natural language strings of varying length and morphosyntactic structure whose primary function is to signal the coherence relations obtaining between a particular unit of discourse and other, surrounding units and/or aspects of the communicative situation and thereby to facilitate the listener's or reader's processing task." (Siepmann 2005:45)

Cette dernière caractéristique est essentielle à la définition du signalement de l'organisation discursive. Elle est peut-être même la plus définitoire des trois...

V.1.2. Point de vue cognitif : des indices textuels

Notre conception du signalement de l'organisation discursive à la surface du texte s'inscrit davantage dans une conception cognitive que dans une définition sémantique. Le lien entre le modèle de séquentialité et les modèles cognitifs est d'ailleurs clair. La définition du signalement de la séquentialité selon Goutsos recoupe complètement celle de différents travaux cognitifs.

> "'Signalling' implies that linguistic devices function as cues to the reader with regard to the way discourse proceeds. More particularly, they help the reader to assign the utterance in which they occur to a continuation or a transition space. For the writer, these devices constitute the resources that are drawn upon in the accomplishment of the tasks of continuity and discontinuity." (Goutsos 1996:517)

Les travaux en linguistique cognitive sur les indices textuels ne cherchent pas à établir des listes de corrélation entre un « marqueur » et une fonction discursive, mais cherchent plutôt à mesurer le poids de tel ou tel élément dans les processus de production et de compréhension des textes. Chaque élément est considéré comme un indice potentiel de l'organisation discursive. Ces indices sont nécessairement « textuels » (*i.e.* présents à la surface du texte) alors que les marqueurs discursifs, étant définis d'un point de vue uniquement fonctionnel, peuvent être non réalisés à la surface du texte. En effet, une relation de discours peut-être interprétée sans pour autant être explicitement marquée par un connecteur (voir en V.1.3 la notion de connecteur implicite).

Pour observer le signalement de l'organisation discursive, les études cognitives se basent très souvent sur des textes 'longs', par comparaison aux travaux en sémantique qui étudient des 'discours' constitués généralement de deux phrases. Deux méthodes sont généralement utilisées selon le point de vue adopté : étude de la production ou étude de la compréhension. Au niveau des processus de production, il s'agit d'étudier la présence ou absence de certaines traces des procédés organisationnels dans des productions humaines. L'étude ne se fixe plus sur quelques formes en particulier, mais sur certains processus organisationnels. En général, les cobayes humains doivent produire des

rédactions par rapport à un sujet bien défini et dont le cadre est orienté par une vidéo regardée ou l'évocation d'un souvenir. Schneuwly (1997), dont nous reparlons en V.1.3, observe comment le fonctionnement de certains organisateurs textuels (comme le connecteur *et*) évolue avec l'âge des locuteurs. La méthode utilisée pour étudier les organisateurs textuels se base sur l'hypothèse que les modes organisationnels s'acquierent et se spécifient selon certains types de texte petit à petit.

Au niveau des processus de compréhension, les tests de rappel permettent de mesurer l'influence de tel ou tel indice sur la bonne compréhension d'un texte. Il s'agit de faire lire à des cobayes humains (généralement des étudiants ou des écoliers) des textes avec et sans les indices étudiés. Ainsi, Lorch (1989) étudie à l'aide de tests de rappel l'effet des « signaux organisationnels » (« *organizational signals* ») sur la mémorisation de textes expositifs. Il a notamment démontré que les titres de sections augmentaient la mémorisation des textes, puisque les rappels étaient plus précis et plus complets lorsque le texte source était structuré par des titres. Gernsbacher (1990, 1996) cherche à mettre en corrélation des formes et des instructions de continuité (« mapping ») ou de déplacement (« *shift* »). Gernsbacher & Robertson (2002) se sont particulièrement intéressées à l'utilisation du déterminant *the*. En se basant sur plusieurs expériences psycholinguistiques (très clairement exposées dans cet article), elles ont démontré que l'article défini (comparé à l'article indéfini) facilitait la construction d'une représentation mentale et notamment le regroupement des informations autour de thèmes communs.

Gaddy *et al* (2001) proposent un intéressant panorama des différentes études cognitives portant sur la mesure de l'influence des indices textuels sur les processus de compréhension. Les « *text cues* » de Gaddy *et al.* (2001) sont des indices qui guident l'attention du lecteur durant le processus de lecture (Gaddy *et al.* 2001:89). Ils jouent un rôle au niveau des focus d'attention réalisés pendant les processus cognitifs mis en œuvre pendant l'acte de lecture. Cette étude se situe dans le Landscape Model (Van den Broek *et al.* 1996) qui voit la lecture comme un processus cyclique (un peu comme pour les piles d'attention de Grosz & Sidner ou les sous-structures du Structure Building Framework de Gernsbacher) où la mémoire de travail sert de tampon à l'information, l'information nouvelle écrasant l'information ancienne et ainsi de suite. Trois catégories d'indices textuels sont pris en compte : des indices linguistiques (marqueurs méta-discursifs, anaphore, etc.), typographiques (surlignement, mise en gras, etc.) et propres à la structure du texte (par exemple : les titres de sections). Les catégories d'indices distinguées par Gaddy *et al.* rejoignent approximativement les quatre catégories d'indices de séquentialité définis par Goutsos (1996), voir infra. (V.1.4).

Tous ces types d'indices tiennent effectivement un rôle particulier dans l'organisation du discours. Ils sont d'ailleurs tous représentés dans ce chapitre (voir sections suivantes). Cependant, comme nous l'avons précisé en introduction, nous ne nous situons pas totalement dans une approche cognitive de d'EOD. C'est pourquoi nous n'avons utilisé ni le terme de marqueur discursif, ni le terme d'indice textuel, même si ce dernier est le plus proche de notre définition des indices de séquentialité.

V.1.3. Vers une définition des indices de séquentialité

Notre définition des indices de séquentialité est très proche de celle des organisateurs textuels (« *Textual Organizers* ») de Schneuwly (1997), même si, d'un point de vue formel, il ne considère qu'un sous-ensemble des indices que nous prenons en compte (nous listons les différents indices de séquentialité pris en compte dans notre étude dans les parties V.2, V.3, V.4 et V.5).

Pour Schneuwly, les organisateurs textuels sont les traces d'opération de séquentialité. Il n'utilise pas le terme de séquentialité mais définit l'organisation discursive comme le résultat d'opérations de connexion et de segmentation –

entendez techniques de continuation et de déplacement ou de rupture (Schneuwly 1997:248). Ces traces peuvent avoir différentes formes, que l'on peut regrouper, pour le français – Schneuwly travaille sur le français – en trois groupes : (1) l'ensemble des conjonctions de coordination et de subordination ; (2) des adverbes ou locutions adverbiales caractérisés par le fait d'être non intégrés syntaxiquement, de ne pas avoir de définition grammaticale claire et d'être fréquemment positionnés en initiale ; (3) des syntagmes prépositionnels – SP de forme plus ou moins complexes.

groupe	***exemples donnés par Schneuwly***
(1) : conjonctions	*et, mais, car, lorsque, parce que, ...*
(2) : adverbes et locutions adv.	*tout à coup, soudain, premièrement, finalement, plus tard, d'une part, c'est pourquoi, ...*
(3) : SP	*à sept heures, pendant huit jours, au sud, après la tombée de la nuit, lors du passage de Luc à Berlin, un jour, le lendemain, ...*

Tableau V.1 : Les trois groupes d'expressions linguistiques pouvant être organisateurs textuels (Schneuwly 1997)

Ces trois groupes de formes se retrouvent dans notre liste des indices pris en compte dans cette étude. D'un point de vue plus sémantique, le groupe (1) correspond aux connecteurs, le groupe (2) à ce que nous appelons « organisateurs textuels » et le groupe (3) aux adverbiaux de *setting*[72].

Dans sa définition des organisateurs textuels, Schneuwly pose deux points qui sont essentiels dans notre étude : les organisateurs textuels sont à étudier dans leur façon de fonctionner ensemble et on ne peut étudier leur fonctionnement sans prendre en compte le type de texte étudié.

> "First, textual organizers have to be analyzed as wholes functioning together, as configurations marking essential aspects of a text or, to put it in another way, as traces of operations on different levels of the production of text. Second, these configurations differ strongly across different text types. [...] What a student masters is not the use of a particular textual organizer, but a certain way of acting verbally, of using a text type that is likely to be efficient in certain communication situations." (Schneuwly 1997:248)

En totale approbation avec Schneuwly, nous posons l'hypothèse que la signalisation des structures textuelles s'effectue par des configurations d'indices plus que par des marqueurs au sens noble du terme – on peut parler d'indices « *soft* »[73]. En effet, une fonction discursive n'est pas marquée explicitement par un trait linguistique mais relève plutôt d'une « influence conjointe de divers facteurs linguistiques » (Jacques & Rebeyrolle 2006), et les traits ou facteurs entrant dans la caractérisation d'un phénomène linguistique peuvent autant appartenir à la structure interne du texte qu'à son appartenance à un type. En d'autres termes, les configurations d'indices varient selon les types de texte (la définition du type d'un texte peut être caractérisée tant par sa visée discursive que par son format, sa longueur par exemple, voir I.4 et III.3.3.b). Le type de texte constitue donc un indice à part entière dans l'identification de ces configurations.

La conception du signalement de l'organisation discursive par des configurations d'indices explique pourquoi de nombreuses relations propositionnelles sont dites 'inférées', car non marquées explicitement par un marqueur. La continuité par défaut est le type même de relation de discours implicite, de même que la plupart des relations rhétoriques. Vergez-Couret (2007) qui s'intéresse au marquage de l'élaboration semble arriver à la conclusion que la relation d'élaboration est plus souvent 'implicite' que marquée par un connecteur explicite. La plupart des projets de modélisation des relations propositionnelles se sont trouvés confrontés à la réalité de la langue, ce qui a aboutit à la

72 Dans la suite de ce chapitre, une sous-section est consacrée à chacun de ces groupes.

73 Cette expression a été formulée par Marie-Paule Péry-Woodley lors de son discours d'introduction au colloque ISDD '06 (International Symposium Discourse and Document) tenu à Caen en juin 2006.

création de « connecteurs vides » (Danlos 2007) ou « connecteurs implicites » (dans le Penn Discourse TreeBank)[74]. Cependant, aucune étude à notre connaissance ne mesure dans quelles proportions le marquage de ces relations est implicite. Il est effectivement plus délicat de repérer et définir un 'marqueur' implicite qu'un marqueur explicite. Le projet du Penn Discourse Treebank construit des protocoles d'annotation afin de caractériser toutes les relations propositionnelles contenues dans un corpus. La première étape de ce projet a été de repérer tous les connecteurs présents dans ce corpus et de leur assigner une relation de discours telle que définie par la SDRT. L'étape en cours de réalisation consiste à caractériser les relations non marquées par un connecteur. Pour ce faire, ils demandent à leurs annotateurs d'insérer le connecteur qu'ils auraient utilisé pour expliciter la relation qu'ils détectent entre les deux phrases. Cette étape est évidemment plus longue (et plus dépendante des compétences de chaque annotateur ?) que la première puisqu'elle ne part pas de formes précises. Pour l'instant[75], seules 3 sections sur 25 ont été annotées au niveau des connecteurs implicites (les connecteurs explicites ont été caractérisés sur les 25 sections)[76]. Le champ d'investigation a été réduit aux relations discursives entre deux phrases non séparées d'un changement de paragraphes (alors que l'annotation des connecteurs explicites est réalisée dès que le connecteur est repéré, quelle que soit sa localisation – entre deux propositions, deux phrases, deux paragraphes, etc.) En 2006, 2003 occurrences de connecteurs implicites ont été repérées contre 18505 connecteurs explicites. Une petite multiplication permet de se faire une idée très approximative – car sans prise en compte des connecteurs implicites intraphrastiques et interparagraphiques – du nombre de connecteurs implicites sur le corpus entier.

Le signalement de l'organisation discursive a fait l'objet de nombreuses définitions aux limites rarement précises. Et pour cause, ce signalement ne correspond pas à un marquage *stricto sensu,* mais à la combinatoire de certains éléments dans certaines positions textuelles et dans certains contextes discursifs. Ces configurations se définissent empiriquement soit par observation du comportement discursif de certains éléments particuliers (comme l'observation du comportement en corpus du supposé marqueur de la relation cause-conséquence *donc*, *cf.* Bestgen *et al.* 2003, 2006), soit par exploration de ce qui existe dans certains contextes discursifs et positions définis. Ainsi, certains indices ni nécessaires ni suffisants apparaissent au fil des travaux, comme le fait d'être en dehors de la phrase, d'être issu d'un processus de grammaticalisation, d'être isolé par des pauses, de se situer en position initiale de phrases ou en séquence avec d'autres indices[77].

V.1.4. Différents types d'indices en position initiale

Les parties qui suivent présentent nos connaissances quant aux indices présents en position initiale. Comme nous l'expliquons au chapitre VI, ce ne sont pas des hypothèses quant aux corrélations forme/fonction qui dirigent nos analyses, mais les données mêmes. Nous n'étudions pas des formes en particulier mais prenons en compte, de façon exhaustive, tout élément apparaissant en position initiale. Nous réalisons donc ici un inventaire plus qu'un argumentaire, et ne mentionnons que superficiellement certains travaux portant sur les indices pris en compte. Certaines parties sont plus fournies que d'autres. Elles correspondent généralement à des indices sur lesquels il n'existe pas d'étude descriptive 'de référence' ayant couvert véritablement le sujet (c'est le cas des titres de sections et

74 Il s'agit généralement d'insérer entre deux propositions non reliées par un connecteur le connecteur « virtuel » le plus représentatif de la relation de discours interprétable entre les propositions.

75 Le rapport sur lequel nous nous appuyons date du 29/03/2006.

76 Nous sommes désolée du manque de précisions quantitatives sur le PDTB corpus, mais il est relativement difficile de cerner la taille du PDTB corpus et des sections qui y sont découpées. Nous supposons que les sections ont la même taille. Des recherches plus approfondies sur le site http://www.seas.upenn.edu/~pdtb éclaireront les curieux.

77 Cette liste est établie par Romera (2004).

des changement de paragraphes). Nous passons parfois très rapidement sur des formes ayant fait et faisant encore aujourd'hui l'objet de nombreux travaux et nombreuses polémiques. Une thèse ne pouvant couvrir tout ce qu'elle aborde, nous préférons généralement renvoyer à un ou plusieurs travaux que nous jugeons 'de référence'. Il est de fait certain que de nombreuses réflexions manquent à notre inventaire, mais notre but est plus de donner les bases nécessaires à la compréhension de notre exploration en corpus et de notre interprétation des données observées.

La catégorisation de nos indices peut se faire selon différents points de vue : du point de vue du modèle de la séquentialité (Goutsos 1996), d'un point de vue plus formel et structurel, et du point de vue de la Systémique Fonctionnelle. Notre distinction relève simultanément de ces trois points de vue : nous cherchons des indices signalant l'organisation séquentielle du discours, en nous basant sur des caractéristiques formelles et structurelles, et en considérant une organisation complexe jouant au niveau des trois composantes méta-fonctionnelles.

Dans son modèle de séquentialité du discours, Goutsos (1996) distingue quatre catégories d'indices de séquentialité (les éléments entre parenthèses ont été rajoutés par nous sauf ceux encadrés de guillemets) :

(1) des indices physiques tels que le changement de paragraphes et l'utilisation des parenthèses (la présence de titres) ;
(2) des indicateurs conventionnels tels que les marqueurs méta-discursifs (Teufel 1999) ou méta-descripteurs (Hernandez 2004) *(Pour résumer, Dans cet article,* etc.*)* et les marqueurs discursifs (*Mais, Alors, Cependant,* etc.) ;
(3) des « configurations formelles » comprenant l'encadrement du discours (les adverbiaux extraprédicatifs) ou la structure des phrases (utilisation de constructions spéciales, d'éléments détachés en initiale, choix des éléments en position Thème) ;
(4) la catégorie des indices de connexion (« *binding patterns* ») comprenant les indices de cohésion (expressions co-référentielles) ou les paires prédicatives (question-réponse, énumération).

Tous ces indices peuvent bien entendu apparaître en position initiale. Au sein des indices de la catégorie (1), nous trouvons les titres et les changements de paragraphes – le changement de paragraphes correspondant au début d'un paragraphe. Les indices de type (2) correspondent soit à ce que nous avons appelé des « adverbiaux textuels », soit aux connecteurs 'purs' (voir plus loin). En (3), nous trouvons les indices issus des stratégies d'encadrement (adverbiaux de *setting*), les constructions spéciales (qui concernent la position initiale en ce sens qu'elles impliquent un Thème spécifique – ThSpe, voir IV.4.2) et bien entendu le type d'élément choisi pour constituer le Thème topical – ThTop – de la phrase. Enfin, la catégorie (4) regroupe les différents indices signalant une quelconque continuation et notamment les indices de co-référentialité que nous étudierons particulièrement au niveau de la forme des Thèmes topicaux.

En dehors de la catégorie (1), tous les indices appartiennent à la composition de la phrase. D'un point de vue syntaxique, nous pouvons caractériser les différents éléments en position initiale par leur intégration syntaxique, *i.e.* leur dépendance ou indépendance vis-à-vis de la proposition principale. D'un point de vue plus structurel, nous distinguons les éléments 'détachés' en position initiale et les éléments sujets. La catégorie des éléments détachés en initiale – INIT – comprend ainsi tout ce qui précède le sujet grammatical, exception faite des connecteurs 'purs' (voir partie V.3.4) qui se trouvent à la limite du détachement.

Selon l'hypothèse de l'encadrement du discours, les éléments extra-prédicatifs (*i.e.* non intégrés à la proposition principale) peuvent étendre leur portée au-delà de leur phrase d'accueil, nous parlons alors d'**adverbiaux**.

> « le fait que les adverbiaux antéposés jouissent d'un potentiel cadratif revêt une importance particulière dans la mesure où ces expressions, jusque là non répertoriées dans les taxinomies de relations cohésives, constituent une des deux grandes familles de marques disponibles dans les différentes langues pour guider le lecteur dans l'accès à la cohérence du discours. » (Charolles 2003:45)

Dans de nombreux travaux de linguistique française, les notions de circonstant et d'adverbial sont synonymes (*cf.* Guimier 1993). Or, si l'on souhaite garder la notion de circonstant dans la composante idéationnelle (un circonstant réalise les circonstances du procès), on ne peut accepter cette synonymie, puisqu'il existe des éléments périphériques de phrase, des reliquats syntaxiques comme le dit Guimier, qui manifestent des éléments fonctionnant au niveau des deux autres composantes. Ainsi, des adverbiaux peuvent exprimer des modalités d'énonciation propres à la composante interpersonnelle (*Malheureusement, Bien évidemment, etc.*) ou des références méta-textuelles de l'ordre de la composante textuelle (*Dans la 3e partie, Premièrement, etc.)* Nous conserverons donc le terme d'adverbial pour faire référence aux éléments extra-prédicatifs, c'est-à-dire qui n'appartiennent pas à la structure syntaxique principale et ne dépendent pas d'éléments appartenant à cette structure principale.

Les **arguments inversés** (les compléments d'objets antéposés dans une construction à sujet inversé e.g. *À cet écart statistique s' ajoute le fait que le passage de la vie active à la retraite ne se fait plus aussi radicalement que par le passé. [ATLAS_1]*) et les **appositions** sont des INIT d'un type un peu particulier. La mise en initiale de ces deux types d'éléments a, c'est certain, une justification d'ordre discursive plus que syntaxique. Cependant, ces deux types d'initiales détachées n'ont jamais vraiment fait l'objet d'étude au niveau discursif. Ainsi, dans notre parcours de recherche, nous n'avons jamais porté d'intérêt réel aux appositions et avons souvent préféré considérer l'argument inversé comme un adverbial circonstanciel. Cependant, Combettes (2005) note que si elles suivent une stratégie textuelle particulière, les appositions peuvent accéder à une capacité de portée. Nos analyses nous font également croire que les appositions contribuent à l'organisation discursive (voir les chapitres VIII, IX et X). Du point de vue des arguments inversés, l'hypothèse de l'encadrement écarte systématiquement les éléments dépendants syntaxiquement de la catégorie des introducteurs de cadre (voir Fuchs & Fournier 2003 précisément sur ce sujet, voir partie V.4.4).

La figure V.1 résume notre catégorisation formelle et structurelle des éléments repérables en position initiale.

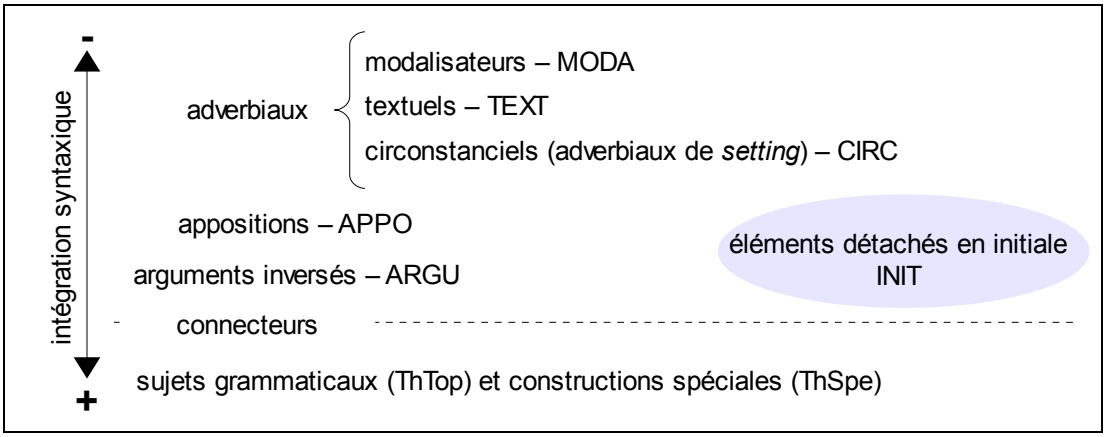


Figure V.1: Catégorisation structurelle des éléments observables en position initiale

Le troisième point de vue pour classer les indices pris en compte dans nos analyses est fonctionnel. C'est ce point de vue qui est adopté dans ce chapitre pour dresser notre inventaire. Les parties suivantes regroupent ainsi les différents indices selon leur rôle au niveau des trois composantes du discours : la composante interpersonnelle, idéationnelle et textuelle (chapitre II).

Mis à part les indices purement textuels (changement de paragraphes, adverbiaux textuels et connecteurs), nous avons fait figurer devant chaque autre catégorie d'indice le préfixe « texto », car il est évident que nos indices sont tous porteurs d'un sens instructionnel au niveau de la composante textuelle. Sinon, ils ne seraient pas dans cet inventaire.

C'est précisément ce sens instructionnel que nous cherchons à décrire. Les indices *texto*-idéationnels et *texto*-interpersonnels jouent sur deux composantes en même temps. La catégorie texto-interpersonnelle est constituée d'éléments exprimant un jugement du locuteur, une modalité d'énonciation, etc. Les éléments texto-idéationnels, en même temps qu'ils organisent le discours, apportent une information nécessaire à la construction de la représentation mentale. Nous trouvons dans cette catégorie des éléments qui réfèrent aux différents objets de discours impliqués dans les procès représentés, *i.e.* les INIT autres que MODA et TEXT et les ThTop. Ces éléments nous intéressent particulièrement pour leur capacité à exprimer les fondations et les entités constitutives du *text-world*, tout en constituant de bons indices de séquentialité.

V.2. Des indices multi-fonctionnels : les titres de sections

Une grande part de cette section est largement inspirée des publications écrites 'à six mains' avec Anne Le Draoulec et Marie-Paule Péry-Woodley pour l'article Hồ-Đắc *et al.* 2001 et avec Marie-Paule Jacques et Josette Rebeyrolle pour les différentes publications et communications résultant de nos agréables collaborations (Hồ-Đắc *et al.* 2004, Jacques *et al.* 2004)[78].

Nous avons à plusieurs reprises associé les titres à l'hypothèse de l'encadrement et notamment au phénomène de portée cadrative (II.2.2, II.3.2 et III.3.3.b). La définition du rôle des titres dans l'organisation discursive peut tout à fait correspondre à celle que nous avons faite de la position initiale. Les titres orientent la section à laquelle ils sont associés. Ils indiquent un déplacement ou une rupture dans le flot du texte, tout en ancrant la section dans l'organisation générale du texte.

Il existe assez peu d'études sur les titres de sections. La plupart des études linguistiques portent sur les titres d'oeuvre en général : titres de romans (Hoek 1981), titres de tableaux (Bosredon 1997), titres de presse (notamment Sullet-Nylander 1998, 2001). C'est dans le domaine de la psychologie cognitive que l'on trouve les quelques études sur les titres de sections. Ces recherches concernent l'impact des éléments de mise en forme matérielle (parmi lesquels figurent les titres de sections) sur la compréhension (Lorch 1989 ; Haladi *et al.* 2002 ; Lemarié *et al.* 2004). Ces dernières recherches vont dans le sens de l'hypothèse que nous formulons quant au rôle discursif des titres, bien que notre approche soit différente. Sachant que leur suppression nuit aux performances des lecteurs en terme de compréhension et d'accès à l'information dans les textes, on peut voir là un indice de ce que les titres de sections participent pleinement à la construction du discours. Nous supposons en effet qu'ils assument des fonctions de structuration et d'organisation du contenu du discours, et pas seulement un rôle d'organisation matérielle du texte. Les titres participent à la construction du *text-world* en tant que signaux textuels (au niveau de la composante textuelle) et en tant qu'unités dotées de contenu sémantique (au niveau de la composante idéationnelle). La métafonction interpersonnelle s'incarne également dans le fait que certains titres ont vocation à « accrocher » le lecteur, à établir un contact complice avec lui, à l'inciter à évaluer positivement le texte, à lui accorder plus d'attention qu'il ne l'aurait fait spontanément. Par exemple, le titre « *a. Au commencement était le droit* » *[GEOPO_12]* – paraphrasant une phrase célèbre – se veut un clin d'œil au lecteur, ce qui vise à établir une certaine connivence et attirer son attention.

D'un point de vue textuel, les titres de sections organisent matériellement le texte en constituant des repères dans le flot du texte. Les titres segmentent le texte, hiérarchisent les sections ainsi délimitées et fournissent une dénomination pour chacune d'entre elles. L'auteur ou le lecteur peut alors faire référence à telle section par son nom, autrement dit, son titre, comme nous l'avons fait quelque fois dans cette thèse.

78 Un article descriptif sur le fonctionnement des titres en discours est en gestation avec Josette Rebeyrolle et Marie-Paule Jacques.

Mais la plupart des titres ne se contentent pas uniquement d'apporter une information sur l'organisation matérielle du texte, ils jouent également un rôle au niveau de l'organisation du contenu du texte, de l'élaboration du *text-world* que celui-ci relate. Confrontées à la réalité de trois sous-corpus (dont le sous-corpus GEOPO utilisé dans cette thèse), Hồ-Đắc *et al.* (2004) distinguent trois modes de participation à la construction du *text*-world : une implication nulle au niveau de la composante idéationnelle (titres à implication zéro), un rôle au niveau de l'introduction des entités impliquées dans le *text-world* de la section (titres référentiels) et une participation à l'orientation thématique de la section (titres thématiques).

Les titres à implication zéro ne participent pas à la construction du monde du texte et n'ont de ce fait un rôle qu'au niveau de l'organisation matérielle du texte : ils marquent son découpage et sa hiérarchisation en sections (ex : *Introduction*). A l'inverse, les deux autres types de titres participent à la construction du monde du texte, mais dans des perspectives différentes. Alors que les titres à implication référentielle installent ou mettent le focus sur un ou des référents du discours dont on va parler dans la section titrée, les titres thématiques posent un cadre pour la section titrée, cadre dans lequel s'inscrit ce dont on va parler ou ce dont on parle déjà. Ainsi, dans la construction du *text-world*, les titres référentiels soulignent le sujet des propos à venir tandis que les titres thématiques posent les fondations qui situent le *text-world* relativement à un domaine d'activité, un domaine de connaissance, un point de vue, une situation spatio-temporelle, etc. Au niveau de l'interprétation, nous avons là encore deux processus différents : attirer l'attention du lecteur sur un ou des objet(s) du discours particulier(s) *vs.* canaliser certaines de ses connaissances d'arrière-plan (*i.e.* orienter l'interprétation de la section). Nous retrouvons ici la distinction entre « *figure* » et « *fond* » (III.1.2.b).

Dans nos analyses, nous nous intéressons particulièrement au fait que le titre ou un de ses éléments constitutifs soit ou non repris lexicalement dans la section qu'il titre. Le fait qu'un titre soit ou non repris nous renseigne sur son implication dans le texte mais également sur la construction même du texte. Jacques *et al.* (2004) remarquent par exemple que dans un corpus regroupant des textes présentant une structure titrée complexe, les titres de plus bas niveau présentent plus de reprises que ceux de haut niveau. Il semble que la fréquence de reprise informe sur le fonctionnement du titre dans le texte. Car l'usage des titres de sections varie d'un type de texte à l'autre, et dans de moindres mesures, d'un auteur à l'autre. Nous verrons ce que nous apprennent nos données sur ce terrain en X.5.

V.3. Des indices textuels

La catégorie des indices textuels rassemble des indices qui ne jouent qu'au niveau de la composante textuelle. Ils n'ont *a priori* pas de contenu propositionnel (composante idéationnelle) et ne jouent pas au niveau de l'implication des participants au discours (composante interpersonnelle). Nous trouvons dans cette catégorie des indices physiques et des indicateurs conventionnels (d'après la terminologie de Goutsos, voir supra). Nous verrons en particulier les marques typo-dispositionnelles de changement de section et de changement de paragraphes, ainsi qu'un ensemble de marques lexicales dont la référence concerne directement le texte (les marqueurs méta-discursifs) ou l'organisation discursive (les adverbiaux textuels et les connecteurs).

V.3.1. Le changement de section

Le changement de section est marqué par la présence d'un titre de section. Cette petite sous-partie rappelle simplement qu'en dehors de la présence en initiale de sections d'un titre, les sections réalisent un découpage du texte en gros segments dont le contenu ou la fonction est spécifié par le titre. Dans le MAT, il est généralement considéré que l'auteur divise son texte en parties, chapitres, sections, sous-sections, etc. Ce n'est qu'ensuite qu'il associe à telle

division un titre (Virbel 1986, Luc *et al.* 2001 et voir la partie II.2.1). Cette conception est sans doute à vérifier par des expérimentations psycholinguistiques.

Par cet acte de segmentation, nous pouvons poser que le changement de section indique une discontinuité dans le discours. Le titre, définitoire du changement de section, marque une rupture dans le flot du texte. Ainsi, il semble incongru de commencer une section par une reprise anaphorique ; et lorsque l'on lit un pronom en début de section[79], il se rapporte à une entité exprimée dans le titre et non dans la section précédente. D'un point de vue cognitif, le changement de section peut être associé à un nettoyage de la mémoire tampon, ce qui signifie que les entités actives en début de section sont nécessairement et uniquement celles réalisées dans le titre. Cela n'exclut pas, bien entendu, que les entités actives dans les sections précédentes restent accessibles. Mais généralement, l'expression de ces entités s'accompagne d'une expression-guide qui informe le lecteur de l'endroit où cette entité a été introduite (par exemple en notant entre parenthèses la partie concernée, *cf. partie X*, *voir partie X*).

Dans nos analyses, nous partons de l'hypothèse que le changement de section équivaut à une rupture même si, d'un point de vue sémantique, il s'agit d'un déplacement, nous supposons que les techniques utilisées relève principalement des stratégies de rupture. Les éléments observés en position initiale de sections (S1) seront donc de bons candidats à l'indication d'une rupture (ce qu'il faudra ensuite vérifier, voir VII.3).

V.3.2. Le changement de paragraphe

La fonction discursive du changement de paragraphes fait l'objet de relativement peu d'études en linguistique. Longacre a été un des premiers à considérer le paragraphe comme une unité grammaticale à part entière (Longacre 1979 : « The paragraph as a grammatical unit »). Son étude est plus une recherche des traits définitoires de l'unité paragraphe qu'une étude de sa fonction dans la construction des textes. Ce n'est que dans des études psycholinguistiques que l'on trouve des informations pertinentes quant à la fonction du découpage en paragraphes. Stark (1988) constitue la première étude expérimentale – et non théorique – de la fonction des changements de paragraphes (Stark 1988 : « What do paragraph markings do »). Nous relatons ici une grande partie de ce travail, difficilement consultable de nos jours. Ensuite, plusieurs études cognitives prennent le relais de Stark (Fayol 1997, Heurley 1994, 1997, Ulbaek 2001), mais les découvertes de Stark restent encore aujourd'hui d'avant-garde.

Les expérimentations psycholinguistiques consistent à mesurer l'effet du découpage en paragraphes soit au niveau des processus d'écriture, soit au niveau des processus de lecture. Dans le premier cas, il s'agit d'étudier le découpage effectué dans des textes produits par des locuteurs différents dans des conditions identiques et avec les mêmes instructions d'écriture (par exemple, écrire un texte permettant aux lecteurs de dessiner une figure géométrique plus ou moins complexe, Heurley 1994). Dans le second cas, il s'agit de confronter des lecteurs à des textes dans lesquels les marques de paragraphes ont été supprimées et de leur demander de les y replacer[80].

Dans tous les cas, les auteurs montrent avec certitude tant au niveau de la production que de l'interprétation : (i) que le découpage d'un texte en paragraphes est complexe à expliquer, (ii) qu'il peut considérablement varier selon les individus (iii) mais qu'il répond tout le temps à un souci de structuration.

79 Hồ-Đắc *et* al. 2004 dénombre 6 titres sur 1041 donnant lieu à une reprise pronominale. Dans notre corpus, nous observons 12 titres de ce type (voir X.5).

80 Pour une description de ces deux types d'expérimentations se reporter à Heurley (1997)

> "Paragraph divisions are based on a number of different criteria, and there must therefore be a variety of possible paragraphing for a single text. Yet it seems that paragraph boundaries are often suggested by the content of the text, which implies that they are not placed at random" (Stark 1988:285)

Dans son expérience, Stark (1988) montre qu'il y a un certain accord entre locuteurs concernant le découpage en paragraphes. Stark (1988) conclut d'une première expérience que les changements de paragraphes sont informatifs (*i.e.* ils ont un sens instructionnel). Lors de cette première expérimentation, plusieurs lecteurs avaient pour tâche de positionner des marques de paragraphe dans des textes où les changements de paragraphes ont été supprimés ou modifiés (ses expérimentations se basent sur des textes plutôt expositifs). Au final apparaît un manque de précision et d'accord entre les lecteurs et l'auteur. Le fait de ne pas pouvoir retrouver avec précision l'endroit où l'auteur avait mis un changement de paragraphes signifie que le changement de paragraphes indique quelque chose qui n'est pas indiqué par le reste du texte. La question est : qu'est-ce qu'il indique ?

Stark souligne que les changements de paragraphes ne facilitent pas spécialement la lecture. L'effacement des changements de paragraphes dans un texte, et même la modification des limites de paragraphe, n'a semble-t-il aucun effet sur l'appréhension du lecteur face au texte manipulé. Les lecteurs lisent les textes modifiés aussi vite et jugent leur cohérence, leur qualité et leur bonne forme à égalité.

> "Given the persistent intuition that paragraph markings make text easier to read, it is surprising that the current study provided no support for this idea. Reading speed and ratings of ease, coherence, and goodness were not affected by the presence or position of paragraph cues." (Stark 1988:297)

Par contre, la présence et la position des marques de paragraphe influence la compréhension, et c'est là l'information essentielle qu'apporte Stark. Tout d'abord, Stark suggère l'existence d'un lien entre la précision de la détection des changements de paragraphes et l'implication du découpage en paragraphes dans l'organisation du texte utilisé pour l'expérience. Ainsi, Stark remarque que dans le texte pour lequel il y a le moins de précision et d'accord, les initiales de paragraphes correspondent beaucoup moins à des discontinuités (d'ordre référentiel principalement).

Ce constat donne lieu à une deuxième expérimentation lors de laquelle les lecteurs doivent déterminer quelles phrases sont importantes dans des textes où l'on a enlevé ou déplacé les marques de paragraphe. Stark mesure alors combien de fois une phrase est jugée importante selon sa position dans le texte. Les résultats de cette expérience sont relatés dans les tableaux V.2 et V.3.

Position textuelle	***Phrases jugées importantes (%)***
début de paragraphe	*46*
milieu de paragraphe	*20*
fin de paragraphe	*20*
début de texte	*24*
fin de texte	*62*

Tableau V.2 : Pourcentage des phrases jugées importantes selon leur position textuelle (Stark 1988)

	changement de paragraphes	
Position de la frontière	***marqué***	***non marqué***
frontière d'origine	*46*	*27*
frontière arbitraire	*21*	*28*

Tableau V.3 : Effet du marquage du changement de paragraphes sur le pourcentage des phrases jugées importantes en initiale de paragraphes (Stark 1988)

Le tableau V.2 fait clairement apparaître que les phrases en début de paragraphe sont jugées plus importantes que les autres, exception faite des phrases de fin de texte qui, dans 62% des cas, sont jugées importantes. Lorsque l'on compare ces jugements selon les textes expérimentaux, on remarque que les phrases en début de paragraphe dans les textes où les frontières de paragraphe ont été modifiées perdent de l'importance. On passe de 46% des

phrases importantes à 21%. Cela signifie bien que ce n'est pas uniquement la marque de paragraphe qui attribue aux phrases qui l'accompagnent une certaine importance dans le processus de compréhension. Ce n'est pas non plus uniquement la composition de la phrase, puisque sans marque de paragraphe, le pourcentage tombe également en dessous des 30%. C'est donc l'association d'une certaine composition phrastique (présence de redénominations, d'adverbiaux, de constructions spéciales, de certaines expressions, etc.) et d'une certaine position textuelle qui présage de l'importance de telle phrase.

> "If the paragraph break occurs at an arbitrary place in the text, then the break doesn't make the initial sentence more important. Readers are not pursuing a blind strategy of judging paragraph-initial sentences to be important. Rather, **the effect of a paragraph cue is an interaction between the cue and the content of what is being cued**." (Stark 1988:297, nous soulignons)

Cette conclusion est en accord parfait avec les hypothèses qui sous-tendent notre méthodologie (voir chapitre VII). Elle rejoint également l'importance que Virtanen accorde à la position initiale de paragraphes.

> "A paragraph boundary may produce a highlighting effect on the paragraph initial sentence: this is where something starts. Signals of the temporal or locative TSCs may therefore be assumed to have the effect of separating what is to follow from what has gone before." (Virtanen 1992:226)

Nous retrouvons dans la citation de Virtanen notre définition de la position initiale en tant que point de départ. La position initiale de paragraphes semble constituer un indice fort dans le marquage de la séquentialité (Goutsos 1997 considère le changement de paragraphes comme un indice de transition), et cela à un niveau local comme à un niveau plus global. Comme le remarque Stark, toujours dans cet article de 1988, les discontinuités thématiques sont fréquentes mais pas obligatoires en initiale de paragraphes. Il peut donc y avoir continuité thématique d'un paragraphe à un autre. Mais cette continuité sera marquée par ce que Stark appelle une variation stylistique : une redénomination (« *over-reference* ») ou une construction divergente (Stark 1988:300). Dans ces situations, le paragraphe indique un regroupement des informations dans un bloc relativement à un critère autre que thématique (rhétorique, d'arrière-plan, etc.) Nous retrouvons ici l'hypothèse de l'encadrement du discours, les paragraphes pouvant alors être confondus à des cadres sous-spécifiés[81].

V.3.3. Les structures énumératives

Les structures énumératives, souvent appelées par raccourci énumérations (les énumérations ne sont en fait qu'une partie des structures énumératives, voir plus loin), ont fait l'objet du travail de thèse de Christophe Luc, dans le cadre du modèle d'architecture textuelle. Luc (2000) propose une description détaillée des structures énumératives. Les structures énumératives se reconnaissent par leur composition : la présence d'une amorce, d'une énumération et parfois d'une conclusion. L'amorce constitue une phrase introductrice, prévenant de l'arrivée d'une énumération et exprimant le critère de regroupement des différents items de l'énumération. L'énumération est composée d'un ensemble d'items (au moins 2) entretenant entre eux des relations diverses (relation de coordination, de subordination, de succession, etc. voir Luc 2000:103). Ce qui fait qu'un item est reconnu comme tel, c'est sa coénumérabilité, *i.e.* le fait qu'il apparaît dans un parallélisme structurel avec d'autres items. Ce parallélisme structurel peut être marqué typo-dispositionnellement par des sauts de lignes accompagnés d'un système de puces ou de numérotations et/ou lexico-syntaxiquement par des organisateurs textuels (autrement appelés Marqueurs d'Intégration Linéaire – MIL tels que *premièrement, ensuite, finalement, etc.* voir les travaux d'Agatha Jackiewicz) ou des parallélismes syntaxiques (voir la

81 Michel Charolles a souvent émis, lors de séminaires et colloques, cette hypothèse d'un rapprochement entre le fonctionnement des cadres et celui des paragraphes. Nous n'avons malheureusement pas connaissance d'un article publié dans lequel cette hypothèse était étayée.

thèse de Nicolas Hernandez (2004) à ce sujet). L'exemple V.1 montre une structure énumérative dont les items sont marqués lexico-syntaxiquement (en gras) introduits par une amorce : « *Les facteurs essentiels sont ici les suivants* ».

(V.1) ***LES ZONES CONTESTÉES DE LA DOMINATION DES ETATS-UNIS*** [titre niveau 1]

Les adversaires rencontrés par les Etats-Unis depuis 1990 se sont rarement montrés coopératifs. [...] Les cas de l'Irak, de la Serbie, de la Somalie, de l'Iran, les embuscades rencontrées en Afghanistan au cours de l'opération Anaconda montrent qu'il est possible de lutter militairement avec les Etats-Unis. Seuls les Somaliens peuvent revendiquer quelque chose qui ressemble à une victoire ; mais les autres ont imposé aux Etats-Unis des coûts inattendus, préservé leurs forces, et souvent survécu à l'affrontement jusqu'à pouvoir hélas colporter entre eux leurs recettes. Ces pays ou entités étaient petits, pauvres, et souvent très en retard militairement. Ces exemples appellent à la prudence. Les facteurs essentiels sont ici les suivants. ***En premier lieu****, la guerre a en général pour les acteurs locaux un intérêt politique de premier ordre, souvent bien plus important que celui des Etats-Unis. Leur tolérance à la souffrance est donc plus grande.* ***En deuxième lieu,*** *en dépit de leur taille réduite, ces acteurs supplantent d'ordinaire les Etats-Unis dans une ressource précise : le nombre d'hommes en âge de combattre. Même s'il n'est plus l'élément déterminant de la guerre terrestre, il reste un facteur critique, notamment en ville, dans la jungle ou en montagne.* ***Troisièmement,*** *les " locaux " disposent en général d'un avantage : ils jouent à domicile. Si les Etats-Unis ont constitué au fil des décennies la mémoire institutionnelle qui leur permet de maintenir leur maîtrise des espaces, les acteurs locaux ont fait un travail similaire sur leur propre pays. Ils connaissent intimement le terrain et la météo, et ont mis au point, sur des décennies, voire des siècles, des tactiques et des stratégies adaptées à leurs milieux.* ***Quatrièmement,*** *nombre des chefs militaires de ces Etats ou entités ont été formés dans le monde développé - pendant la guerre froide, la formation militaire fut souvent utilisée comme instrument d'influence politique. Ils ont appris les tactiques en vigueur en Occident, comme l'usage des armes occidentales, et les meilleurs d'entre eux peuvent tourner ces connaissances contre les Etats-Unis. Certains rapports montrent d'ailleurs que les adversaires des Etats-Unis ont échangé leurs expériences.* ***Cinquièmement,*** *l'arsenal nécessaire au combat rapproché, à terre, dans les airs à basse altitude ou dans les eaux territoriales est beaucoup moins coûteux que les armements nécessaires à la guerre dans les " espaces communs ". En outre, la diffusion des capacités économiques et technologiques civiles trouve son parallèle dans le domaine militaire : de nouveaux fabricants apparaissent, cherchant des débouchés à l'export, et l'arsenal pour le combat rapproché connaît un perfectionnement constant. Tous ces facteurs se renforcent et contribuent à créer une " zone contestée ". Dans une telle zone, les interactions entre les Etats-Unis et les forces locales vont souvent prendre la forme d'un véritable affrontement. Tout ceci n'annonce pas forcément une défaite américaine, mais nombre de difficultés.* *[GEOPO_10]*

A la lecture de cet exemple, la relation entre structure énumérative et indices de séquentialité en position initiale devient évidente. Qu'il s'agisse d'énumérations indexées par une puce ou par un MIL, la puce ou le MIL apparaît en position initiale de phrases. Dans notre mémoire de DEA (2000), nous faisions l'hypothèse que l'organisation discursive d'un texte pouvait être représentée par une structure énumérative globale, à l'allure d'une table des matières. Cette hypothèse pouvait également s'adapter à des zones plus locales. Ainsi, une TSC temporelle peut être représentée sous forme énumérative (comme nous l'avons fait avec l'exemple II.6 p.49).

L'importance que nous accordons à la position initiale est intimement liée à cette hypothèse de représentation. Chaque item d'une énumération correspond à une étape dans la présentation d'un fait. Quelles que soient les relations entre items (et elles sont complexes, voir Luc 2000), le changement d'item peut être associé à un déplacement.

Jackiewicz (2005) attribue aux organisateurs textuels la capacité (i) de découper et baliser des segments de texte, (ii) de regrouper différents éléments sur la base d'une propriété ou d'un ensemble de propriétés (« c'est généralement l'amorce qui typiquement exprime le principe qui fédère les items » *op.cit.* p.106) et (iii) d'ordonner les items en série. A part la dernière capacité, nous retrouvons dans cette définition fonctionnelle des organisateurs textuels tous les aspects de la segmentation textuelle telle que définie au chapitre III. Nous pouvons donc associer la présence d'une puce ou d'un MIL à l'indication d'un déplacement[82], au niveau intentionnel, idéationnel ou textuel, toutes les composantes semblent possibles.

82 La notion d'« étape » utilisée par Virtanen (1992) prend ici toute son ampleur, voir III.2.4.

V.3.4. À la limite du 'purement textuel' : les connecteurs et autres adverbiaux textuels

Les connecteurs et adverbiaux textuels (« *linking adverbials* » Biber *et al.* 1999) sont, comme leur nom l'indique, des éléments qui connectent deux unités entre elles. Les unités peuvent être de taille variable : des propositions, des phrases, des ensembles de phrases, des paragraphes.

De plus, un connecteur peut connecter une phrase à un ensemble de phrase, par exemple en fin de paragraphe ou de section comme dans l'exemple V.2 où le connecteur *néanmoins* contraste avec plusieurs phrases précédentes (à partir de la phrase mise en gras).

(V.2) *Le régime syrien est également conscient du fait que l'opposition à la tutelle syrienne ira croissant une fois le Liban débarrassé de l'occupation israélienne. Il procédera sans doute à une révision "rationnelle" de sa politique libanaise. Par la suite, le degré d'interférence de la Syrie dans les affaires intérieures de son petit voisin dépendra, en large partie, de la capacité des hommes politiques libanais à rompre avec cette tradition historique consistant à entraîner de façon active les acteurs extérieurs dans leurs conflits internes.* ***S'ils réussissaient à se prendre en main et à résoudre les problèmes politiques et sociaux de leur pays sans "assistance" étrangère, on pourrait s'attendre à ce que le redéploiement, sans cesse reporté, des troupes syriennes ait enfin lieu dans un contexte de "réduction graduelle" de la domination politique de Damas sur le Liban.*** *La réduction graduelle signifie que la Syrie veillera à maintenir une certaine influence sur les affaires politiques et de sécurité de ce pays, principalement en plaçant des hommes de confiance aux postes les plus sensibles au sein de l'armée libanaise et du Deuxième Bureau, en soutenant des forces politiques ayant fait la preuve de leur loyauté envers Damas et enfin en opposant son veto à l' ascension" politique de personnes connues pour lui être ouvertement hostiles et qui brigueraient des postes gouvernementaux haut placés. De plus, la série d'accords conclus entre les deux États et qui couvrent tous les domaines de la coopération (tels que la sécurité, le commerce, le travail, l'agriculture, la santé et le partage des eaux de l'Oronte) garantira la pérennité des intérêts de la Syrie au Liban et préservera le caractère privilégié de ses relations avec ce pays même avec la fin de son système de tutelle actuel.* ***Néanmoins,*** *une réduction plus hâtive et même désordonnée de la présence et de l'influence syriennes n'est pas à exclure au cas où se produirait un changement de régime brutal à Damas. [GEOPO_13]*

La définition des connecteurs tout comme celle des marqueurs discursifs est fonctionnelle, ce qui rend leur catégorisation formelle périlleuse. Cependant, identifier une liste fermée de connecteurs est nécessaire à notre étude. Nous nous intéressons principalement au signalement de l'organisation discursive (qui ne se joue pas uniquement au niveau des relations propositionnelles), notamment à celui des continuités et discontinuités au niveau idéationnel qui se construisent par des procédés d'orientation et de connexion (III.1.2.b). Or, les connecteurs ne jouent pas au niveau de la composante idéationnelle et ne permettent que rarement d'orienter l'interprétation de grandes zones de texte, leur rôle étant souvent limité au niveau l'inter-propositionnel. Il nous faut donc pouvoir identifier ces éléments afin de les distinguer des éléments initiaux tels que les INIT qui eux, peuvent jouer au niveau de la composante idéationnelle et peuvent porter sur des grands pans de texte, étant de bons indices de déplacement (à l'inverse des connecteurs, signes d'une continuité au niveau idéationnel). Ainsi, l'exemple V.3 montre un connecteur (*Cependant*) suivi d'un adverbial temporel *(fin 50 et début 49,)* qui pose un circonstant qui oriente le reste du texte en précisant que l'auteur va relater des faits ayant en commun la même localisation temporelle. Le connecteur joue à un autre niveau, explicitant la relation de contraste entre la phrase introduite par *Cependant* et le discours antérieur.

(V.3) Désormais, la guerre civile est inévitable, car César sait qu'il ne pourra pas abandonner ses pouvoirs et son armée avant d'être sûr de son élection comme consul pour 49 : telle est la fameuse "question de droit" de l'expiration des pouvoirs qui va empoisonner la vie politique romaine et mener à la guerre civile. César et Pompée se renvoient la balle, luttent par tribuns et consuls interposés, chacun refusant d'abandonner ses pouvoirs tant que l'autre les garde. Le Sénat est en fait favorable au départ des les deux protagonistes. ***Cependant, fin 50 et début 49,*** *les aristocrates confient à Pompée le soin de "défendre la République" et le commandement des légions en Italie. César rassemble ses troupes, envoie au Sénat un véritable ultimatum, et, lorsque les tribuns qui lui sont favorables doivent quitter Rome où la loi martiale est déclarée, il a le prétexte qu'il lui fallait pour franchir le Rubicon. Contre l'armée des Gaules, le Sénat et*

Pompée disposent des légions d'Espagne et des 130 000 hommes que Pompée est autorisé à lever en Italie. [PEOPL_6]

Notre délimitation de la catégorie des connecteurs s'est donc faite de façon empirique lors de l'élaboration de notre programme d'annotation. Face à la réalité des corpus, nous avons petit à petit identifié différents facteurs caractéristiques de ce que nous avons appelé les 'connecteurs purs'. Ces connecteurs purs sont des mots généralement seuls, délimités ou non du reste de la phrase par une virgule et situés en toute première position. Ils correspondent à ce que Quirk *et al.* (1972) définissent par le terme « *conjuncts* » dans leur grammaire anglaise :

> "Conjuncts, [...] serve a connector function within the flow of the text, signalling a transition between ideas.
>
> *If they start smoking those awful cigars, **then** I'm not staying.*
> *We've told the landlord about this ceiling again and again, and **yet** he's done nothing to fix it.*
>
> At the extreme edge of this category, we have the purely conjunctive device known as the conjunctive adverb (often called the adverbial conjunction):
>
> *Jose has spent years preparing for this event ; **nevertheless**, he's the most nervous person here.*
> *I love this school ; **however**, I don't think I can afford the tuition*." (Quirk & Greenbaum 1993:126 et http://grammar.ccc.commnet.edu/grammar/adverbs.htm)

Les connecteurs 'purs' correspondent à des conjonctions de coordination et des « conjonctions adverbiales » (*cependant, de plus, en conséquence, toutefois,* etc.) qui servent à créer des relations entre idées plus complexes que les conjonctions de coordination. La liste des connecteurs 'purs' repérés est donnée en annexe I. Les caractéristiques retenues pour leur repérage sont les suivantes :

- être le premier élément de la phrase, détaché ou non du reste de la phrase par une virgule et appartenir à la catégorie :
  - des conjonctions de coordination (excepté *ni*) ;
  - des conjonctions adverbiales : *d'abord, puis, pourtant, ensuite, enfin, certes, cependant, aussi, ainsi, néanmoins, alors, en effet, d'ailleurs, de plus* ;
- ou être une locution construite avec une conjonction de coordination suivie d'un certain groupe d'adverbes (la liste de ces adverbes est donnée dans l'annexe H (H.3.2))
- être le premier élément de la phrase nécessairement détaché du reste de la phrase par une virgule et appartenir à la catégorie des adverbes (exceptés ceux de forme -ment).

Comme le montre cette liste, nous rangeons les adverbes *puis*, *ensuite*, *enfin*, *alors*, *donc* dans la catégorie des connecteurs vu leur absence de contenu idéationnel. Cette position est critiquable si l'on considère les cas où ces adverbiaux dénotent une relation d'ordre temporel (*puis* = « après cet événement »). Cependant, ces cas sont très rares. Bras *et al.* (2001) montrent clairement que *puis* est essentiellement un connecteur et non un circonstant temporel. Borillo *et al.* (2004) renforcent cette conclusion en menant une étude comparative entre *puis* et *un peu plus tard* dans des textes narratifs. Cette dernière étude prouve que *puis* fonctionne bien au niveau de la structure argumentative et non temporelle, contrairement à *un peu plus tard* qui n'a pas cette possibilité de double jeu. Un travail identique sur l'adverbial *alors* est proposé par (Bras & Le Draoulec, 2007).

> "It is particularly interesting to notice that a priori temporally equivalent adverbials may have such different effect on Discourse Structure. On the one hand, we see that *puis* is a real discourse relation connective […]. On the other hand, (if) *un peu plus tard* is not a discourse relation connective". (Borillo *et al.* 2004:330)

Contrairement à ces connecteurs *temporels*, nous ne catégoriserons pas comme « connecteurs » les adverbiaux anaphoriques et déictiques *là*, *ici* et *maintenant*. Et ce pour les raisons inverses que précédemment : les cas où ils dénotent une localisation spatio-temporelle (*là* = à cet endroit et *maintenant* = de nos jours) sont plus fréquents que lorsqu'ils ne font que connecter.

Dans une autre catégorie, nous avons la notion d'**adverbial textuel**. Ce type d'adverbial, proche de la famille des connecteurs, permet d'ouvrir des espaces de discours. Ils peuvent correspondre à des marqueurs méta-discursifs qui situent un segment par rapport à sa localisation dans le texte (ex : *dans la partie X*), à des MIL qui balisent et ordonnent des items d'énumération (*tout d'abord, deuxièmement, finalement*) ou à des expressions autres qui permettent elles aussi de répartir les informations dans des blocs relativement à un critère organisationnel.

Cette notion d'adverbial textuel vient de la distinction faite par Charolles (1997) entre les univers de discours et les espaces de discours (voir III.3.3.a). Les introducteurs d'espaces sont définis par Charolles (1997:27) comme des expressions qui « portent sur les aspects métalinguistiques de l'énonciation (*bref, en somme, en un mot, en fin de compte,* etc.) ou de l'organisation du discours (*d'une part, d'autre part, premièrement, deuxièmement, d'un côté/de l'autre,* etc.) Ces expressions sont à même d'intégrer plusieurs propositions et contribuent bien évidemment au partitionnement de l'information. » La possibilité de portée est la principale différence entre les connecteurs et les adverbiaux textuels. Alors que les connecteurs 'purs' peuvent relier deux syntagmes ou deux propositions syntaxiquement dépendantes et, par conséquence, être situés à l'intérieur d'une phrase, les adverbiaux textuels sont généralement détachés en initiale du reste de la proposition. Dans leur grammaire anglaise, Quirk & Greenbaum (1993) parlent d'éléments « *Disjuncts* » en opposition aux « *Conjuncts* » (nos connecteurs 'purs'). Les éléments disjoints, à l'inverse des éléments conjoints, orientent le reste de la phrase.

> "A disjunct frequently acts as a kind of evaluation of the rest of the sentence. Although it usually modifies the verb, we could say that it modifies the entire clause, too." (Quirk & Greenbaum 1993:126)

La distinction entre connecteurs et adverbiaux textuels n'est pas toujours reconnue. Dans certains modèles, les connecteurs sont intégrés à la catégorie des adverbiaux circonstanciels, comme pour la terminologie et les principes d'identification de Guimier (1993). Dans d'autres modèles, les adverbiaux textuels sont intégrés à la catégorie des connecteurs, mais des connecteurs d'un type particulier puisque détachés du reste de la proposition. Mais malgré cette différence d'ordre syntaxique, la limite entre certains adverbiaux et certains connecteurs reste délicate à définir, surtout pour les adverbiaux portant sur les aspects métalinguistiques de l'énonciation (*bref, en somme,* etc.) Nous tenons à maintenir cette distinction car elle permet de faire une différence dans notre processus d'annotation entre des éléments fortement susceptibles d'avoir une portée cadrative et des éléments dont la portée se réduit à la zone de connexion entre deux phrases.

V.3.5. Des constructions de mise en arrière-plan : sur l'usage du On...

Certaines constructions syntaxiques ont la particularité de mettre en position finale le sujet réel de la phrase. Cette catégorie regroupe les constructions passives, les constructions impersonnelles et les constructions en *On...* La distinction entre constructions actives et constructions passives n'est pas retenue dans notre étude. Non pas que la forme passive ne peut avoir une incidence sur l'effet de séquentialité du discours. Mais son étude ne rentre pas dans notre champ d'investigation. Nous avons en effet circonscrit notre étude aux indices présents en position initiale. Or, que ce soit dans une construction passive ou dans une construction active, c'est un thème topical qui sert de sujet grammatical.

Par contre, les constructions en « *On... »* ou régies par un sujet de forme impersonnelle ont une incidence directe sur la composition de la position initiale, puisqu'elles positionnent en sujet grammatical un élément qui ne porte aucune valeur idéationnelle ou interpersonnelle. Les constructions impersonnelles sont présentées en V.5.2 puisqu'elle

semblent tout à fait indiquées pour exprimer un jugement de façon indirecte. Les constructions en « *On... »* sont présentées ici.

Lors de notre élaboration du processus d'annotation, certaines constructions fréquentes et pourtant rarement répertoriées dans les grammaires nous sont apparues : les constructions ayant pour sujet grammatical le pronom « *on* » ou « *nous* ». Notre première idée était d'étiqueter ces constructions comme étant des commentaires, *i.e.* un regard du locuteur sur ses propos, comme dans l'exemple V.4.

(V.4) *On imagine mal comment faire accepter de telles mesures aux parlementaires représentant les États concernés par cette rationalisation. [GEOPO_3]*

La grammaire méthodique du français de Riegel *et al. (*1994:143) définit, au niveau de la catégorie des modalisations, les commentaires comme des énoncés « par lesquels le locuteur exprime ses engagements à l'égard de ce qu'il est en train de dire ». Les exemples qu'il donne sont uniquement l'insertion d'un adverbial modalisateur en début de phrase ou dans une construction du type « *je pense franchement que ...* » qui se recoupe approximativement avec quelques uns de nos énoncés en *On...* comme celui donné en (V.4). Mais il est vite apparu que la dénomination « commentaire » ne correspondait qu'à une infime partie de ce type de construction.

Les exemples suivants montrent des constructions en *On...* que l'on ne peut pas véritablement associer à un commentaire[83].

(V.5) *On cite le chiffre de 1 milliard de dollars supplémentaire pour les programmes spatiaux, demandé cette année par rapport à le projet 2002, ce qui amènerait le montant total à 8 milliards de dollars. [GEOPO_3]*

(V.6) *On assiste peut-être enfin à la naissance de un débat de nature stratégique sur l'espace militaire. [GEOPO_3]*

(V.7) *On peut penser qu'il voudra sans état d'âme mettre fin aux dérives budgétaires qu'a connues l'agence récemment - notamment avec un dépassement de budget de 800 millions de dollars pour la station internationale. [GEOPO_3]*

(V.8) *On peut également supposer que les pressions politiques sur la Syrie se renforceront, notamment de la part de les États-Unis, pour un retrait ou un redéploiement significatif de ses troupes au Liban une fois le Hezbollah désarmé. [GEOPO_6]*

(V.9) *On dit qu'il existe au moins vingt-trois structures décisionnelles différentes en exercice au sein de l'Union, et les stratégies de lobbying doivent ainsi s'adapter à la fois à l'état du règlement et à l'équilibre des forces politiques en présence. [GEOPO_16]*

Dans ces exemples, le pronom o*n* permet soit une mise au passif du sujet réel dans des constructions où le passif serait difficilement acceptable (V.5), soit la présentation d'un fait (V.6 et V.9), soit une alternance au pronom impersonnel *il* (V.7 et V.8).

Cet effet peut également se faire par l'emploi du pronom *Nous* comme en (V.10).

(V.10) *Nous savons que Bartók a beaucoup composé dans son enfance et son adolescence, et que son entrée à l'Académie de musique de Budapest est marquée par un silence de trois ans. [PEOPL_15]*

Le « *on* » *est* souvent utilisé, en français usuel, en synonyme du *nous*. Cependant, dans les exemples de (V.5) à (V.8), le remplacement du « *on* » par un « *nous* » est tout à fait acceptable.

Tous ces exemples glanés dans notre corpus d'étude (nous avons recensés 836 constructions en On/Nous...) nous ont poussée à définir ces constructions d'un double point de vue très général : une définition formelle d'une part (le sujet grammatical de ces constructions correspond au pronom *on* ou *nous*) et fonctionnelle d'autre part (ces constructions permettent la mise en arrière-plan du sujet réel). Nous gardons bien évidemment à l'esprit le fait que la définition formelle peut également correspondre à des constructions méta-discursives comme en (V.11) ou des constructions présentatives en (V.12) et (V.13).

83 Bien que la plupart des exemples sont issus du sous-corpus GEOPO, les constructions en *on...* ne montrent pas de variations significatives selon les différents sous-corpus (voir VIII.5).

(V.11) *Nous examinerons plus loin les perceptions et les interprétations syriennes et israéliennes de ce 'rendez-vous manqué' avec l'histoire.* [GEOPO_4]

(V.12) *À l'extrême gauche de l'échiquier politique, on retrouve des opposants au retrait tels que le député Yossi Sarid (Meretz) – l'un des plus virulents critiques de l'opération 'Paix en Galilée' – qui redoute dans ce cas de figure un déluge de katioushas sur le nord d'Israël, contraignant l'armée israélienne à revenir en force au Liban en y lançant une invasion massive, terrestre et aérienne.* [GEOPO_4]

(V.13) *Jacopo Cicognini, dans un Convitato di pietra en prose (1650 ?), tout en conservant le dénouement fatal, en a surtout exploité les ressources burlesques. On lui doit notamment l'invention du 'catalogue' des innombrables femmes abusées par don Giovanni ; l'échange de vêtements entre le valet (devenu Passarino) et son maître ; enfin la dissonance finale entre le cri de don Giovanni saisi par l' Enfer et celui du valet réclamant ses gages.* [PEOPL_1]

Face à cette grande hétérogénéité, nous définissons ces constructions comme étant des constructions spéciales permettant la mise en arrière-plan du Thème et parfois la mise en avant d'une modalité d'énonciation ou d'une implication du locuteur.

V.4. Des indices texto-idéationnels

Cette catégorie d'indice rassemble des formes qui possèdent à la fois un sens instructionnel permettant de guider le lecteur dans la construction de son *text-world* et un sens propositionnel participant à la construction même du *text-world*. Nous trouvons dans cette catégorie des éléments détachés (les adverbiaux circonstanciels, les appositions), les différentes formes que peut prendre le sujet grammatical et l'utilisation de certaines constructions spéciales thématisantes ou focalisantes. Dans la terminologie de Goutsos, cette catégorie rassemble des « configurations formelles » et des indices de connexion.

V.4.1. Les adverbiaux circonstanciels – CIRC

Les adverbiaux circonstanciels sont des éléments extra-prédicatifs qui expriment les circonstances dans lesquelles les propos tenus sont jugés pertinents par l'auteur. Ces circonstances sont à même de poser les fondations (notamment spatiales et temporelles) d'un text-world (voir la partie II.1.3). Ainsi, les adverbiaux circonstanciels définissent les limites représentationnelles des univers de discours, nous parlions alors de setting.

La notion de setting selon Dik a également été présentée dans la partie IV.2 portant sur la fonction d'orientation propre aux éléments en position initiale. Les adverbiaux de setting correspondent à des expressions locatives qui situent selon un critère d'interprétation (un temps, un lieu, un domaine spécifique, une source d'énonciation particulière, une certaine thématique, etc.) les propos tenus. Ces expressions sont aussi appelées des "scene-setting topic" par Lambrecht (1994:118), "stage topics" par Erteschik-Shir (1997) ou encore "chinese-style topic" par Chafe qui les définit comme des éléments qui établissent un cadre spatial, temporel ou spécifique à l'intérieur duquel la prédication principale se déroule ("a spatial, temporal or individual framework within which the main predication holds", Chafe 1976:53).

Le fait que dans ces trois terminologies les *settings* soient associées à la notion de topique souligne le lien qui existe entre la position initiale et l'expression topique. Nous avons déjà souligné cette confusion dans le chapitre précédent, les notions de Thème et de topique étant parfois confondues. Cependant, si l'on considère l'exemple V.14, nous ne pouvons pas affirmer que le paragraphe est construit « à propos » des trois années allant de 1992 à 1995. Comme le souligne Charolles (2003:36) « les SP [introducteurs d'univers, comme ici *En trois années, de 1992 à 1995,*] sont instrumentalisés pour la répartition des informations et l'organisation du discours, ils ne constituent pas, à

proprement parler, le topique du texte, ils servent simplement, en coulisse, à ordonner les contenus communiqués [...] ».

(V.14) ***En trois années, de 1992 à 1995,*** *le paysage migratoire entre le Sud de l'Angleterre et le Nord-Ouest de la France s'est fortement modifié. Il existe une forte dichotomie spatiale en relation avec les mouvements de population. On remarque d'une part, des espaces où la dynamique migratoire est forte et d'autre part, des espaces moins concernés par ces mouvements. Les 43 départements et comtés pris en compte connaissent presque tous une progression de leur population. Ils peuvent cependant être divisés entre deux sous-groupes : [...]* *[ATLAS_1]*

Les références temporelles et spatiales constituent évidemment les composantes 'typiques' permettant de poser un univers de discours puisqu'elles forment habituellement l'ancrage 'par défaut' des objets du discours. Toutefois, comme le souligne notre définition de l'orientation par les *settings* (ou celle de Chafe), d'autres types de références peuvent introduire un univers de discours, c'est notamment le cas des références à un domaine de connaissances spécifique. Nous parlons alors de références « notionnelles ». Les adverbiaux notionnels introduisent des cadres définis selon un domaine d'activités et de connaissances spécifique (*Dans le domaine de la biologie*), un point de vue particulier (*en général*), ou encore un ensemble de concepts particuliers à un domaine précis (*En hôtellerie homologuée*). Cette terminologie se retrouve chez Le Querler (1993:178) qui parle de « cadrage notionnel » pour identifier ces cadres dont l'IC réduit le domaine dans lequel les propos tenus sont pertinents. Nous renvoyons également à la thèse de Denis Vigier qui parle de ces univers de discours sous l'appellation de « cadres praxéologiques » (Vigier 2003, 2004).

L'exemple V.15 montre deux adverbiaux circonstanciels notionnels qui, sous deux formes bien différentes, situent les informations apportées dans le domaine (1) du nucléaire et (2) de la biologie.

(V.15) *Le Conseil de sécurité des Nations Unies a créé en 1991 une Commission spéciale (UNSCOM) en charge du désarmement de l'Irak, tout en confiant le volet nucléaire à l'AIEA. Huit années plus tard, quel bilan peut-on dresser ?* ***En ce qui concerne le nucléaire,*** *l'AIEA estime avoir épuisé dès 1995 sa mission de destruction et d'enlèvement des matières prohibées.* ***Dans le domaine biologique,*** *l'UNSCOM a procédé à l'élimination de [...]*

Notons que le « En ce qui concerne le nucléaire », équivalent au «as-for» de l'anglais, est généralement considéré comme une construction thématisante, ou construction détachée. Dans ces constructions, l'élément détaché a pour rôle d'introduire le référent topique et non un simple critère d'interprétation qui oriente le reste de la phrase. Dans l'hypothèse de l'encadrement du discours, de telles constructions introduisent ce que Charolles (1997) appelle des champs thématiques c'est-à-dire des cadres introduits par un introducteur de type « concernant X », « à propos de X », etc. et dont l'élément introduit par l'IC (le « X ») constitue le topique de tout le cadre. Nous voyons bien à la lecture de l'exemple que le topique de la phrase n'est pas le nucléaire mais l'« AIEA ». Pour que la construction détachée introduise le topique de la phrase, il faut que celui-ci soit repris dans la proposition principale, comme dans « En ce qui concerne le nucléaire, il pollue autant que... ».

Dans l'exemple V.15 comme dans celui qui suit, nous voyons bien qu'il y a similarité entre des introducteurs de cadre de forme « en X » (généralement associés à l'ouverture d'un cadre notionnel) et des introducteurs de cadre de forme « concernant X » (généralement associés à l'ouverture d'un champ thématique).

(V.16) ***LE TOURISME EN 1998*** [titre niveau 1]
Au niveau national [titre niveau 2]
La zone transmanche [titre niveau 2]
En hôtellerie de plein air, *les régions Bretagne et Pays-de-la-Loire comptent le plus de nuitées, plus du quart sont étrangères sauf en Ile-de-France où la clientèle est essentiellement étrangère, les enquêtes montrent que les français préfèrent se loger chez un ami ou dans la famille.*
Concernant l'hôtellerie homologuée, *l'Île-de-France avec Paris concentre le plus grand nombre de nuitées (52 millions), les autres régions comptent entre 2 et 6 millions de nuitées. La clientèle est plus internationale dans le Nord-*

Pas-de-Calais qu'en Bretagne et Pays-de-la-Loire.
***En général,** on peut distinguer trois types de régions [...].[ATLAS_3]*

Cette similarité nous permet de fusionner la catégorie des champs thématiques à celle des cadres notionnels. Nous avons ainsi trois types d'univers de discours : les univers temporels, les univers spatiaux et les univers notionnels.

L'importance des adverbiaux circonstanciels dans l'organisation du discours n'est pas uniquement limitée à l'hypothèse de l'encadrement du discours. Si l'on considère la dimension cognitive de telles expressions (i.e. la mise en place d'une représentation ou d'une sous-représentation), il est clair que les adverbiaux de *setting* peuvent indiquer un déplacement au niveau d'une continuation idéationnelle. De plus, s'ils apparaissent en séquence, les adverbiaux circonstanciels peuvent tout à fait marquer une TSC relative au type de circonstances exprimée (une TSC temporelle si les circonstances organisatrices sont temporelles, une TSC spatiale si elles sont spatiales, une TSC thématique si elles sont notionnelles[84]).

> "Sentence-initial adverbials of time and place in narrative and descriptive texts tend to form chains of text-strategic markers which have two basic functions in the discourse. They help to create coherence and at the same time they signal text segmentation." (Virtanen 2004:82)

Les adverbiaux circonstanciels sont donc des éléments extra-prédicatifs en position initiale dont la fonction principale est de poser les circonstances qui vont orienter l'interprétation de la suite du discours. Les circonstances peuvent correspondre à différents « **rôles sémantiques** ». La liste des rôles sémantiques circonstanciels n'est pas fermée et ne met en accord aucune grammaire. Cependant, certains rôles semblent sûrs, comme les circonstants de temps et les circonstants de lieu. A ces deux rôles, nous en ajoutons un troisième, celui des circonstances notionnelles. Les autres rôles sémantiques (manière, condition, conséquence, cause, but, etc.) ne seront pas distingués dans notre annotation. D'une part, un découpage trop détaillé des différents rôles sémantiques rendrait les quantités d'observables trop faibles pour effectuer des analyses statistiques pertinentes (voir les chapitre VI et VII). D'autre part, les rôles sémantiques cités entre parenthèses ne peuvent pas réellement poser les *settings* d'un *text-world* (à part les circonstants de condition, que nous avions annotés au départ, mais dont la quantité était vraiment trop faible pour les retenir). D'un point de vue représentationnel, il semble en effet difficile de fonder une sous-représentation mentale sur une manière, une cause, etc., surtout si l'on se situe dans l'étude de textes expositifs (notre découpage aurait certainement été différent dans des textes narratifs ou procéduraux).

V.4.2. Les appositions – APPO

La construction appositive ou « construction attributive détachée » (Riegel *et al.* 1994) est définie par Neveu (1998 :67) comme une construction syntaxique articulant nécessairement deux constituants : un segment support et un segment apport. Le lien entre les deux segments est réalisé par l'incidence du segment apport sur le segment support, le premier apportant une information supplémentaire sur le second. Les appositions que nous considérons correspondent à ce que Combettes appelle des « constructions détachées descriptives » (Combettes 1998, 2005). Dans l'exemple V.17, le segment apport que nous appelons apposition correspond au segment « *trop occupé par la politique de l'immédiat* ». Celui-ci réalise une prédication seconde, *i.e.* en arrière-plan de la prédication principale

84 Nous utilisons ici le terme « TSC thématique » pour souligner la possibilité d'avoir deux niveaux thématiques. Un niveau topical qui correspond au niveau phrastique (ce à propos de quoi est la phrase) et un niveau que nous qualifions de thématique qui correspond à un niveau plus global (la thématique dans laquelle se situe les phrases qui suivent).

(*César n'a pas le temps de...*) une seconde prédication est exprimée : *César est trop occupé par la politique de l'immédiat*.

(V.17) ***Trop occupé par la politique de l'immédiat****, César n'a pas le temps d' organiser de façon systématique ses pouvoirs, d'autant que jusqu'à la fin subsiste, dans le camp adverse, une "légitimité" qui conteste la sienne. [PEOPL_]*

Il existe également des constructions détachées non descriptives telles que « *à Paris il s'est mis à fumer* » qui situent le sujet grammatical (voir Combettes 2005:33). Ces constructions sont la plupart du temps considérées dans notre thèse comme des circonstants (voir le chapitre VII et la description du programme de caractérisation en annexe H). Cette distinction entre circonstants et appositions est généralement reconnue et repose sur la relation entre l'élément détaché et la proposition principale. Le segment appositif détaché est en relation d'identité référentielle avec le segment support et donc ne sert pas à son identification référentielle comme le font les circonstants en restreignant les conditions sémantiques de sa réalisation.

> « Les prédicats spatiaux du type *rue Saint-François de Paule* ou *à l'Université de Cologne*, et les prédicats temporels du type *en l'année 1930-31* ou *(le) mercredi 3 janvier*, doivent en toute rigueur être tenus pour étrangers à cette catégorie, puisqu'en dehors du détachement, qui est ici aléatoire et peut ne pas être réalisé, ils ne partagent aucun trait fonctionnel avec les segments appositifs : absence de co-référence, absence de fonctionnement incidenciel de type adjectival et donc absence de support actanciel. Ils manifestent au contraire clairement une incidence de type adverbial, c'est à dire une incidence externe de second degré. » (Neveu 1998:75)

Les appositions sont considérées comme « le résultat du détachement de constituants étroitement liés à la prédication principale, dans une relation du type attributif. Une gradation s'établit entre les trois énoncés : *il est sorti furieux -> il est sorti, furieux -> furieux, il est sorti*, gradation qui conduit d'une prédication principale à une prédication seconde ; l'antéposition – pour des raisons qui relèvent davantage d'une problématique stylistique – si elle conduit à la formation d'une prédication seconde et rend ainsi autonome la forme adjectivale, n'entraîne pas, en raison du sémantisme du verbe principal, la création de relations circonstancielles et conserve au groupe antéposé une valeur attributive » (Combettes 2005:37).

Les appositions permettent donc de réaliser une prédication seconde (d'arrière-plan) qui spécifie certaines caractéristiques de l'entité exprimée dans le segment support.

> « Le segment appositif détaché fournit un prédicat second, relativement mobile, greffé sur tout ou partie de la prédication première, et dont la contribution sémantique, extrêmement variable, […] *a priori* ne peut pas modifier la vériconditionnalité de la prédication d'ancrage » (Neveu 1998:68)

Les types syntaxiques nous permettant de caractériser les appositions sont les mêmes que ceux retenus par Neveu :

(i) des constructions substantives de forme SN (Det+nom, nom, Det+NP, NP). Riegel *et al.* (1994:191) note que les « appositions nominales suivent toujours leur GN de rattachement sauf celles qui sont dépourvues de déterminant et qui peuvent précéder le GN sujet.

Enfin, maîtresse des échéances diverses, Damas peut, tour à tour, les utiliser pour renforcer et consolider ses instruments de contrôle (...) ou, au contraire, pour les geler et les annuler afin de éviter des tests difficiles ou gênants selon la conjoncture (...) [GEOPO_6]

(ii) des constructions adjectives :

- soit des syntagmes adjectivaux – ADJ

Amoureux précoce, il épousa à dix-huit ans, en novembre 1582, Anne Hathaway, de huit ans plus âgée que lui, qui lui donna une fille, Suzanne, le 26 mai de l'année suivante [PEOPL_8]

- des constructions participiales avec participe passé – PPA ou avec participe présent – PPR.

Devenu Premier ministre, Shimon Pérès était mû par un sentiment d'urgence ou Parlant couramment anglais, il fut représentant de Philips pendant de nombreuses années [GEOPO_4]

- des SP

En péril d'hérésie, il est pourtant trop brillant pour n'être pas encouragé, distingué, et devient préfet de la confrérie de la Vierge [PEOPL_2]

- des SN

L'air hagard, l'autre semblait indifférent à la scène. [exemple de Neveu]

Les appositions ne peuvent vraisemblablement pas présenter une portée. En effet, leur incidence se focalise sur un segment support précis de la phrase. Elles ne font pas partie des adverbiaux ou modifieurs de phrase. Concernant le marquage de la séquentialité du discours, les appositions indiquent plus une continuité thématique qu'une discontinuité.

> « Le fonctionnement des constructions détachées doit être rapproché d'un rôle de lien référentiel. En effet, le sujet de la prédication seconde coïncide souvent avec le référent thématique du contexte antérieur ; la construction détachée apparaît ainsi comme une sorte de constituant 'intermédiaire', qui maintient l'identité référentielle tout en introduisant une nouvelle caractérisation. » (Combettes 2005:35)

Cependant Combettes (2005) remarque que certaines appositions peuvent avoir une portée cadrative (et non sémantique), comme dans l'exemple qu'il fournit :

Buffalo Grill traverse une période vaches maigres. Sur les dix premiers mois de l'année, près du quart du chiffre d'affaire est parti en fumée (...)
***Peu habituée à être exposée**, cette société, qui n'avait jusqu'à présent (...), se retrouva soudain au coeur d'une tempête juridico-médiatique. Elle tenta, non sans maladresse et cafouillage, d'établir un contre-feu (...)*
***Soucieuse de redresser son image**, Buffalo Grill a demandé à sa nouvelle agence de communication de mener une enquête (...) D'où il ressort que les traces de la crise sont profondes. (...) [exemple issu de Combettes 2005:42]*

Dans cet exemple, effectivement, les deux appositions permettent de renvoyer à deux aspects successifs de l'événement : le « manque d'expérience » puis l'envie de réagir. Il faut remarquer que cet effet va de pair avec un changement de paragraphes (indice de déplacement) et deux expressions co-référentielles marquées (indices elles aussi de déplacement) : un SN démonstratif permettant une reformulation du topique *Buffalo Grill* et une redénomination.

V.4.3. Des instructions dans les expressions (co-)référentielles

Il est généralement admis qu'il existe des corrélations entre la forme d'une expression référentielle et l'accessibilité de l'entité à laquelle elle réfère. Plus encore, aux différentes formes correspondent différentes instructions quant à cette accessibilité. Ainsi, le locuteur a à sa disposition une palette de formes. Utiliser une forme associée à un degré d'accessibilité relativement bas pour exprimer le topique permet alors d'indiquer un déplacement dans une continuité référentielle.

De nombreuses études en linguistique cognitive cherchent à établir des échelles de marquage de l'accessibilité des référents. Il s'agit de classer les expressions référentielles selon le statut cognitif du référent au moment où l'expression le désigne. Aux deux pôles de ces échelles, nous trouvons d'un côté des expressions désignant des référents totalement accessibles car actifs dans le *text-world*, et de l'autre des expressions désignant des référents totalement nouveaux (et donc inaccessibles) pour le *text-world*. Prince (1981) est la première à proposer une telle échelle d'acceptabilité des Topiques. Nous avons présenté ce type d'échelle en II.1.2. Le classement des centres anticipateurs de la théorie du centrage pour prédire le centre préféré susceptible de faire le lien avec la phrase suivante se base sur ce même type d'échelle (III.3.2)[85]. Le critère de base du classement du Centrage est la fonction syntaxique.

85 Prince fait partie de cette équipe de chercheurs à la base de laTthéorie du centrage (Marilyn Walker, Aravind Joshi & Ellen Prince).

Pour la plupart des échelles d'accessibilité, le critère de classement réside dans la forme de l'expression référentielle : dans le type de déterminant, la complétude du syntagme et la distinction entre forme pronominale et forme pleine. Deux échelles sont plus couramment utilisées : l'échelle relative à la Hiérarchie du Donné (Gundel *et al.* 1993) et l'échelle de la Théorie de l'Accessibilité d'Ariel (1990). Dans ces deux théories, un lien certain est établi entre la forme et le statut cognitif. L'idée générale est la suivante : telle forme va guider l'interprétation du lecteur en lui indiquant le statut cognitif du référent désigné.

> « les différents déterminants et les différentes formes pronominales signalent des informations distinctes à propos de l'état de mémoire ainsi que de l'attention (statut cognitif), en tant qu'elles font partie de leur sens conventionnel. » (Gundel *et al.* 2000:82)

Selon ces théories, un pronom va automatiquement désigner un référent très accessible. Pour la Théorie de l'Accessibilité, il désigne un référent hyper-accessible. Pour la théorie de la Hiérarchie du Donné, il correspond à un référent actif et sur lequel est mis le focus d'attention. Nous ne nous étendrons pas davantage dans la description de la Hiérarchie du Donné puisque nous utilisons, dans cette thèse, la Théorie de l'Accessibilité dont l'échelle du marquage de l'accessibilité (du degré d'accessibilité le plus élevé au moins élevé) est le suivant :

> Extremely high accessibility markers (gaps, PRO and *wh* traces, reflexives, and agreement) > Cliticized pronoun > Unstressed pronoun > Stressed prnoun > Proximal demonstrative (+NP) > Distal demonstrative (+NP) > Proximal demonstrative (+NP) + modifier > Distal demonstrative (+NP) + modifier > First name > Last name > Short definite description > Long definite description > Full name > Full name + modifier

D'après Ariel (1990, 2001, 2004), le marquage de l'accessibilité est universel*, i.e.* la correspondance entre une forme et un degré d'accessibilité est le même dans toutes les langues. De ce point de vue, les différences entre les langues ne se situent pas au niveau des correspondances mais au niveau de l'existence ou non des formes en question. Donc, quelle que soit la langue, l'usage d'un pronom marquera un degré d'accessibilité plus grand que l'usage d'une description définie complète (supposition qui semble tout à fait plausible).

Dans notre travail, nous devons adapter l'échelle selon différents critères :

1. la possibilité d'avoir telle forme en position sujet, à l'écrit, en français (élimination des pronoms réfléchis et toniques, des gestes),
2. la possibilité de repérer et caractériser automatiquement ces formes (élimination des anaphores zéro),
3. la lisibilité des résultats et donc le choix pour des catégories permettant des analyses comparatives*, i.e.* présentant suffisamment de données pour permettre des analyses statistiques.

Ce qui aboutit à l'échelle d'accessibilité donnée en figure V.2.

Pronoms et SN possessifs	7	Forte accessibilité Λ
Démonstratif (+SN)	6	|
Démonstratif (+SN) + modifieur	5	|
Nom propre repris	4	|
Description définie réduite ou reprise	3	|
Description définie complète	2	|
Nom propre nouveau	1	V
Description indéfinie	0	Faible Accessibilité

Figure V.2 : Échelle du marquage de l'accessibilité d'Ariel adaptée à l'étude du français écrit

Nous supposons la même idée qu'Ariel : l'idée que les expressions référentielles portent une instruction procédurale. Cependant, nous ajoutons à cette idée un facteur textuel : les sens instructionnels varient selon la

situation discursive (genre de texte, type de texte, stratégie textuelle employée). Cette hypothèse est véritablement confortée par l'étude de Catherine Schnedecker (2005) qui montre que dans des portraits journalistiques, le sens instructionnel du nom propre est particulier, puisqu'il propose une alternance au pronom personnel de troisième personne.

L'étude des corrélations entre forme et degré d'accessibilité permet d'établir des échelles de marquage de l'accessibilité des référents. Ces échelles ne prennent pas en compte l'idée de configurations d'indices. Leur conception considère la forme d'une expression comme le marquage même du degré d'accessibilité. De plus, les classements effectués restent encore à l'état expérimental. D'ailleurs, des études récentes défendent de plus en plus l'idée que la forme n'est qu'un facteur dans le marquage de cette accessibilité. Reboul (1997) récuse fermement la théorie d'Ariel en insistant sur le fait que c'est essentiellement la fonction référentielle des expressions qui permet d'établir l'échelle d'accessibilité et non l'inverse (*i.e.* une instruction relative à l'accessibilité du référent 'inscrite dans' les expressions référentielles qui explique les différentes façon de référer). La position textuelle et le type de texte semblent grandement influencer les corrélations mises en place. Schnedecker (2005) témoigne tout particulièrement de l'influence de la variation langagière dans l'usage des noms propres répétés.

Nous gardons donc des réserves quant à l'usage de telles échelles. Ainsi, pour notre étude, nous nous appuyons conjointement sur les degrés d'accessibilité et sur les différentes fonctions que peuvent jouer certaines expressions référentielles dans le signalement des (dis)continuités. Nous distinguons dans les expressions référentielles trois caractéristiques pouvant avoir une incidence sur le signalement de la séquentialité :

1. le fait qu'elles soient une pronominalisation ;
2. le fait qu'elles constituent une reprise lexicale ;
3. et dans le dernier cas, leur détermination.

V.4.3.a) Pronominalisation

L'indication d'une continuité référentielle se fait essentiellement par des expressions co-référentielles, *i.e.* qui réfèrent à un même référent. Parmi les techniques de co-référence l'anaphore est celle qui, certainement, a suscité le plus grand nombre d'études (nous renvoyons aux différents travaux de Georges Kleiber ainsi que ceux de Francis Cornish, Michel Charolles ou Francis Corblin). L'anaphore correspond à une relation de continuité référentielle qui s'effectue par un phénomène de « saturation ».

> « Une expression est anaphorique si une part de son interprétation est une valeur non fixée, qui requiert d'être identifiée à une valeur du même type fournie par son contexte d'usage. La fixation de cette valeur sera appelée saturation. » (Corblin, 1995:137)

Les formes les plus 'efficaces' pour co-référer sont les formes pronominales. Les pronoms personnels de troisième personne et les pronoms démonstratifs permettent de marquer un référent déjà saillant comme étant un point d'ancrage.

> « À l'instar du pronom *il*, le pronom *ça* constitue généralement une trace de point d'ancrage. Le démonstratif se distingue toutefois du pronom personnel, d'une part, parce qu'il renvoie à un référent saisi comme une entité no nommée, et, d'autre part, parce que son fonctionnement indexical confère une saillance particulière au référent auquel il renvoie. » (Grobet 2002:162)

L'utilisation d'un pronom est un indice très fort d'une continuité, mais pas de n'importe quel type de continuité, d'une continuité topicale. Il est impossible qu'un pronom indique une discontinuité.

> « Il n'y a pas seulement réalisation d'une co-référence à moindre frais. *Il* n'est pas utilisé uniquement pour indiquer qu'il s'agit de Fred [dans l'exemple V.18], mais avant tout pour marquer un fait crucial de

> cohérence : que l'on va (continuer de) parler d'un référent déjà saillant lui-même ou présent dans une situation saillante et que l'on va en parler en continuité avec ce qui l'a rendu saillant. » (Kleiber 1994:99)

Pour bien illustrer l'indication très forte d'une continuité topicale que porte le pronom *il*, Kleiber compare l'exemple :

(V.18) *Fred enleva son manteau. Il avait trop chaud.*

à celui-ci où l'on observe une reprise lexicale complète (une redénomination) :

(V.19) *Fred enleva son manteau. Fred avait trop chaud.*

et explique que dans (V.19), la redénomination *Fred* « ne peut que redonner le référent comme en première mention et conduit à l'effet contraire de celui qu'accomplit le pronom : même si la co-référence est maintenue, la continuité se trouve en quelque sorte rompue et le second énoncé *Fred avait trop chaud* n'a plus besoin d'être une suite de la situation « Fred enleva son manteau », puisque Fred n'est pas présenté la deuxième fois comme étant engagé dans une telle situation. » (Kleiber 1994:100). Cette explication nous permet de passer au deuxième grand groupe des expressions co-référentielles : les expressions avec reprise lexicale.

V.4.3.b) La reprise lexicale

Comme l'a très clairement expliqué Kleiber, les expressions co-référentielles avec reprise lexicale ne sont pas des indices de continuité 'pure', puisqu'elles 'n'emportent pas' avec elles l'indication d'une continuité au niveau de la situation d'énonciation. En d'autres termes, une reprise lexicale peut indiquer un déplacement au niveau des *setting*. Dans notre terminologie, nous pouvons attribuer aux reprises lexicales une certaine capacité à indiquer un déplacement.

Au niveau cognitif, il est généralement admis qu'une redénomination entraîne une charge de travail plus importante qu'une pronominalisation.

> « L'emploi en position sujet dans un énoncé non-initial au sein d'un segment[86], d'un nom propre répété codant un référent dont le niveau de focus est élevé (c'est le Cr de cet énoncé) donne lieu à des temps de lecture nettement plus grands par rapport à ceux enregistrés lors de l'emploi d'un pronom dans la même position. » (Cornish 2000:19)

Cette remarque doit cependant être pondérée par le fait que, dans les cas où la première mention n'est pas expressément saillante, il est moins coûteux d'avoir une redénomination. Ainsi, dans l'énoncé :

> *Berthe a giflé Paul. Il avait à nouveau menti [énoncé donné par Kleiber 1994:84]*

Il semble que l'énoncé :

> *Berthe a giflé Paul. Paul avait à nouveau menti.*

soit lu moins vite que le précédent, car *Paul* est moins saillant que *Berthe*. Le temps de lecture plus long lorsqu'il y a redénomination s'explique uniquement lorsque le référent antécédent est saillant, comme l'explique Cornish (2000:20).

> « Une façon d'expliquer les temps de lecture plus élevés dans la condition 'nom-nom' serait de poser l'existence, chez l'allocutaire, de l'une ou l'autre des deux démarches suivantes : étant donné que le pronom attendu codant le Cr n'a pas été employé, le sujet mobiliserait des ressources cognitives supplémentaires, soit afin d'instancier un nouveau référent ayant le même nom que celui du référent topical ; soit en interprétant cette répétition comme un signal pour ouvrir un nouveau sous-topique, une

86 Cornish fait ici référence aux segments de discours tels que le conçoit le modèle de Grosz & Sidner (1986) et que Cornish (2000:9) définit comme suit : « Le segment de discours peut être exprimé par un seul énoncé ou bien par un groupe d'énoncés qui, pris ensemble, mettent en œuvre une sous-intention de la part de l'énonciateur dans le discours à l'œuvre. Par exemple, camper la scène dans le récit d'une histoire, raconter un épisode dans un récit, ou établir et justifier une prémisse au sein d'une argumentation ».

> unité de discours nouvelle où le référent le plus saillant de l'unité précédente exige d'être réintroduit en même temps que de nouvelles coordonnées de temps, d'espace et de perspective. »

La présence d'une redénomination peut ainsi correspondre à l'indication d'un déplacement. Mais ce n'est pas toujours le cas. Selon Catherine Schnedecker (1997, 2003, 2005) qui s'intéresse particulièrement aux cas de redénominations par les noms propres, la redénomination peut désambiguïser une situation référentielle floue.

> « Il est des expressions référentielles littéralement remarquables en ce qu'elles répètent, comme on le dit traditionnellement, l'étiquette lexicale qui a introduit le référent. Du fait qu'elles font apparaître une constante référentielle, ces expressions servent d'ancrage, permettant à l'interprète de 'raccrocher les wagons' quand le parcours référentiel est […] particulièrement fluctuant.» (Schnedecker, 1997:30)

Cette remarque est également faite par Virtanen (1992:114), mais uniquement au niveau de portions de texte narratives.

Pour Virtanen, il y a risque d'ambiguïté lorsque :

- la distance entre les deux expressions co-référentielles est trop grande, nécessitant alors une charge cognitive trop lourde à l'allocutaire pour faire le lien ;
- d'autres participants participent également à ce même moment du discours.

Virtanen note également qu'une redénomination peut correspondre au commencement d'une nouvelle unité textuelle[87], même s'il n'y a pas risque d'ambiguïté. (Virtanen 1992:102), ce qui rejoint notre première association entre redénomination et indice de déplacement.

Une dernière remarque, très importante, est à faire sur l'emploi de la redénomination à travers les différents types de texte. Schnedecker (2005) montre que, dans des textes à site mono-référentiel, la redénomination peut tout à fait constituer une alternance au pronom. Lors de nos analyses concernant les redénominations par nom propre, nous devrons rester très vigilante sur le type de texte observé.

Pour finir cette petite section, il faut noter que les reprises lexicales ne sont pas toujours des redénominations. En effet, dans de nombreux cas (surtout dans les cas de dénominations par descriptions et non pas nom propre), la reprise lexicale est l'occasion d'effectuer une réduction de terme, notamment dans les écrits scientifiques où les dénominations peuvent correspondre à des expressions relativement complexes (cf. Jacques (2003).

Notre distinction entre descriptions complètes et descriptions réduites s'inspire complètement de la thèse de Marie-Paule Jacques (2003) qui observe les facteurs discursifs permettant la réduction des termes complexes. Le phénomène de réduction entre évidemment dans les phénomènes d'anaphore. Dans le cas d'une anaphore pronominale, le lecteur va identifier dans la sous-représentation en cours une entité saillante dont la valeur peut saturer la valeur quasi-vide du pronom. Dans le cas d'une description réduite, le lecteur va « chercher dans le contexte les éléments à l'aide desquels «saturer» la référence du SN, dont le N est, pour une occurrence de terme réduit, une forme identique à la tête d'un terme complexe. » (Jacques 2003:176). Ces descriptions réduites (ou « termes réduits ») sont généralement déterminées par un démonstratif ou un défini, qui instruit le lecteur qu'il a affaire à une expression dont la valeur a besoin d'être saturée. Ainsi, les SNdem et SNdef, lorsqu'ils co-réfèrent, sont associés au phénomène anaphorique de réduction de terme autrement appelé « effacement de l'expansion d'un terme complexe » (Jacques 2003).

Jacques observe en corpus que les démonstratifs déterminent principalement des descriptions réduites :

87 Virtanen définit ses unités textuelles en suivant le modèle de la grammaire fonctionnelle (Dik 1997). Dans ce cadre, différentes unités sont organisées hiérarchiquement, selon des critères interprétatifs. La plus grande unité délimitée est le texte, puis les s(sous(sous))sections, les épisodes et enfin, les déplacements (*moves*).

> « Dans 83 % des cas, le SN démonstratif est à interpréter comme un terme réduit, le co-texte antérieur comportant soit le terme complexe correspondant (69 % des occurrences), soit d'autres indices contextuels permettant sa récupération dans le modèle du discours (14 % des occurrences) » (Jacques 2004:179)

Nous en arrivons à la distinction cruciale à faire entre les différentes façons de déterminer une description.

V.4.3.c) Détermination des groupes nominaux

Les déterminants constituent des indices de l'accessibilité des référents au sein du discours. Selon les différentes théories de l'accessibilité (Prince 1981, Gundel *et al.* 1993, Ariel 1990), nous avons l'échelle :

SN possessifs > SN démonstratifs > SN définis > SN indéfinis

Les SN possessifs sont sans ambiguïté des expressions co-référentielles qui permettent une progression thématique à thèmes dérivés, comme dans l'exemple V.20 où les deux SN possessifs sont dérivés du référent principal : le président syrien Assad.

(V.20) *Contrairement aux affirmations de certains observateurs, le président Assad n'a plus besoin de maintenir le pays dans un état de guerre pour des raisons de légitimité interne.* ***Son régime*** *jouit d'une popularité plus grande aujourd'hui qu'il y a dix ou quinze ans.* ***Sa gestion du processus de paix*** *– en engageant son pays sur la voie de la paix régionale mais sans s'y précipiter tête baissée – semble recueillir l'assentiment des partisans du pouvoir comme de ses opposants. La perspective d'une paix avec Israël a dès le départ été présentée à l'opinion publique comme "la paix des braves" et surtout pas comme une mise au rabais des aspirations nationales du peuple syrien. [...]* *[GEOPO_5]*

Les SN indéfinis sont généralement associés à un phénomène de rupture dans les progressions thématiques, comme le montrent les trois SN indéfinis de l'exemple V.21.

(V.21) *De l'avis de ces militaires, l'enlisement de Tsahal au Liban-sud commence à affecter sérieusement le moral des troupes alors que l'assurance et la combativité du Hezbollah ne font que se renforcer sur le terrain. À l'inverse, l'ALS censée au départ être la cheville ouvrière de tout le dispositif israélien à le sud est devenue au fil de le temps et plus précisément depuis deux ans un allié de moins en moins fiable et de plus en plus difficile à gérer et à contenir.* ***Plusieurs sources, israéliennes et autres,*** *font état de défections de plus en plus nombreuses en son sein de jeunes combattants qui vont grossir les rangs du Hezbollah et/ou se transforment en agents doubles transmettant au Hezbollah des renseignements sur les mouvements et les opérations tactiques des troupes israéliennes.*
Contre les tenants de cette thèse, ***un noyau dur d'officiers*** *continue à défendre fermement le maintien de la zone de sécurité comme un moindre mal.* ***Un retrait sans garantie de sécurité, même avec menaces de représailles massives en cas de attaques de le Hezbollah sur le nord d'Israël****, serait un coup de poker aux risques incontrôlables, qui exposerait directement les populations civiles. Les combattants du Hezbollah s'étendraient tout au long de la frontière et tenteraient des opérations d' infiltration en territoire israélien. Le retrait porterait également un coup fatal au prestige de Tsahal vis-à-vis de l'opinion publique israélienne mais également arabe, contrainte pour la première fois de se replier sous la pression d' une guérilla de quelques milliers d' hommes. [...]* *[GEOPO_4]*

Entre ces deux extrêmes, nous avons les SN définis – SNdef et les SN démonstratifs – SNdem. Les déterminants démonstratifs et, plus relativement, les déterminants définis constituent des indices fiables de relation anaphorique.

Manuélian (2004) offre une étude en corpus de la co-référentialité de SNdef et SNdem. Cette étude – qui précède l'élaboration d'un outil de génération – se base sur l'annotation manuelle de tous les SNdef et SNdem présents dans un extrait de l'année 1987 du Journal Le Monde[88] (65 000 mots).

L'annotation réalisée par Manuélian consiste à déterminer pour chaque SNdef et SNdem s'il correspond à une première mention, à une anaphore associative ou à une deuxième mention. Dans les cas de deuxième mention, Manuélian distingue les cas de co-référence directe par reprise lexicale et les cas de co-référence indirecte*, i.e.* sans

88 Ce corpus est extrait du corpus ATILF (annoté morphosyntaxiquement).

reprise lexicale, comme dans « *Hélène Manuélian (2004) propose... Cette étude...* » où *« Cette étude* » est une deuxième mention par co-référence indirecte de « *Hélène Manuélian (2004)* ».

		SNdef	SNdem
Nombre d'occurrences		8 863	557
Première mention*		78%	17,5%
Anaphore associative		4,5%	0,2%
Deuxième mention = co-référence		18%	82%
	directe	7%	16,5%
	indirecte	11%	65,5%

*La première mention correspond à un emploi référant à un objet non encore mentionné dans le contexte (Manuélian 2004:83)

Tableau V.4 : Emploi (co)référentiel des SNdef en des SNdem (Manuélian 2004)

La différence entre SNdef et SNdem apparaît immédiatement. Les SNdef correspondent essentiellement à des premières mentions alors que les SNdem sont à 82% utilisés pour co-référer, et principalement par anaphore indirecte. Notons également que les SNdef sont 16 fois plus nombreux que les SNdem ! ce qui fait que, d'un autre point de vue, ce sont majoritairement des SNdef qui expriment une deuxième mention (1595 occurrences) et non des SNdem (457 occurrences).

Ces résultats se retrouvent également dans la thèse de Dupont (2003) qui a réalisé une analyse manuelle sur le caractère co-référentiel des SNdem et SNdef, en comparaison à l'emploi des pronoms. Son annotation consiste à mesurer la distance entre l'expression co-référentielle et son antécédent. Les résultats de son travail de thèse sont présentés dans le tableau V.5.

	pronoms			SN démonstratifs			SN définis			TOTAL	
antécédent dans...	Nb	%	%'	Nb	%	%'	Nb	%	%'	Nb	%
la même phrase	110	20,8	93,2	4	4,8	3,4	4	2,7	3,4	118	15,6
la phrase précédente	320	60,5	82,1	50	59,5	12,8	20	13,5	5,1	390	51,6
le même paragraphe	75	14,2	26,7	17	20,2	10,8	65	43,9	41,4	157	20,8
un autre paragraphe	24	4,5	26,7	13	15,5	14,4	53	35,8	58,9	90	11,9
TOTAL			529			84			148		755

Les % correspondent au profil-colonne et les %' au profil-ligne

Tableau V.5 : Fréquence de l'utilisation des pronoms, SN démonstratifs ou SN définis par rapport à la distance de l'antécédent (Dupont 2003:90)

Comme on le voit dans ce tableau, les SNdem sont plus proches, dans leur comportement anaphorique, des pronoms puisqu'ils co-référèrent généralement à une expression située dans la phrase précédente. Ces deux études en corpus nous permettent de dire que les SNdem sont des descriptions co-référentielles plus 'fiables' que les SNdef puisque les SNdef peuvent fréquemment ne pas être des descriptions co-référentielles.

La distinction entre descriptions co-référentielles définies et démonstratives est source d'un très grand nombre de travaux en linguistique française (Corblin (1987, 1995), Charolles (2002), Kleiber (1981), De Mulder (1994, 1997)). Cette distinction repose essentiellement sur la gestion des « circonstances d'évaluation » (Kleiber 1990b), *i.e.* les circonstances actives lors de la situation énonciative dans laquelle le référent est saillant. Ainsi, les premières semblent signaler qu'il y a reprise d'un référent et de ses circonstances d'évaluation, tandis que « le contenu nominal du SN démonstratif (re)classifie le référent qu'il désigne comme un individu particulier de la classe des N » (Corblin, 1995:66), sans maintien des circonstances d'évaluation.

> « Contrairement au défini qui appréhende le référent comme unique du fait de la relation qu'il doit entretenir avec un des participants de la scène, le démonstratif implique une rupture avec cette scène. La saisie du référent se faisant par investigation de la situation d'énonciation, il y a détachement avec tout ce qui peut toucher à la façon dont les participants conçoivent les événements auxquels ils sont mêlés. De là il découle que les SN démonstratifs installent le référent dans le focus d'attention des destinataires comme une entité nouvelle ou, en tout cas, en faisant abstraction de la représentation que les destinataires peuvent ou pourraient en avoir dans les circonstances où intervient l'échange. » (Charolles 2002:120-121)

De ce fait, une reprise par SNdem s'associe souvent à un effet de déplacement ou de rupture (« Les démonstratifs, des indices de changement de contexte » De Mulder 1997:120), ce qui n'est pas le cas des reprises par SNdef.

V.4.4. Des constructions thématisantes ou focalisantes : les phrases « thétiques »

Cornish (2005) définit les phrases thétiques comme des énoncés qui « servent à rapporter un événement ou à présenter un nouveau référent ou un nouvel état de choses au sein d'un discours » et qui sont caractérisée par le fait que, dans ces constructions, le verbe principal ne prédique pas. Cette absence de prédication correspond à une absence de topique au sens de Lambrecht (1994). Dans ces constructions, le verbe sert à « indiquer l'existence, la localisation, l'apparition ou la disparition du référent du terme sujet » (Cornish 2005:78). Il ne sert pas à apporter un commentaire sur le terme sujet. De fait, le terme sujet n'est pas le topique de la phrase.

Cette absence de topique confère à ces constructions un effet de discontinuité dans la séquentialité du discours.Gómez-Gonzáles

> « Au niveau discursivo-pragmatique, leur raison d'être est de servir à présenter un entité, une proposition ou un état de choses en tant qu'élément d'information nouveau pour le discours. En tant que telles, elles ne présupposent qu'un contexte discursif minimal, et peuvent de ce fait apparaître au tout début d'un discours, ou d'une unité de discours ; à ce titre, elles peuvent servir à indiquer la fin de l'unité précédemment en vigueur, et le début d'une nouvelle ; mais aussi l'existence d'une description ou explication parenthétique ou d'arrière-plan. » (Cornish 2005:75)

Plusieurs types de constructions thétiques se distinguent, chacune jouant avec l'ordre des constituants phrastiques, ce qui permet d'avoir en position Thème autre chose que le sujet grammatical et, dans des termes plus syntaxiques, de mettre en fonction de sujet autre chose que le sujet logique. Ainsi, le Thème peut correspond à une prédication complète comme dans les constructions clivées (la Systémique Fonctionnelle parle alors de « *predicated Theme* »), un complément du verbe antéposé comme dans les inversions locatives, ou l'expression de la simple existence d'un objet de discours comme dans les constructions présentationnelles. Nous ne nous étendrons pas dans la description de telles constructions, vu le nombre d'études qui y sont consacrées, notamment dans le domaine de la structure informationnelle des phrases (voir à ce sujet tous les travaux de Lambrecht cités dans la bibliographie). Nous citerons simplement quelques auteurs qui ont nourri notre appréhension de telles constructions.

Les **constructions clivées** ont fait l'objet de plusieurs études en linguistique (Prince 1978, Lambrecht 1994, Delin 1989, Gundel 2002, Gómez-González 2001). Cependant, les clivées ont plutôt fait l'objet de travaux sur la construction des phrases et leur structure informationnelle. Au niveau informationnel, les clivées permettent de mettre le focus d'attention sur un élément particulier. De ce fait, peu d'études cherchent à définir le rôle discursif de ces constructions. Il faut cependant citer le travail d'Hasselgård (2004a) qui pose l'hypothèse d'une similarité de comportement entre les

adverbiaux circonstanciels et les clivées qui mettent le focus sur des localisations spatiales, temporelles ou notionnelles[89], comme dans :

(V.22) ***C'est dans le Midi méditerranéen et aquitain et en Île-de-France*** *que les enfants d'employés sont en plus grand nombre - jusqu'à près de 20% des enfants en Corse, à Nice, en Seine-Saint-Denis ; par contre, les départements ruraux du Centre-Ouest et du Centre-Est en comptent souvent moins de 10%. [ATLAS_2]*

Hasselgård définit les clivées comme des constructions dont la fonction peut être très diverse. Ainsi, les clivées peuvent :

- signaler un contraste,
- introduire un nouveau topique,
- effectuer un déplacement d'un topique à un autre,
- indiquer la fermeture d'une portion de texte, ou
- signaler la mise en Thème d'un élément.

À la lecture de cette liste, les constructions clivées apparaissent davantage comme des indices de discontinuité (déplacement ou rupture) que des indices de continuité, et cela sans doute davantage lorsque c'est une circonstance qui est mise en focus. Le fait d'utiliser une construction clivée permet d'effectuer un déplacement en douceur, selon les termes d'Hasselgård.

> "The function of transition seems particularly prominent with clefted adverbials. Clefts with this function typically have a given, often anaphoric, adverbial in cleft focus position, while the cleft clause introduces a topic for the subsequent discourse. The speaker/writer thus achieves a smooth transition between two topics, juxtaposing them by means of a relational clause, and launching the new topic unobtrusively in a subordinate clause." (Hasselgård 2004a:12)

Les **constructions présentationnelles** (ou « existentielles ») sont un autre type de constructions thématisantes qui permettent de mettre au devant de la scène (du *text-world*), un objet de discours particulier.

> "The underlying rationale is that the interpersonal dimension of existential-there constructions does not reside in their transmitting of the existence or the non-existence of something, but rather in their conveying an instruction to present something on the scene" (Gómez-González 2001:263)

De telles constructions peuvent également être associées à l'indication d'une discontinuité dans le discours. Elles n'ont pas de rôle actif dans le déroulement des progressions thématiques puisque l'élément en position Thème est uniquement structurel.

> « Le présentatif '*there*' [dans des énoncés existentiels de type : *'there are a lot of constraints of interference between received spots*], tout en fournissant un ancrage pour l'expression des nouvelles informations, n'a pas d'identité propre. À notre avis, un tel élément sans référent ne peut fournir une base de départ pour un enchaînement phrastique ultérieur et ne joue donc pas de rôle actif dans la progression thématique. Le fait que, selon la norme écrite, l'expression d'un sujet grammatical reste obligatoire ne change rien à la nature redondante d'un '*there*' ou d'un '*it*' impersonnel dans les énoncés existentiels » (Carter-Thomas 2000:67-68)

Ces constructions servent généralement à introduire dans le discours de nouveaux référents, comme dans l'exemple V.23.

(V.23) ***Il existe deux éventualités, quoique peu probables, qui pourraient bouleverser cette situation.*** *La première est celle d'un retrait unilatéral des forces israéliennes du Liban – en application de la résolution 425 du Conseil de sécurité des Nations unies – avec menace de représailles massives en cas de attaques en territoire israélien. L'autre possibilité serait une escalade de la violence entre les forces israéliennes et le Hezbollah qui pourrait amener Israël à lancer une*

89 Hasselgård (2004a) remarque que, lorsque le focus est mis sur des circonstances, dans 74,5% des cas, il s'agit d'une localisation spatiale ou temporelle (localisation temporelle dans 45% des cas).

attaque de large envergure contre le Liban et contre des cibles syriennes. Dans ces deux cas de figure, la Syrie [...] [GEOPO_5]

Troisième cas de construction perturbant l'ordre canonique, le cas des inversions du sujet. Catherine Fuchs (1997, Fournier & Fuchs 1998, Fuchs & Fournier 2003) et Karen Lahousse (2003a, 2003b) ont consacré de nombreuses études au sujet de ces constructions, ainsi que Francis Cornish dans un article de 2001.

Les inversions ou **constructions à sujet inversé** correspondent à la mise en initiale soit du verbe principal, soit d'un argument inversé exprimant généralement une localisation (que nous abrégeons ARGU). Le premier cas correspond à une inversion focus ; le second cas à une inversion ordinaire ou « inversion locative » (Cornish 2001a). Concernant l'inversion focus, il s'agit de mettre le sujet grammatical en position focus (en fin de phrase et donc, avec un degré de dynamisme communicatif élevé). Ce type d'inversion – plus rare – peut permettre l'introduction d'un nouveau référent, ce qui les rapproche d'une indication de discontinuité, mais à un niveau purement topical. En effet, les *settings* actifs ne sont pas déplacés. Et il est généralement possible de continuer une chaîne de référence en 'sautant' au dessus de la construction inversée, qui ne constituerait qu'une pause dans la progression thématique.

(V.24) *L'école, c'est la solitude parmi les autres et le début de la conscience de solitude, c'est le premier monde où l'on est jeté sans clés ni indices, où les autres sont déjà un mystère que l'on pressent noir sans comprendre. Solitude encerclée déjà, car les autres sont ensemble, solitude humiliée déjà avec ce père douteux, et l'insécurité commence avec l'humiliation pressentie.* ***Vient la première injustice, les coups de règle sur les doigts, qu'on n'a pas mérités, et le visage sinistre et cruel du bourreau, penché contre soi****. Il ne sera plus seul dans la solitude, mais avec les héros-victimes, depuis le Christ jusqu'à Parnell. Il s'oriente vers une pose de solitude. [PEOPL_2]*

Concernant l'inversion locative, Fuchs & Fournier (2003) remarquent que la localisation mise en initiale dans ce type d'inversions ne peut avoir de pouvoir cadratif.

> « Dans la construction XVS (avec le sujet postposé), X peut être un complément plus ou moins fortement intégré au prédicat selon les cas, mais il n'est jamais totalement indépendant de la relation prédicative (*cf.* l'absence de virgule) et n'a pas d'autonomie référentielle : ce n'est pas un repère au plan énonciatif. C'est pourquoi il est intrinsèquement non cadratif : il participe d'une situation prototypiquement statique.
> Cette construction revient, fondamentalement, à poser une relation entre X (en position thématique) et S (en position rhématique) par l'intermédiaire d'un V qui fonctionne comme un simple relateur. » (Fuchs & Fournier 2003:91)

Selon ce point de vue, les constructions inversées locatives ne peuvent pas, elles non plus, indiquer un déplacement. Il semble que la principale fonction discursive de telles constructions soit la mise en arrière-plan du référent sujet, afin d'indiquer au lecteur qu'il « devra accorder un degré moindre de concentration et d'attention au référent du sujet » (Cornish 2001a:112). Du point de vue de la structure informationnelle, la post-verbalisation du sujet indique que celui-ci ne peut être le topique de la phrase. Dans les cas d'inversion locative, il peut y avoir mise en place d'un topique scénique, mais dont la portée est réduite à la phrase

Enfin, dernier type de construction thématisante, les dislocations à gauche. Knud Lambrecht est certainement celui qui a le plus travaillé sur ces constructions. Deux variantes de dislocation existent : les dislocations à gauche et les dislocations à droite. Un exemple type de dislocation est la célèbre phrase : « *Ils sont fous, ces romains* », qui constitue une dislocation à droite que l'on pourrait transformer en la dislocation à gauche suivante : « *Ces romains, ils sont fous* ». Ce type de construction est très associé au registre oral. Cependant, nous trouvons quelques exemples de dislocations dans notre corpus écrit. Lambrecht (2001) définit les dislocations comme des constructions qui servent à marquer le topique, *i.e.* à indiquer que parmi plusieurs possibilités, c'est cette entité qui va être le topique de l'énoncé. Pour cela, l'entité doit avoir été énoncée dans le discours précédent ou être inférable par la situation discursive ou le discours précédent. Dans notre corpus, les dislocations ont la forme de la première phrase de l'exemple V.24. Elles

permettent la mise en place d'une continuité topicale, sans pour autant rompre avec le discours précédent (puisque l'entité doit avoir été énoncée). Ces constructions semblent très contraintes par le type de texte dans lequel elles apparaissent et donnent au texte un style plus littéraire, voire poétique.

V.5. Des indices texto-interpersonnels

Il est assez rare de trouver dans la littérature des descriptions du fonctionnement de la modalité du point de vue de l'organisation du discours. La plupart des ouvrages s'intéressent davantage a l'expression de la modalité (quelles formes linguistiques peuvent exprimer une modalité) qu'à l'impact de l'expression de la modalité sur l'environnement discursif. Parmi les expressions recensées, nous trouvons généralement des verbes spécifiques *(devoir, croire, vouloir)* et des adverbes *(certainement, heureusement)*. L'auteur peut également user de constructions spéciales lui permettant, indirectement dans le cas des textes expositifs, de porter un certain jugement sur les informations qu'il relate.

V.5.1. Les adverbiaux modalisateurs – MODA

Les adverbiaux modalisateurs sont des expressions qui traduisent l'attitude d'un locuteur par rapport à son énoncé, c'est-à-dire le doute, la certitude, la critique... toute marque de jugement ou de sentiment du locuteur. De nombreuses dénominations existent pour référer à ce type d'expressions. Biber *et al.* (1999), dans leur grammaire de l'anglais, parlent de *stance adverbials.* Leur terminologie est celle qui se rapproche le plus de la nôtre (ils distinguent des adverbiaux circonstanciels, connecteurs et modalisateurs) :

> "Stance adverbials convey speaker's comment on what they are saying (the content of the message) or how they are saying it (the style)" (Biber *et al.* 1999:764)

Dans les grammaires françaises, ces adverbiaux sont souvent intégrés aux circonstants, à l'intérieur de la catégorie des « compléments interprétatifs ». Melis (1983 161-168) distingue les compléments assertifs qui « permettent au locuteur de prendre position sur la véridicité de son discours » (*probablement, manifestement, vraisemblablement, sans doute,* etc.) des compléments évaluatifs qui permettent au locuteur d'exprimer une appréciation subjective sur son discours (*curieusement, étrangement, malheureusement*, etc.) Nous ne ferons pas cette distinction, considérant que tout jugement, qu'il soit assertif ou évaluatif relève de l'expression de l'implication du locuteur, ce qui nous intéresse ici. En effet, l'expression de la composante interpersonnelle peut participer au signalement d'un déplacement, comme l'illustre l'exemple V.25 que nous avons déjà traité en IV.5.

(V.25) *Malheureusement, certaines difficultés structurelles et conjoncturelles sont apparues et cette étude officielle n'a pas encore vu le jour. Le groupe de réflexion chargé de l'étude était une sous-commission particulière du Policy Coordinating Committee sur l'espace (PCC-space) créé par le NSC. Dans les faits, les efforts du NSC en matière spatiale n'ont pas été suffisants. L'autorité au sein des sous-groupes n'était pas clairement attribuée au NSC. Ed Bolton, Director for Space au NSC sous l'autorité de Frank Miller, n'était pas de rang suffisant pour imposer des compromis aux différentes agences réunies dans le PCC-Space. Surtout, les événements du 11 septembre ont axé les priorités du gouvernement sur l'action et non sur la réflexion. [GEOPO_2]*

Nous avons de fortes raisons de croire que c'est le positionnement de cet adverbial modalisateur en initiale de paragraphes qui lui confère une portée et non la présence seule du modalisateur. Il est en effet très probable qu'en d'autres positions textuelles, un adverbial comme « *Malheureusement* » ne puisse délimiter un cadre de discours. Nous vérifierons cette supposition par l'analyse de nos données.

V.5.2. Des constructions modalisantes : les constructions impersonnelles

Les constructions impersonnelles permettent soit de porter de façon indirecte un jugement sur un fait (V.22) ; soit de remplir la fonction de sujet grammatical lorsque aucun référent ne semble adéquat (V.20). Ce type de construction est également employé pour présenter un référent au devant de la scène, ce qui correspond alors à notre catégorie des « constructions présentationnelles ».

(V.26) *Il convient donc de faire le point sur les acquis de la PESD en rappelant que les questions débattues aujourd'hui correspondent à des préoccupations anciennes et que le processus se poursuit en dépit de les vicissitudes de la politique impériale des États-Unis et des clivages qui sont apparus au sein de l'UE entre les 'atlantistes' et les tenants d'une 'Europe européenne'. [GEOPO_23]*

Ces constructions sont particulières aux langues à sujet obligatoire telles que le français, qui nécessite un sujet pour tout verbe fini, même si celui-ci n'a aucune valeur référentielle. Dans des langues sans sujet obligatoire tel que le vietnamien, les pronoms impersonnels ne sont pas nécessaires et de ce fait n'existent pas, comme le montrent les traductions en (V.20) et (V.16)[90].

(V.27) mưa[91]
Il pleut

(V.28) *Phàm lăng là xây tự sinh thời vua, chớ không phải khi vua băng hà rồi mới xây.*
Il est dit que le tombeau se construit du vivant du roi et non après sa mort.

Phàm	*lăng*	*là xây tự*	*sinh thời*	*vua,*	*chớ không phải*
En général	*tombeau royal*	copule + *se construire*	*durant sa vie*	*roi,*	*et non pas*

khi	*vua*	*băng hà*	*rồi*	*mới*	*xây.*
lorsque	*roi*	*mourir*[92]	*déjà*	*juste*	*construire*

Les constructions impersonnelles qui nous intéressent particulièrement dans notre étude sont celles qui permettent au locuteur d'exprimer indirectement un jugement sur un fait. Leur fonction est alors très proche de celle associées aux constructions en *on/nous* vues précédemment. Les constructions impersonnelles sont clairement associée à l'expression de la composante interpersonnelle et sont désignées en SF par l'étiquette « *Thematized comment* ». Ces constructions permettent de commencer la phrase par le point de vue du locuteur, à la façon des adverbiaux modalisateurs.

> "[Thematized comment] still involved a grammatical operation (the use of 'it' as a place-holder), which serves to set up as the starting point of the message the speaker's own comment. One's own attitude is natural starting point, and thematized comment is extremely common in many kind of discourse." (Thompson 2004:152)

En mettant en sujet un pronom impersonnel, ces constructions permettent un positionnement énonciatif du locuteur à la façon des constructions en *on/nous...* Les impersonnelles permettent à la fois l'expression d'un jugement et une distanciation entre le locuteur et les faits rapportés. Elles sont de ce fait très courantes dans le discours journalistique , comme le remarque Charaudeau (2006).

90 Nous ne faisons bien évidemment pas une étude sur la typologie des langues. Ces traductions nous permettent de souligner la présence uniquement syntaxique du pronom impersonnel *il* (à l'inverse du pronom personnel *il*). Elles proviennent du document fournit lors du baccalauréat de 1995 pour l'épreuve de vietnamien langue étrangère. Elles nous permettent en plus de faire un clin d'oeil à notre langue paternelle.

91 Le verbe *mưa* est catégorisé verbe impersonnel dans les grammaires vietnamiennes. il peut également avoir un emploi personnel : *trời mưa* (le ciel pleut). IL peut également référer au nom *la pluie*.

92 Le mot *băng hà* signifie mourir, mais uniquement dans les situations où la personne qui meurt est un roi.

> « L'enjeu de crédibilité exige de celui-ci [le sujet énonciant] qu'il ne prenne pas parti. D'où une délocutivité obligée qui devrait faire disparaître le *Je* sous des constructions phrastiques impersonnelles et nominalisées. Ce n'est pas à proprement parler de l'objectivité, mais c'est le jeu de l'objectivité par l'effacement énonciatif. »

Nos données confirment cet état de fait, puisque l'un de nos sous-corpus, plus journalistique (GEOPO) montre significativement plus de constructions impersonnelles.

V.6. Récapitulatif des indices de séquentialité en position initiale

Le tableau V.6 reprend les différents éléments que l'on peut repérer en position initiale et leur associe une certaine capacité à indiquer une continuité ou une discontinuité dans la séquentialité du discours.

	dénomination	***abréviation***	***indice de...***
mise en forme	changement de section	S1	discontinuité (rupture)
	changement de paragraphes	P1	discontinuité globale (déplacement)
	titre	-	discontinuité (rupture)
	puce	-	discontinuité locale (déplacement)
connecteurs		Connect	continuité
INIT *élément détaché en initiale*	adverbial circonstanciel	CIRC	discontinuité
	adverbial modalisateur	MODA	discontinuité ?
	adverbial textuel	TEXT	discontinuité
	apposition	APPO	continuité
	argument inversé	ARGU	continuité
ThTop *Thème topical*	pronominalisation	PRO	continuité
	reprise lexicale	_R	continuité ou déplacement
	détermination indéfinie	SNindef	rupture
ThSpe *Construction Spéciale*	construction clivée	Cliv	discontinuité
	construction présentationnelle	Present	discontinuité
	Construction à sujet inversé	SujInv	continuité
	Dislocation à gauche	Disloc	continuité
	Construction avec On/Nous...	On...	discontinuité ?
	Construction impersonnelle	ILimp	discontinuité ?

Tableau V.6 : Récapitulatif des différents candidats au rôle d'indice de séquentialité

Partie 3.

Mise en Oeuvre

« Stratégies »

Chapitre VI
Étude de l'organisation discursive et linguistiques de corpus

Sommaire

Les linguistiques de corpus se donnent pour objet de décrire certains aspects du langage en se basant sur sa réalité, *i.e.* sur des textes existants. Une étude linguistique en corpus décrit des fonctionnements linguistiques en se basant sur une interprétation de faits réellement observés. Cette approche repose sur l'idée de partir à la découverte, d'aller explorer des réalités langagières pour y découvrir des régularités linguistiques. Il s'agit de vérifier et mesurer la réalité de faits théoriques. Et c'est là un des avantages des linguistiques de corpus : permettre de « poser la question de l'articulation de la performance et de la compétence [...]. Cela remet en cause la traditionnelle et stricte distinction entre acceptabilité et non-acceptabilité. » (Habert *et al.* 1997:9).

> « La linguistique de corpus prend le langage comme elle le trouve » (G. Sampson, 1994 traduit par Habert *et al. op.cit.*)

En plus de remettre en cause la distinction entre acceptabilité et non-acceptabilité, le fait de prendre le langage comme il se réalise implique deux conséquences essentielles : l'ouverture du champ d'investigation à des « expressions banales et peut-être trop courantes pour faire l'objet d'une analyse linguistique qui s'attache souvent à

des *détails* » (Habert *et al. op.cit.*) et la nécessité de tenir compte – et donc de rendre compte – de la variation langagière.

La description de la variation langagière amène à la question et à la définition des registres et des genres langagiers. L'opposition écrit-oral, la corrélation entre des types d'organisation et des genres discursifs, les contrastes sociolinguistiques constituent ainsi des problématiques centrales aux linguistiques de corpus. Nous assistons ces derniers temps à un retour sur le devant de la scène de la problématique des genres discursifs et des registres de langue *(cf.* Condamines 2003*)*. La partie VI.2.3 revient plus précisément sur la question du genre. Un aspect fondamental de notre hypothèse concerne effectivement l'idée que nous avons différents modes de segmentation et articulation entre ces modes selon le 'type' de texte considéré.

Pour étudier certains aspects de l'organisation discursive, nous avons mis en place une méthodologie basée sur une exploration en corpus. Cette exploration consiste en une analyse exhaustive de tout élément qui compose la position initiale, ce qui s'inscrit complètement dans les linguistiques de corpus pour lesquelles un des principes fondamentaux est de ne pas présélectionner des données jugées *a priori* pertinentes (selon notre compétence linguistique).

> "data are used exhaustively : there is no prior selection of data which 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 [Computer Corpus Linguistics]" (Leech 1992:112)

Dans cette citation, Leech fait référence non pas aux linguistiques de corpus, mais aux linguistiques de corpus computationnelles. Il est évident que de nos jours, l'informatique fait partie intégrante des approches en corpus ; et Leech est un des pères fondateurs de cette nouvelle vague numérique des linguistiques de corpus. Dans le domaine de l'EOD, on peut même se demander si les linguistiques de corpus sont possibles sans informatique. Nous reparlerons de cet aspect en partie VI.3.

VI.1. Des méthodes d'investigation et des hypothèses

Deux méthodes d'investigation sont envisageables selon la place que l'on donne aux hypothèses linguistiques. La question à se poser est de savoir si c'est le modèle qui guide les méthodes d'investigation ou si ce sont les résultats d'observations en corpus qui construisent le modèle. On retrouve cette distinction dans l'opposition entre une approche *hypothesis-driven vs. data-driven*[93] (Leech 1992, Biber 1999, Tognini-Bonelli 2001, Rayson 2002). D'un côté, il s'agit d'asseoir un modèle théorique sur des données recueillies en corpus. De l'autre, ce sont les données qui permettent de définir des phénomènes linguistiques. Nous nous situons clairement dans ce deuxième type d'approche, en gardant toutefois à l'esprit qu'aucune recherche ne commence sans intuition *a priori*.

VI.1.1. *Hypothesis-driven* or *data-driven approaches ?*

Pour Tognini-Bonelli (2001), dans une approche *hypothesis-driven*, les linguistes ont recours aux corpus pour asseoir leur modèle théorique sur des données réelles, pour apporter la preuve des fondements d'un modèle construit *a priori* .

93 Leech (1992) repris par Tognini-Bonelli (2001) parlent plutôt de *corpus-based vs. corpus-driven approaches.* Nous retenons la terminologie de Biber (1999) et Rayson (2002) qui nous semble plus parlante.

> "The term *corpus-based* [hypothesis-driven] is used to refer to a methodology that avails itself of the corpus mainly to expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study" (Tognini-Bonelli 2001:65)

Elle récuse fermement cette approche en stipulant qu'elle ne sert qu'à attester par une liste d'exemples des définitions théoriques, sans prendre en compte les variations réelles de la langue, *i.e.* la langue dans sa réalité. Le corpus ne sert donc pas à construire le modèle, mais à fournir des preuves de sa validité. Et si la réalité n'est pas aussi adéquate à la théorie qu'on aurait pu l'espérer, ce n'est généralement pas le modèle qui est remis en cause (*cf.* Tognini-Bonelli 2001:68-77). La position de Tognini-Bonelli est quelque peu brutale. En effet, le fait de partir d'hypothèses fortes ou de s'inscrire dans un modèle théorique donné paraît essentiel à toute étude. La distinction est ici à faire concernant l'apport du corpus vis-à-vis de ces hypothèses ou de ce modèle. Le travail de Bestgen *et al.* (2003, 2006) a ceci de tout à fait remarquable qu'il propose une réelle évaluation linguistique basée sur corpus, sans pour autant tomber dans la caricature du corpus 'décoratif'. Cette évaluation n'a pas pour unique but de conforter des hypothèses, mais surtout de porter un regard différent sur un objet linguistique.

À l'opposé de l'approche hypothesis-driven, Tognini-Bonelli défend une élaboration des modèles à partir des données selon une démarche totalement guidée par les données.

> "In a corpus-driven [=data-driven] approach, the commitment of the linguist is to the integrity of the data as a whole, and descriptions aim to be comprehensive with respect to corpus evidence. The corpus, therefore, is seen as more than a repository of examples to back pre-existing theories or a probabilistic extension to an already well defined system" (Tognini-Bonelli 2001:84)

Dans le même ordre d'idée, Rayson (2002) défend l'idée que l'approche « *data-driven* » permet de découvrir des faits que l'on ne soupçonnait pas *a priori*.

> "The problem with [hypothesis-driven] approach is that during the investigation, we can search only for evidence, or lack of evidence, for what we expect to find. The alternative to hypothesis-driven research is *data-driven* research, in which we are informed by the corpus data itself and allow it to lead us in all sorts of directions, some of which we have never thought of." (Rayson 2002:1)

Le choix d'une étude guidée par les données est d'autant plus tentant qu'il se développe à l'heure actuelle des techniques de traitement automatique d'investigation en corpus. Ces techniques réalisent des calculs statistiques qui établissent des mesures de similarité entre zones de textes contiguës. Nous donnons ici l'exemple de deux techniques en vogue : le *TextTiling* (Hearst 1997) et l'Analyse Sémantique Latente (Landauer *et al.* 1998).

La méthode de segmentation thématique intitulée *TextTiling* (*cf.* Hearst 1997, Ferret & Grau 1998, Ferret & Grau 2000, Ferret *et al.* 1998, Ferret 2002) est un exemple de ces méthodes automatiques pouvant s'appliquer au niveau discursif. Il constitue d'ailleurs un des rares algorithmes utilisé en TAL pour segmenter le discours. Cet algorithme réalise un découpage d'un texte en groupes de paragraphes successifs portant sur un même thème en se basant sur une mesure de similarité lexicale entre séquences consécutives de mots. Tous les 20 mots, l'algorithme mesure la similarité entre les 100 mots apparaissant à gauche et à droite. Un minimum local de cette mesure est considéré comme l'indice d'une zone de changement thématique et une frontière est alors définie à l'emplacement de la limite de paragraphe la plus proche. Les mots ayant joué un rôle important dans le maintien à une valeur élevée de la mesure de cohérence lexicale entre deux minima sont mis à profit pour caractériser le thème de la région considérée. Si l'algorithme décrit permet en effet une segmentation thématique pertinente à l'échelle de groupes de paragraphes, rien ne garantit que deux zones non contiguës traitant d'un même thème soient caractérisées par des listes de mots identiques, ce qui rend difficile la détection de proximité thématique entre segments non consécutifs.

Cette méthode permet de réaliser sur de gros corpus des calculs de proximité et de mesurer, segmenter les textes. Cependant, elle ne constitue pas en l'état un outil pour l'EOD puisqu'elle ne permet pas à l'utilisateur d'explorer le texte, de tester des hypothèses, de prendre en compte des facteurs autre que la similarité d'apparition. Ce type de méthode peut cependant être d'une grande utilité pour la validation d'une analyse linguistique préalable. Il s'agit alors de vérifier sur des quantités de données plus grandes et avec des méthodes automatiques des résultats déjà acquis par des méthodes empiriques traditionnelles.

L'Analyse Sémantique Latente – LSA (*cf.* Landauer *et al.* 1998 et http://lsa.colorado.edu/) permet justement ce type de validation. La LSA se base également sur un algorithme statistique qui calcule les similarités 'de sens' entre des mots ou des portions de textes délimitées. Ce calcul se base sur l'hypothèse que deux unités qui affichent des contextes similaires sont sémantiquement proches. L'algorithme compare les contextes droit et gauche de chaque mot ou portion de texte et représente la différence de ces contextes par deux vecteurs. L'angle formé par les deux vecteurs permet de comparer les mots ou portions étudiés[94]. Dans leurs travaux sur le marquage de l'implication du locuteur dans les relations de causalité, Bestgen *et al.* (2003, 2006) ont utilisé cette méthode pour tester sur des quantités de données plus importantes les résultats de leurs analyses manuelles. Après avoir annoté à plusieurs juges le degré d'implication du locuteur dans l'utilisation d'une centaine d'occurrences de connecteurs de causalité, ils ont eu recours à la LSA pour « implémenter les procédures d'analyse ». Cette implémentation a alors permis de « rendre [ces procédures] indépendantes de l'analyste et de les appliquer à des vastes ensemble de données, avec des centaines et même des milliers d'occurrences d'un même phénomène linguistique ». Cette implémentation repose bien entendu sur une hypothèse linguistique forte : le degré d'implication est associé au degré de voisinage sémantique.

VI.1.2. Des hypothèses aux données, des données aux hypothèses

Comme le souligne Degand & Bestgen (2004), l'implémentation des procédures d'analyse nécessite, avant toute technique de TAL, « une série d'hypothèses linguistiques (provenant d'études empiriques antérieures) ». Il est effectivement inévitable et naturel de partir d'hypothèses qui permettent d'observer, d'approcher un objet d'étude. Toute recherche est guidée par des hypothèses. Cela ne signifie pas automatiquement ne rechercher que les preuves de ces hypothèses. L'utilisation de corpus nous permet d'explorer la langue, d'observer ce qui s'y passe, en vue de découvrir certaines régularités alors peut-être modélisables. Elle nous permet également d'observer les variations langagières de ces régularités. Notre angle de vue ne peut être innocent. Les données observées ne sont pas prises au hasard et les méthodes pour les repérer et les analyser ne sont pas sans lien avec des hypothèses qui sous-tendent notre investigation linguistique. Croire que les données sont les seules sources valables pour asseoir une description est dangereux. S'ils ne sont pas associés à une connaissance théorique du phénomène observé et une prise en compte des résultats d'études non empiriques, les résultats d'une approche *data-driven* ne fournissent rien d'autre qu'une observation de la langue. Leur interprétation ne peut pas se faire en attribuant un pouvoir absolu aux données et en faisant fi de toutes les théories et hypothèses construites en amont. Une interprétation peut aboutir à des généralisations peut-être logiques mais pas pour autant linguistiques, *i.e.* explicatives de la langue. Nadau (1999:322) parle alors d'« induction sans rigueur, que les logiciens décrivent comme une projection de prédicats qui ont réussi ».

94 L'interprétation des résultats est la suivante : plus le cosinus de l'angle est proche de -1 ou +1 (ce qui correspond à un angle nul ou plat, soit 0° ou 180°), plus les mots ou portions pris en compte présentent des contextes similaires. Ils sont alors considérés comme étant des « voisins » sémantiques. A l'inverse, plus le cosinus de cet angle est proche de 0 (ce qui correspond à l'angle droit, soit 90° ou 270°), plus les mots ou portions pris en compte présentent des contextes différents et donc, d'après le modèle, un sémantisme différent.

Notre démarche se situe spécifiquement dans un va-et-vient entre données et hypothèses. D'un côté, nous posons deux hypothèses : (1) l'organisation du discours est marquée à la surface du texte et (2) le type d'organisation et la nature de son marquage varient selon les genres et types discursifs. Cette dernière hypothèse constitue le socle de notre méthodologie : observons les variations de nos données en jouant des facteurs influant *a priori* sur l'organisation du discours (voir VII.3). Les variations observées peuvent ensuite être mises en relation avec le modèle de séquentialité correspondant à notre vision de l'organisation discursive (chapitre III).

Cette méthodologie est détaillée dans le chapitre suivant. Elle apporte une réponse aux difficultés rencontrées pour concilier linguistique (de corpus) computationnelle et « recherche ouverte » (Péry-Woodley 2005:189, voir section VI.1.4). Nous nous inscrivons totalement dans la perspective des travaux de Teufel (Teufel 1998, 1999 Teufel & Moens 2002[95]) qui propose une méthode de « zonage » du texte. Cette méthode consiste à repérer automatiquement dans des articles scientifiques les articulations rhétoriques générales (par exemple les zones où l'auteur expose ses objectifs, celles où il critique une autre approche, etc.[96]). Les indices permettant de délimiter les différentes zones sont purement formels : position textuelle, longueur des phrases, reprise des éléments du titre dans la zone, présence de mots récurrents, morphologie des verbes, présence de citations, présence de marqueurs méta-discursifs. La définition des indices s'est effectuée dans une approche mêlant à la fois hypothèses fortes, jugement humain et analyses de corpus *data-driven*[97]. Après avoir posé le modèle linguistique définissant de façon opérationnelle les différentes zones argumentatives (voir III.1.2.a), 80 textes ont été 'découpés' par des annotateurs humains. Ensuite, une évaluation de l'accord entre annotateurs est effectuée pour mesurer l'homogénéité des annotations[98] (si l'accord est trop faible, le modèle définitoire est à remettre en cause). Puis, un programme informatique mesure, sur les 80 textes annotés, la probabilité de certains traits (choisis par hypothèses linguistiques ou non[99]) à être corrélés avec un certain type de zone. Il 'suffit' ensuite de réinjecter ces mesures de probabilité dans les textes, en calculant pour chaque phrase sa probabilité d'être dans telle ou telle zone en associant les probabilités de chaque indice présent dans cette phrase. Il s'agit ensuite d'évaluer l'efficacité du programme en comparant les zones délimitées automatiquement et délimitées manuellement, et de l'améliorer si besoin est (en ajoutant des traits linguistiques à prendre en compte par exemple).

Malgré notre incapacité à employer de tels moyens (surtout humains), nous nous inscrivons modestement dans la même veine : une méthodologie basée sur des traits purement formels rend l'analyse discursive plus respectable du point de vue de la linguistique computationnelle qui laisse pour l'instant l'EOD de côté, jugée trop subjective[100].

VI.1.3. Analyses qualitatives et/ou quantitatives ?

En dehors de la différence entre approches *hypothesis-driven* vs. *data-driven*, deux modes d'analyse sont applicables aux données linguistiques : des analyses quantitatives et des analyses qualitatives.

> "The difference between qualitative and quantitative corpus analysis, as the terms themselves imply, is that in qualitative research no attempt is made to assign frequencies to the linguist features which are identified in the data. Whereas in quantitative research we classify features, count them and even construct more complex statistical models in attempt to explain what is observed in qualitative research

95 La page web de Simone Teufel est généreusement fournie : http://www.cl.cam.ac.uk/~sht25/.

96 Nous revenons sur les sept types de zones définies par Teufel & Moens (2002) en III.1.2.a.

97 Teufel 1999 décrit entièrement cette méthode et en présente les résultats.

98 Teufel & Moens (2002) expliquent brillamment comment mettre en place et évaluer une campagne d'annotation multijuge.

99 Les algorithmes d'apprentissage peuvent s'appliquer à n'importe quel trait linguistique, il faut juste que celui-ci soit identifié. Ainsi, il peut s'agir d'unités lexicales précises, de temps verbaux précis, ou de traits plus complexes comme la présence d'une reprise d'un élément du titre.

100 Ces propos sont largement inspirés de ceux tenus par Simone Teufel au cours d'un séminaire CLLE-ERSS (mars 2007, Toulouse).

> 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." (McEnery & Wilson 2001:76)

Concernant l'EOD, il est encore difficile de parler de choix entre analyses qualitatives et analyses quantitatives étant donné la difficulté de considérer de façon opérationnelle un phénomène discursif. En effet, la multiplicité et la diversité des indices à prendre en compte pour décrire de façon stable un phénomène discursif laisse encore les linguistes et informaticiens très prudents dans une quelconque définition de structures discursives (voir la première partie de ce chapitre ainsi que le chapitre V sur les indices de séquentialité). Cependant, les choses vont vite et entre le début et la fin de nos travaux de thèse la mise en oeuvre de 'chantiers' d'annotation discursive associant connaissances linguistiques et techniques informatiques devient de plus en plus envisageable.

Piérard & Bestgen (2005) soulignent la nécessité de comprendre la structure du discours pour pouvoir repérer les éléments linguistiques qui signalent cette structure du discours. Connaître l'organisation discursive d'un texte suppose « une analyse linguistique fine ou le recours à des juges auxquels on demande d'indiquer les ruptures thématiques qu'ils perçoivent. La complexité et le coût de ces procédures manuelles rendent l'étude de grands corpus impraticable. La conséquence de cette situation est que l'on dispose actuellement de peu de données issues de grands corpus qui attestent qu'un élément donné est plus fréquemment employé en situation de rupture ou de continuité thématique. »

De plus, il y a une tradition en EOD et en France pour l'analyse qualitative où l'on cherche à cerner le fonctionnement et l'étendue d'un phénomène en s'appuyant sur des exemples et contre-exemples issus de textes réels ou créés. Ce type d'analyse n'utilise pas les corpus pour mesurer la réalité langagière d'un phénomène linguistique mais pour vérifier que les phénomènes théorisés sont possibles dans le système de la langue (même si ce phénomène est, dans la réalité du système, un phénomène très rare ou très contraint) – nous retrouvons ici la critique évoquée par Tognini-Bonelli au sujet de l'approche *hypothesis-driven*. Le problème avec ce type d'analyse est que l'on reste dans une description d'exemples sans généralisation des résultats obtenus. Ainsi, certains travaux peuvent décrire de façon très fine un phénomène linguistique qui, finalement, ne joue qu'un très faible rôle dans l'organisation discursive dans la réalité langagière, les exemples appuyant l'analyse étant au final relativement rares.

> "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 & Wilson 2001:76)

Notre intérêt pour l'utilisation d'analyses quantitatives n'est pas une question de validation mais de découverte. L'observation des variations quantitatives nous permet de dégager des régularités pertinentes pour décrire l'organisation des textes sans entrer dans leur interprétation (voir à ce sujet l'article de Viprey 2005). Il s'agit de proposer une solution pour étudier sur de gros corpus des phénomènes discursifs. Les méthodes mises en place pour repérer les différents éléments constitutifs de la position initiale et observer leur comportement au fil des textes font l'objet du chapitre chapitre VII. Elles s'appuient sur les concepts et calculs statistiques présentés ci-dessous.

VI.1.4. L'EOD en corpus : une recherche ouverte

Aujourd'hui, l'EOD est souvent jugée trop subjective pour être implémentable. La place accordée aux intuitions dans la construction du modèle y est effectivement très importante. Trop peu d'études existent pour nous offrir des

prises fiables et valides auxquelles se raccrocher[101]. En morpho-syntaxe par exemple, de nombreuses années ont été nécessaires pour arriver à s'accorder autour de la définition des unités d'analyse - définition qui reste tout de même variable selon le cadre théorique. En EOD, nous en sommes encore loin.

L'organisation du discours est composée d'une grande variété de structures. Ces structures peuvent être de grain différent et jouer sur différents niveaux. Nous avons l'exemple type des titres de sections qui organisent le texte dans son entier (à un niveau global) tout en étant impliqués dans la construction de la représentation mentale au niveau le plus local (de phrase en phrase).

De plus, ces différentes structures sont entrelacées. Ainsi, un titre peut participer à une chaîne de référence, poser ou fermer un cadre. De même, un cadre de discours peut chevaucher une chaîne de référence, c'est-à-dire s'ouvrir au cours d'une chaîne et se refermer après la fin de cette chaîne. Le cas inverse est bien entendu possible et peut-être même plus fréquent.

Enfin, comme nous l'avons défini au chapitre V, la signalisation de ces structures et de leur entrelacement s'effectue par des configurations d'indices plus que par des marqueurs au sens noble du terme, les indices pertinents pouvant être de nature linguistique et extralinguistique.

Un des buts de l'EOD est précisément de mettre au jour ces configurations d'indices qui poussent le lecteur à inférer telle ou telle segmentation, telle ou telle structuration. De fait, il n'y a pas de catégories prédéfinies à tester. Dans cette optique, Péry-Woodley (2005) parle de recherche « ouverte ».

> « Un aspect fondamental des travaux sur le discours est précisément la recherche ouverte des corrélats linguistiques des principes d'organisation. [...] 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. typodispositionnelles –, absence de marque). » (Péry-Woodley 2005:189)

Cet aspect de l'EOD va *a priori* à l'encontre des méthodes en linguistiques de corpus qui se basent généralement sur des analyses quantitatives (voir VI.1.3). En effet, pour étudier la réalisation d'un phénomène linguistique, il faut avoir défini préalablement les traits formels et parfois les tests qui permettent de l'identifier de façon stable. L'article de Degand & Bestgen (2004) a spécifiquement pour sujet cette difficulté de s'accorder entre êtres humains autour d'une définition opératoire. L'attribution d'un degré d'implication du locuteur dans des relations de causalité est un bel exemple de cette complexité d'identification d'un phénomène discursif. La plupart des phénomènes discursifs répondent de cette complexité d'identification, leur signalisation étant fréquemment implicite (voir V.1). Notre approche de l'EOD en corpus vient justement en réaction à cette complexité d'identification, en essayant de faire émerger à partir des données mêmes des régularités relatives aux différents modes de structuration définis au chapitre III. Pour faire émerger ces régularités, trois étapes essentielles sont à distinguer :

1. la constitution du corpus ;
2. l'annotation des données, ce qui comprend leur repérage et leur caractérisation ;
3. l'analyse et l'interprétation des données.

101 Notons toutefois que les études dans ce domaine sont en pleine expansion (études sur des corpus anglais généralement).

VI.2. Corpus pour l'EOD

Les corpus correspondent à des « collections de textes (éventuellement un seul texte) constituéess à partir de critères linguistiques et/ou extra-linguistiques pour évaluer une hypothèse linguistique ou répondre à un besoin applicatif » (Condamines 2003:32). Le choix de tel ou tel corpus prend en compte plusieurs critères :

- sa taille ;
- l'échantillonnage de ce qui le compose et son format ;
- les critères extralinguistiques et linguistiques qui le caractérisent.

VI.2.1. Taille du corpus

Les critères de choix concernant la taille des corpus sont autant d'ordre théorique que pratique : on a d'un côté un objectif (évaluation d'une hypothèse ou besoin applicatif) et de l'autre côté des outils et méthodes permettant de l'atteindre. Une analyse sur de gros corpus ne peut se passer d'outils de TAL et malheureusement, les outils actuels ne permettent pas réellement de pister, visualiser et traiter des phénomènes d'ordre discursif. Ce manque d'outillage ajouté à la nature multi-indicielle et variable du marquage discursif explique le recours habituel de l'EOD à des analyses manuelles sur des corpus de faible volume. Car seule l'interprétation humaine est à même d'identifier de tels phénomènes, comme les articulations rhétoriques entre segments, la délimitation des chaînes de référence ou des cadres de discours, et ce n'est d'ailleurs pas une tâche humaine facile (imaginez un court moment devoir annoter tous les progressions thématiques et les relations rhétoriques identifiables dans ce court extrait du sous-corpus GEOPO... sans commentaire).

(VI.1) ***Contrôle des importations*** [titre niveau 4]

Le contrôle des importations pétrolières représente l'autre face de l'interventionnisme pétrolier américain. Dès les années 1930, et plus encore après 1945, le pétrole du Venezuela et du Mexique, puis du Moyen-Orient, exerçait une forte pression sur le marché intérieur américain. La mise en place de barrières protectionnistes s'imposait comme une nécessité sous peine de ruiner le système de proration : les deux faces de l'interventionnisme pétrolier sont donc étroitement liées. Concrètement, la protection prit la forme de quotas et de taxes. Les quotas furent d'abord "volontaires" (1949-1958), puis obligatoires dans le cadre du Mandatory Oil Import Program (1959-1973). En 1932, le Revenue Act imposa, pour la première fois, des taxes sur les importations pétrolières (pétrole brut et certains produits raffinés) ; elles furent progressivement réduites à la faveur d'accords avec le Venezuela et le Mexique, et de la signature du GATT en 1947. Cette forte réduction des taxes fut à l'origine de la réglementation par les quantités (quotas) à partir de 1949. Les taxes ne furent pas pour autant abolies, et furent réorganisées en 1962 dans le cadre du Trade Expansion Act. La pénétration du pétrole importé fut néanmoins très importante sur cette période (cf. Figure 10, p. 31) ; elle eut été nettement supérieure en situation de libre-échange.

La petite taille des corpus a pour conséquence fâcheuse la difficulté à généraliser les résultats observés au delà des textes 'lus'. L'analyse reste au niveau du ou des textes étudiés (nous avons déjà évoqué cela dans la partie VI.1.3), textes devenant, au fur et à mesure de l'analyse, familiers à l'analyste[102]. Car l'utilisation de corpus de petite taille pose la question de la représentativité[103]. Et la réponse consiste parfois à fabriquer ou prendre ailleurs quelques exemples pour asseoir ses hypothèses. Gómez-González (2001) exprime clairement ce dilemme entre petits et grands corpus pour l'analyse de phénomènes discursifs :

> "[One of the reasons which justifies the choice of LIBMSEC [49 285 mots] as the corpus for this investigation] is its relatively small size as compared to other tagged machine-readable corpora, which makes it suitable for a manual (clause by clause) analysis of syntactic Themes across different categories, given that our characterization of syntactic Themes precludes any kind of automatic data

102 Nous avons noté la difficulté de généraliser des analyses qualitatives en VI.1.3.

103 Nous renvoyons à l'ouvrage de Condamines 2005 et notamment à son introduction pour une explication panoramique de la conception de représentativité et des critères de constitution des corpus.

> searching. Indeed, the available text retrieval programs cannot automatically find instances of syntactic Themes because these are not characterized by one specific element itself." (Gómez-González 2001:193)
>
> "The first problem concerning the LIBMSEC is that owing to its small size it did not contain examples of all (sub)types and (sub)classes of syntactic Themes postulated above. This made it occasionally necessary to resort to made-up examples or token taken from the relevant literature." (Gómez-González 2001:196)

Notre exploration repose sur un corpus relativement volumineux de 700 000 mots, présenté en partie VI.3.2. Notre choix pour un 'gros' corpus ne peut être justifié sans celui d'une automatisation des méthodes d'investigation. Ce n'est que parce que nous nous permettons une approche *data-driven* et une automatisation de l'étape d'annotation que nous pouvons travailler sur de telles quantités de données. Nous pouvons dès lors réaliser des analyses quantitatives et reproduire ces analyses sur d'autres données.

VI.2.2. Échantillonnage et Format

Si l'objet d'étude est discursif, on peut difficilement envisager de travailler sur des extraits de textes, nous avons à maintes reprises précisé l'unité fonctionnelle et organisée que compose le texte. Mais la question de l'échantillonnage du corpus implique également la question de la taille de textes constitutifs du corpus. Comme nous l'avons expliqué dans la partie I.4, les textes courts affichent une organisation sans doute moins complexe que des textes 'longs'. De plus, notre méthodologie utilise le facteur 'position textuelle' pour observer des variations au niveau de l'organisation discursive. L'idée étant que le marquage de la séquentialité du discours en position initiale diffère selon le niveau de segmentation considéré (de section en section, de paragraphe en paragraphe et de phrase en phrase), prendre des textes courts aurait été hors de propos (la partie suivante présentant notre méthodologie).

Si l'on souhaite appliquer des méthodes d'extraction ou de repérage automatique ou si les analyses sont envisagées de façon quantitative et ce sur un assez grand volume de données, le format électronique des corpus est nécessaire. Mais la question du format est plus complexe et réside principalement dans la question de l'annotation du corpus qui dépend autant des outils TAL à disposition que des informations que l'on veut extraire. Les outils d'annotation actuellement intéressants pour l'EOD appartiennent aux domaines de la morphosyntaxe[104] et de la syntaxe[105]. Ces outils nous permettent de rechercher dans les corpus des expressions à partir des lemmes, des morphèmes qui les composent, ou de leur catégorie et fonction syntaxiques. On peut aussi utiliser des ressources lexicologiques exogènes afin d'annoter toutes les expressions appartenant à un champ sémantique spécifique, par exemple, tous les noms de lieux géographiques. Les ressources exogènes sont externes à ce qui est contenu dans le corpus d'étude et se distinguent des ressources endogènes qui sont construites à partir du corpus[106]. Par exemple, la délimitation de segments thématiques peut se faire soit en utilisant uniquement des phénomènes de récurrence lexicale présents dans le corpus (c'est la base du *text-tiling*), soit en ayant recours à des dictionnaires ou des bases

104 On parle communément d'étiqueteurs (*tagger)*. L'étiqueteur le plus utilisé de nos jours semble être TreeTagger (http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html) développé par l'*Institute for Computational Linguistics* de l'Université de Stuttgart dans le cadre d'un projet intitulé 'textual corpora and tools for their exploration'. En français, il existe un autre étiqueteur de qualité : Cordial développé à Toulouse par l'entreprise 'Synapse développement' http://www.synapse-fr.com/

105 Deux types d'outils TAL existent au niveau syntaxique : les parseurs et les analyseurs de dépendances. Les premiers analysent la grammaticalité d'un énoncé et sont fréquemment utilisés en génération de texte. Les seconds permettent de délimiter les constituants immédiats d'un énoncé et de dégager les relations syntaxiques entre ces constituants. Parmi ces analyseurs de dépendances, nous utilisons Syntex (Bourigault & Fabre 2000, Bourigault 2007) développé au sein de l'ERSS – Toulouse. (http://www.irit.fr/RFIEC/syntex/).

106 Cécile Frérot (2005) propose une étude comparative de l'apport de ressource exogènes *vs.* ressources endogènes dans une tâche de TAL (la résolution des ambiguïtés de rattachement prépositionnel).

lexicales permettant de repérer tous les termes sémantiquement liés à un thème précis (par relation de synonymie, d'hyper/hyponymie, etc.)[107]

Travailler sur des corpus annotés implique nécessairement la prise en compte (i) du taux d'erreur de l'outil, (ii) de l'orientation du programme :

(i) Pour chaque outil, une évaluation de son taux de précision et taux de rappel est effectuée (le taux de précision correspond au pourcentage d'annotations correctement effectuées par rapport au nombre total d'annotations effectuées ; le taux de rappel est égal à la proportion d'annotations correctement effectuées par rapport à toutes les annotations qu'il aurait dû y avoir). Par exemple, pour Syntex, le taux de précision obtenu en 2000 était de 86%, le taux de rappel de 60%.

(ii) Certains cas ambigus peuvent être résolus différemment pour l'annotation syntaxique et pour l'annotation discursive. Un travail concernant le problème du rattachement syntaxique des circonstants a précisément montré de telles divergences (Hồ-Đắc & Frérot 2004). Les circonstants, lorsqu'ils sont détachés de la structure propositionnelle intéressent particulièrement l'analyste en discours et le syntacticien ; le premier pour les étudier, le second pour ne pas les rattacher. Dans des cas ambigus de rattachement comme « on a observé des crues dans l'Aveyron », le 'syntexticien' (*i.e.* le développeur d'un outil d'analyse des dépendances syntaxiques) aura tendance à rattacher *dans l'Aveyron* au verbe *observer* et ainsi à en faire un constituant intra-prédicatif, tandis que l'analyste en discours préférera le définir comme un extra-prédicatif si le constituant se trouve dans une TSC spatiale construite autour par succession de circonstants de lieux.

Au niveau discursif, il existe quelques corpus annotés discursivement. Deux grands groupes de travaux se distinguent : l'annotation des relations anaphoriques et l'annotation des relations rhétoriques. Depuis le début du siècle, différents projets ayant pour objectif la création de corpus de référence pour l'EOD se mettent en place. Quel que soit le projet, les annotations effectuées sont toutes issues d'un traitement humain. La constitution de tels corpus nécessite des ressources humaines telles que, de nos jours, seuls des projets sur la langue anglaise peuvent financièrement y accéder. Sans entrer dans le détail de ces projets – les articles de référence sont très clairs et généralement en ligne – nous en citons quelques uns en particulier pour leur importance et leur avenir certain. Le *RST annotated corpus*[108] (Carlson *et al.* 2003) est un corpus constitué de 385 textes (176 000 mots) issus du journal *Wall Street*. Chaque texte a été annoté selon le modèle de la RST, *i.e.* toutes les propositions relationnelles – au niveau phrastique ainsi qu'au niveau du texte entier – ont été caractérisées par différents annotateurs humains formés pour l'occasion. Le *Penn Discourse Treebank*[109] (Miltsakaki 2004) propose une évolution du *Discourse Treebank Corpus* vers une annotation des relations de cohérence entre deux phrases adjacentes (qu'il y ait ou non connecteur). Et enfin, le *Timebank corpus*[110] propose une annotation des relations temporelles liant les différents événements exprimés dans environ 300 articles informatifs.

Au niveau de la segmentation textuelle à proprement parler, des projets commencent à se mettre en place sur l'annotation des informations propres à l'architecture des textes. Ainsi, on peut annoter les unités textuelles telles que les titres, les sauts de sections, de paragraphes, les éléments en italique, etc. Les corpus de ce type restent encore assez rares, ce qui est entre autres dû au fait de la non prise en compte de ces caractéristiques par les outils TAL disponibles jusqu'à aujourd'hui. En effet, les outils de TAL prennent généralement en entrée des textes bruts, ce qui efface toute trace d'annotations préalables concernant par exemple la mise en forme. Ce manque est en train de changer depuis l'arrivée du format XML (extensible Markup Language, 1998) qui offre des méthodes simples et

107 Gaume (2004) propose une représentation en graphe des liens de synonymie entre près de 50 000 entrées (nom, verbe, adjectif) du lexique français. Cette ressource est libre d'accès à l'adresse : http://erss.irit.fr/prox/.

108 URL : http://www.isi.edu/~marcu/discourse

109 URL : http://www.seas.upenn.edu/~pdtb

110 URL : http://www.cs.brandeis.edu/~jamesp/arda/time/timebank.html

universelles pour l'annotation de données et donc un nouveau format à utiliser par les outils TAL. Par exemple, LinguaStream[111] (développé par Bilhaut & Widlöcher 2006) peut recevoir en entrée un fichier XML, ce qui permet de conserver des annotations préalables (manuelles ou générées par un autre outil). La génération de corpus annotés discursivement au format XML est d'ailleurs une des préoccupations applicatives de notre travail, ce format pouvant offrir des outils de visualisation intéressants pour l'analyse de l'organisation discursive[112].

VI.2.3. Critères extralinguistiques et linguistiques : le genre et le type

Les derniers critères de choix entrant dans la constitution d'un corpus d'étude sont sans doute les plus importants. Le titre de cette sous-section met en parallèle genre textuel et critères extralinguistiques d'une part, et type textuel et critères linguistiques d'autre part, ce qui présage une double classification des textes : une classification par genre et une classification par type.

Cette partie est très largement inspirée de deux écrits récents axés sur la problématique de la typologie textuelle : l'article d'Anna Trosborg (1997) intitulé « Text Typology : Register, Genre and Text Type » et le rapport du projet TEXTCOOP (LIPN Paris 13) portant précisément sur la classification des textes : « Typologie textuelle, état de l'art et application » (Gayral *et al.* 2007)[113]. Cette problématique n'est pas nouvelle et les propos tenus ici l'ont déjà été, notamment par Douglas Biber (1988-1999). D'ailleurs, ces deux écrits se situent complètement dans la lignée des travaux réalisés par Biber. Nous avons préféré relater les propos tenus par ces écrits qui constituent, de notre point de vue, une bonne synthèse de la problématique à l'heure actuelle.

Trosborg (1997) distingue trois axes définitoires de l'usage langagier : les registres, les genres et les types. Un texte appartient à un registre par son utilisation d'une certaine variété de la langue. Ainsi, nous distinguons le registre familier, politique, religieux, scientifique, etc. L'appartenance d'un texte à un registre particulier se retrouve par l'usage particulier d'un vocabulaire et d'une syntaxe. Par exemple, le registre familier accepte des constructions et des vocables que le registre scientifique ne saurait tolérer.

À côté de ces considérations portant sur la variété langagière utilisée (et qui, souvent, sont assimilées aux catégorisations en genre ou en type, sans distinction précise), nous trouvons des considérations sociologiques externes au texte lui-même, considérant un texte non pas pour la variété de langue qu'il utilise, mais pour la fonction communicative qu'il remplit. Un texte appartient à un certain genre s'il remplit la fonction sociale associée à ce genre. La classification par genre consiste à catégoriser les textes « selon leur relation au monde social et selon leur conditions de production et de réception » (Gayral *et al.* 2007:4).

> "By means the concept of genre we can approach texts from the macro-level as communicative acts within a discourse network or system" (Trosborg 1997:7)

La classification par genre est de nature ouverte : les textes peuvent correspondre à autant de genres qu'ils peuvent remplir de fonctions dans notre société et la liste de fonctions que peut remplir un texte est sans limite : article de presse, recette de cuisine, publicité, petite annonce, carte postale, article scientifique, thèse, roman, nouvelle,

111 LinguaStream est libre et disponible à l'adresse : http://www.linguastream.org/

112 Notre programme permet de générer des annotations discursives. Les annotations générées sont présentées au chapitre VII. et un extrait de corpus annoté est donné en annexe K. Le sous-corpus GEOPO issu de notre corpus est disponible (version brute et annotée) à l'adresse : http://w3.univ-tlse2.fr/erss/textes/pagespersos/hodac/Corpus. Nous remercions l'IFRI d'avoir accepté la mise à disposition de ce corpus qui peut désormais être utilisé par toute la communauté scientifique. Nous remercions également Franck Sajous pour sa participation et sa disponibilité dans la mise en ligne de GEOPO.

113 Je remercie chaleureusement Marie-Paule Jacques pour m'avoir fait part de ce rapport ainsi que de l'article de Trosborg (1997).

'article' de dictionnaire, notice, rapport médical, programme politique, texte de prière, texte de loi ... Au vu de ces exemples, on comprend pourquoi aucun inventaire des genres existant n'a été établi. « Un tel inventaire est par définition impossible. En effet, dans la mesure où les genres se définissent par rapport à une pratique sociale et culturelle et que les pratiques sociales et culturelles sont en constante évolution, les genres sont en constante évolution et leur description a toujours de fait un temps de retard » (Gayral *et al*. 2007:7).

La définition des genres de textes est purement socio-culturelle. le **genre** est une « catégorie de textes fondée sur une pratique sociale établie, définie *a priori*. La catégorie est reconnue et validée par le fait qu'elle peut se dénommer. » (Gayral *et al*. 2007:6). Le genre d'un texte rappelle que le texte est un objet fonctionnel. D'un point de vue linguistique, le genre constitue une définition du texte inopérante puisqu'elle ne définit en rien la façon de construire un texte (elle définit la façon dont on va utiliser un texte). Pour faire face à cet état de fait, plusieurs projets cherchent à associer aux différents genres des traits internes au texte (Gayral *et al.* citent Malrieu & Rastier 2001 ainsi que Branca-Rosoff 1999). Il s'agit d'associer un genre de texte à un **type de texte**. Cependant, il est de bon sens de ne pas croire en une telle association, nous y reviendrons plus loin.

Le type correspond à une « catégorie de textes fondée sur l'existence de traits linguistiques communs ou d'un critère pertinent au regard d'un objectif (applicatif ou autre). Les types ne correspondent pas obligatoirement à des pratiques sociales définies. Ils peuvent émerger *a posteriori*, par l'analyse d'un corpus de texte. » (Gayral *et al.* 2007:6). cette définition fait apparaître deux façons de typer les textes. La première, empirique, relève d'un calcul purement formel de traits de surface. La seconde, plus théorique, consiste à poser des hypothèses sur le 'type' de texte, type au sens commun du terme. Nous retrouvons ici la classification en textes de « type narratif », « procédural », « expositif » etc. que nous avons utilisée pour caractériser le 'type' de texte qui nous intéresse.

> « Soit la typologie est induite du corpus, par recueil et analyse selon des méthodes statistiques d'un certain nombre de traits lexicaux, morphologiques, syntaxiques, etc. ou encore par apprentissage. [...] Soit la typologie correspond à une élaboration plus théorique, qui n'est pas nécessairement appuyée sur un corpus précis. L'objectif qui guide le chercheur est alors de rendre compte de ce que peuvent avoir en commun une variété de textes » (Gayral *et al.* 2007)

La distinction entre typologie induite et typologie théorique correspond plus ou moins à la distinction faite plus haut entre approche *data-driven* et approche *hypothesis-driven*. Notre travail se situe précisément à l'articulation de ces deux approches. La constitution de notre corpus s'est d'abord fondée sur une typologie théorique. Notre restriction au type expositif en constitue un premier pas (voir I.4.2). Mais plus encore, nos analyses se basent justement sur l'hypothèse de variations au niveau organisationnel des trois sous-corpus distingués (voir VII.1). Ce sont précisément ces variations supposées qui nous permettent, par analyses statistiques, de découvrir des indices du marquage de l'organisation de ces textes, et de réaliser, par contre-coup, une typologie induite. En transformant quelque peu les propos de Gayral *et al.*, l'objectif qui nous guide est de rendre compte à la fois de ce que peuvent avoir en commun l'ensemble des textes d'un sous-corpus et de ce que peuvent avoir en opposition les trois sous-corpus.

L'idée d'une typologie induite est largement représentée par les travaux de Biber (1988, 1995). La méthode de Biber est basée sur l'hypothèse qu'un texte porte en lui des collocations de traits qui permettent de le caractériser. Ces collocations marquent des fonctions discursives primaires relatives (i) au rôle en discours que la forme remplit (par exemple, le déterminant démonstratif signale qu'il y a nouvelle identification du référent à faire et changement de contexte) (ii) aux contraintes situationnelles et de production que cette forme reflète (par exemple, une forte présence de déictiques et de pronoms à la première et deuxième personne pour le discours face-à-face) (iii) aux contraintes de

production que cette forme reflète (par exemple, des formes lexicales de l'hésitation pour un discours en direct) et (iv) aux aspects sociaux ou situationnels que cette forme connote (exemple des régionalismes).

> "[Biber's Method] is based on the assumption that frequently co-occurring features have at least one shared communicative function. It is claimed here that there are relatively few primary functions in English, and that frequent co-occurrence of a group of linguistic features in text is indicative of an underlying function shared by those features." (Biber 1988:63-64)

Pour évaluer son hypothèse, Biber procède à des analyses factorielles qui permettent de regrouper un grand nombre de traits dans des dimensions alors représentatives d'un comportement général du texte étudié. Ces dimensions ne sont pas construites *a priori*, mais élaborées selon des méthodes statistiques qui regroupent un certain nombre de traits linguistiques suivant leur fréquente collocation en corpus (pour un description détaillée de sa méthode se reporter à Biber 1988:79-85). Biber (1988) analyse 67 traits linguistiques au travers de 481 extraits de textes écrits et parlés. Cette expérience est réitérée sur le Coréen, le Tuvaluan[114] et de Somalien (Biber 1995). Cinq dimensions principales émergent de la première analyse. Ces dimensions sont des échelles de mesure qui présentent deux pôles opposés. La première distingue l'aspect informatif de l'aspect plus impliqué, la seconde l'orientation narrative *vs.* non narrative, la troisième distingue le fait que le texte relève d'une référence élaborée (*i.e.* indépendante de la situation) ou d'une référence dépendante de la situation, la quatrième dimension concerne l'expression de la persuasion selon qu'elle est déclarée ou non déclarée et la dernière oppose le style abstrait du style concret. À travers ces dimensions, une typologie des textes peut être établie selon que les textes sont plus ou moins caractérisés par chacune d'entre elles.

Face à de tels travaux, certains linguistes mettent en avant l'hétérogénéité textuelle pour réfuter toute possibilité d'un typologie des textes. Adam (1992, 1997, 1999) trouve « profondément erroné de parler de types de textes[,] l'unité texte [étant] trop complexe et trop hétérogène pour présenter des régularités linguistiquement observables et codifiables. » (Adam 1999:82). Cette position repose sur la théorie compositionnelle des séquences (Adam 1999) qui dit qu'un texte est un ensemble de séquences, chaque séquence pouvant répondre à un « fait de régularités » tel que narration, description, argumentation, explication, dialogue. Nous sommes tout à fait d'accord avec cette conception d'une hétérogénéité du discours, défendant l'idée de portions de textes différemment structurées. Cependant, la notion de type qu'il réfute ne correspond pas à notre définition puisqu'elle répond à une typologie *a priori*. La typologie définie par Adam, mais auparavant Werlich (1976) ou encore Longacre (1976, à laquelle nous avons référé pour définir les textes expositifs en I.4.2) est une typologie que l'on peut qualifier de « procédurale » en ce sens qu'elle se base sur la distinction entre différents processus de production linguistique (la narration, la description, etc.) La typologie définie par Biber est une typologie induite se basant sur la composition même des textes (*i.e.* les mots et constructions qui les constituent).

De plus, la construction d'une typologie des textes par repérage de configurations d'indices de surface ne réfute pas le fait qu'un texte soit hétérogène. C'est ce que montre Biber & Finegan (1994), en appliquant les dimensions mises en évidence dans Biber (1988) sur des articles du domaine de la recherche médicale afin de montrer la variation intra-textuelle de ces textes découpés selon quatre sections ayant des visées discursives différentes (introduction, méthode, résultats, discussion). Les auteurs concluent alors que la variation intra-textuelle de ces textes, même si elle peut être mise en évidence par l'analyse multi-dimensionnelle, est beaucoup moins grande que celle observée entre textes de « registres » différents.

114 Langue des îles Tuvalu proches des îles Fidji (Polynésie).

> "The present study shows that there are systematic linguistic differences associated with micro-purpose variation within experimental research articles, but that those differences are small relative to the full range of variation among English registers." (Biber & Finegan 1994)

La notion de registre utilisée par Biber & Finegan (mais aussi par Biber en général) correspond à une définition hyperonymique des genres. En fait, elle est issue de la classification existante dans le corpus utilisé par Biber pour élaborer sa typologie induite : un corpus construit à partir du *LOB corpus* (anglais écrit) du *LONDON-LUND corpus* (anglais oral). Son corpus est composé de 481 textes (960 000 mots) présentant 23 genres tels qu'écrits scientifiques, revues de presse, biographies, romans de science-fiction, lettres personnelles, conversations face à face, émissions de radio, etc. (se reporter à Biber 1988, 1992a , 1995 ou à Hồ-Đắc 1999 pour un résumé en français de la 'méthode Biber').

Les travaux de Biber ne cherchent pas à associer un type à un genre mais à aboutir à une classification efficace des textes. L'idée n'est pas de définir tel genre de texte par tel type, mais de dégager des configurations de traits (des dimensions) sur lesquelles peut s'appuyer une classification de différents fonctionnements textuels et discursifs. La question d'une classification par genre reste posée ; car un genre peut être associé à plusieurs modes organisationnels : plusieurs types de textes. De même, un type de texte peut être associé à plusieurs genres.

> "Genres and text types are clearly to be distinguished, as linguistically distinct texts within a genre may represent different text types, while linguistically similar texts from different genres may represent a single text type" (Trosborg 1997:12)

L'analyse proposée dans notre thèse ne cherche pas à mettre en place une typologie des textes, encore moins à associer un type de texte à un genre. Cependant, elle soutient l'hypothèse que la structuration du texte peut varier d'un texte à l'autre ainsi qu'à l'intérieur même d'un texte, en s'intéressant notamment à ce qui compose la position initiale, c'est-à-dire aux différents contenus et aux différentes formes qui tiennent cette position. Plusieurs travaux concernant la position de thème rejoignent cette idée, notamment ceux de Fries (1995b, c) qui fait les hypothèses suivantes :

> "(i) different patterns of Thematic progression correlate with different genres, *i.e.* patterns of thematic progression do not occur randomly but are sensitive to genre ; and
>
> (ii) the experiential content of Themes correlates with what is perceived to be the method of development of a text or text segment
>
> (iii) the experiential content of Themes of a text correlates with different genres, and
>
> (iv) the experiential content of the Themes of a text correlates with different generic elements of structure within a text." (Fries 1995c:319).

Cette hypothèse se retrouve également dans les travaux sur les stratégies textuelles qui accordent une grande importance à la position initiale de phrases mais aussi de paragraphe. Ainsi les (portions de) textes présentant des stratégies textuelles organisées selon une continuité temporelle tels que des écrits historiques ou des biographies se distinguent des autres par une grande fréquence d'adverbiaux temporels en position initiale.

Pour prendre en compte la variation textuelle, trois sous-corpus vont être à la base de notre exploration qui, d'après nos intuitions, se distinguent entre autres par l'utilisation plus ou moins grande de certaines stratégies textuelles. Notre analyse quantitative permettant d'établir une typologie *a posteriori* des trois sous-corpus affirmera ces différences entre groupes de textes (voir les caractéristiques de notre corpus d'étude en partie VII.1).

VI.3. Des outils pour l'analyse de corpus

Nous avons plusieurs fois précisé notre culture « TAL » qui a influencé notre méthodologie. Mais en plus de cette culture, la tendance actuelle lie indéniablement les linguistiques de corpus aux traitements automatiques. Cela devient évident lorsque la taille des corpus dépasse plusieurs centaines de milliers de mots. Cette automatisation de la constitution de nos observables entraîne un changement dans l'approche de l'objet d'étude : dans la façon de définir l'objet d'étude et dans la façon de concevoir la méthodologie d'analyse (nous y reviendrons).

Pour automatiser des processus d'annotation, il est nécessaire d'avoir une **définition opérationnelle** des éléments à annoter, c'est-à-dire une définition qui puisse être traîtée par une machine (inapte à l'interprétation). C'est cette nécessité qui peut expliquer la relative frilosité de certains linguistes vis-à-vis des méthodes de linguistiques de corpus. En effet, l'apparition de corpus annotés discursivement et d'outils pour l'annotation discursive des corpus est toute récente. Nous assistons ainsi à la naissance d'outils informatiques prenant en compte des aspects essentiels à l'organisation discursive (prise en compte de la linéarité du texte[115] et du découpage en sections et paragraphes, possibilité de faire se chevaucher les annotations, de prendre en compte simultanément plusieurs facteurs de natures très diverses, etc.)

> "commercially available corpus analysis tools are not very helpful for investigations of discourse-level features. Standard concordancing packages are designed to produce a listing of specified target words with their immediate sentential contexts. Because such tools are not designed for complex grammatical or semantic analysis, they are also not suitable for discourse analysis. For example, concordancing packages provide no means for identifying all the nouns in a text, let alone classifying those nouns as known versus new referents." (Biber *et al.* 1998:107)

Les besoins pour développer l'EOD en corpus sont de deux types. D'une part, il apparaît nécessaire de pouvoir prendre en compte simultanément des indices identifiés par leur morphologie, leur fonction syntaxique, leur localisation dans le texte, leur sens, leur proximité par rapport à un autre indice, etc. D'autre part, il manque cruellement de méthodes d'analyse de ce type de données. Une première idée est la mise en place d'outils de visualisation permettant par exemple de repérer des zones de concentration d'indices et de pouvoir naviguer dans le texte entier tout en zoomant sur certains endroits pertinents par rapport à la présence de certaines formes. Nous avons en partie défini l'EOD par cette nécessité permanente de pouvoir mettre en contexte large des traits fins[116].

Cette actuelle pauvreté d'outils TAL appliqués à l'EOD va de pair avec l'aspect 'recherche ouverte' exposé précédemment. En effet, si le linguiste n'arrive pas à définir clairement les objets qu'il veut observer, l'informaticien ne peut élaborer un outil. Cependant, la dichotomie entre linguiste et informaticien s'efface avec le temps. Ainsi, des collaborations actives entre linguistes et informaticiens aboutissent à la création de plateformes d'expérimentations (Linguastream[117] en est un bel exemple), et les TAListes deviennent peu à peu des linguistes/informaticiens ou informaticiens/linguistes.

L'utilisation de processus automatiques pour accéder aux données discursives permet un regard nouveau sur l'EOD. La possibilité de réaliser des analyses probabilistes ou d'appliquer des algorithmes d'apprentissage ouvre des nouvelles portes. Ces méthodes encore peu répandues en recherche linguistique permettent de 's'abstraire' du

115 On peut remarquer que les premiers corpus (anglais) constitués pour les linguistiques de corpus ne contenaient que des extraits de textes, l'unité d'analyse étant évidemment la phrase et non le texte.

116 À ce sujet, et c'est regrettable, il est évident que Frantext (seul gros corpus d'œuvres littéraires en langue française muni d'un outil de consultation) est difficilement utilisable pour l'EOD puisque les contextes se limitent à 300 mots autour du candidat recherché pour les textes publics, et à 300 caractères pour les textes sous droits (URL : http://atilf.atilf.fr/frantext.htm).

117 URL : http://www.linguastream.org/

contenu du texte pour envisager le texte comme une sorte de matrice de mots interconnectés. Le risque majeur de ces nouvelles méthodes est d'oublier qu'un texte est un contenu organisé, ce que nous observons dans les premiers travaux de segmentation thématique (Hearst 1994, 1996, Ferret & Grau 1998, 2000) ou dans certaines utilisations de la LSA (Kintsch 2002). Cependant, l'ouverture actuelle des laboratoires d'informaticiens et de linguistes permet d'éviter ce risque. Et il est alors parfaitement envisageable de développer ces méthodes pour l'EOD. D'ailleurs, ce sont peut-être ces méthodes qui permettront de mettre au jour les configurations d'indices *soft* que le lecteur (et l'analyste) humain n'arrive pas à circoncire et identifier. Les travaux de Teufel (1999, Teufel & Moens 2002) offrent un des plus beaux exemples de réunion entre des intuitions linguistiques, des annotations linguistiques humaines, des analyses probabilistiques, le tout pour aboutir à la détection automatique de la segmentation rhétorique générale d'articles scientifiques (voir III.1.2.a et VI.1.2).

Pour repérer des configurations d'indices signalant l'organisation du discours (définies au chapitre précédent), il faut que ces indices aient été annotés, c'est-à-dire repérés et caractérisés. Une première possibilité consiste à utiliser des corpus préalablement annotés. De telles annotations permettent la construction de patrons (ou expressions régulières) pour identifier les indices pertinents pour l'analyse. Ce type d'annotation peut s'effectuer de deux façons qui ne diffèrent que dans la forme : soit avoir recours à des outils d'analyse interactifs qui aident le linguiste à repérer et catégoriser ses observables en proposant des candidats (Biber *et al.* 1998:112-116) ; soit créer des programmes d'extraction basés sur des patrons assez génériques et réaliser des corrections manuelles sur les résultats obtenus. Dans les deux cas, il faut concevoir un outil propre à l'étude[118], ce qui implique la collaboration d'informaticiens dans le projet de recherche ou l'aptitude du linguiste à programmer des outils pour l'expérimentation.

Une autre possibilité – celle qui est testée ici – consiste à poser des hypothèses sur les variations de l'organisation du discours. Par exemple, un exposé scientifique est sans aucun doute organisé différemment d'un roman ou d'un manuel. De même, on peut fortement supposer que le rôle d'une phrase dans l'organisation du discours n'est pas le même selon que la phrase se situe en début ou en fin de section. Ces deux exemples présentent chacun un indice qui peut influer fortement sur l'identification d'une relation de discours : le type de texte et la position textuelle. Ces deux facteurs de variation constituent le point de départ de nos analyses. L'identification des configurations d'indices marquant la séquentialité du discours s'effectue alors en mesurant les variations textuelles et la position textuelle de différents traits linguistiques (dans notre cas, il s'agit des éléments présents en position initiale, cf. chapitre VII).

VI.4. Concepts et calculs statistiques

Le texte n'étant pas un sac de phrases anarchique, l'observation des éléments qui le constituent doit tenir compte de l'organisation qui le sous-tend. Une analyse quantitative ne peut se baser uniquement sur la fréquence des données. Il faut situer les données, prendre en compte leur contexte linguistique et extra-linguistique. Il est alors possible de mesurer des regroupements, des variations significatives.

Pour mesurer des variations, un modèle théorique est nécessaire. Ce modèle constitue un moyen pour calculer l'écart entre les données effectivement observées et les données théoriques. Les données théoriques répondent à l' « hypothèse nulle », c'est-à-dire à l'hypothèse que la fréquence du phénomène dans l'ensemble du texte et sa

118 La phase d'annotation des projets de corpus annotés discursivement cités plus haut (le *RST corpus*, le *Penn Discourse Treebank corpus*, le *Timebank corpus*) s'effectue toujours par 'orientation' de l'annotateur humain. Un outil d'annotation est généralement utilisé, outil qui peut proposer des annotations probables aux annotateurs afin de les guider dans leur travail. Pour plus de précision, voir les articles de référence cités ainsi que les URL données.

fréquence dans le sous-ensemble considéré sont proportionnelles. L'hypothèse nulle consiste donc à supposer que la répartition des données est aléatoire : les fréquences dépendent seulement de la taille des 'parties'[119] considérées.

Pour prouver par une enquête statistique que telle ou telle configuration n'est pas due au hasard, quatre étapes sont nécessaires :

1. « la construction d'un modèle théorique ;
2. l'observation de la répartition réelle et des écarts qui existent entre celle-ci et le modèle théorique ;
3. l'application à ces écarts d'un test statistique qui les appréciera en probabilité ; si la probabilité est forte, les écarts seront déclarés non significatifs : l'hypothèse nulle ne peut être rejetée et aucune conclusion linguistique ou stylistique ne peut être retirée de l'expérience ; si la probabilité est faible, les écarts ne peuvent être attribués au seul jeu du hasard : le fait linguistique ou stylistique existe ;
4. l'interprétation des écarts entre modèle et observation, si ceux-ci ont été reconnus significatifs. » (Muller 1968:44)

VI.4.1. Constitution du modèle théorique : fréquence, proportion moyenne

La première étape de l'investigation statistique consiste à collecter les fréquences (en nombre d'occurrence) des différentes variables à l'étude (la liste de ces variables *i.e.* de nos observables est donnée en VII.2). Cette collecte des fréquences nous permet, pour chaque variable, d'obtenir sa proportion moyenne d'apparition. Pour illustrer ces deux notions, prenons trois variables relatives à la forme du sujet grammatical des phrases :

- **PRO3** équivaut au nombre de phrase comportant un sujet grammatical de forme pronom personnel de 3e personne (PRO3 = *elle(s)* + *il(s)* – les *il* impersonnels) ;
- **NP** équivaut au nombre de phrase comportant un sujet grammatical de forme nom propre ;
- **SNindef** équivaut au nombre de phrase comportant un sujet grammatical de forme SN indéfini.

Le tableau VI.1 indique la fréquence et la proportion moyenne de ces trois observables dans notre corpus.

	Corpus entier	
	Fréquence (Nb de phrases)	***Proportion (%)***
PRO3	2 221	*9,6*
NP	1 592	*6,9*
SNindef	1 447	*6,2*
Autres Sujets	17 957	*77,3*
Nb total de phrases	23 217	*100,0*

Tableau VI.1 : fréquence et proportion moyenne

Sur les 23 217 phrases de notre corpus, la variable PRO3 affiche une **fréquence** de 2 221 occurrences (2221 phrases de notre corpus contiennent un sujet grammatical de forme PRO3), ce qui représente 9,6% des phrases (2221/23217). L'hypothèse nulle revient à considérer que cette proportion est identique quel que soit le contexte d'apparition de la phrase. Ainsi, dans le modèle théorique de cette expérience, n'importe quel corpus, texte, section, ou paragraphe présente 9,6% de phrases ayant pour sujet grammatical un PRO3. De même, si l'on ne prend que les premières phrases de chaque texte, de chaque section ou de chaque paragraphe. 9,6% équivaut à la **proportion moyenne** de la variable PRO3. Il paraît immédiatement évident que cette hypothèse nulle est impossible, et cela, car on fait l'hypothèse que le fait d'avoir pour sujet grammatical un PRO3 n'est pas de l'ordre du hasard mais répond à des stratégies textuelles et plus généralement, à des procédés discursifs de continuité référentielle. Nous pouvons effectuer le même commentaire pour les NP et les SNindef.

119 Le terme 'partie' n'est pas, ici, à mettre en relation avec la notion de segment de texte délimité par des intentions auteuriales. Il peut s'agir par exemple de comparer la fréquence des observables contenus dans deux fenêtres de *n* lignes délimitées au hasard.

VI.4.2. Observation des écarts au modèle théorique

Observons maintenant les écarts que peut prendre la fréquence d'apparition de nos trois variables selon la position de la phrase dans le paragraphe : première phrase de paragraphe (P1), deuxième phrase et plus (P2&+) et dernière phrase (DP). Le tableau VI.2 indique, pour chaque variable et selon les différentes positions textuelles envisagées, leur fréquence d'apparition en sujet grammatical (*f*) et la proportion que cela représente *(p)*.

	P1		P2&+		DP		Toutes phrases	
	f	p	f	p	f	p	f	p
PRO3	185	*0,035*	1 541	*0,114*	495	*0,111*	2 221	*0,096*
NP	407	*0,077*	963	*0,071*	222	*0,050*	1 592	*0,069*
SNindef	293	*0,055*	871	*0,065*	283	*0,064*	1 447	*0,062*
Autres sujets	4 398	*0,833*	10 112	*0,750*	3 447	*0,775*	17 957	*0,773*
Nb total de phrases	5 283	*1,000*	13 487	*1,000*	4 447	*1,000*	23 217	*1,000*

Tableau VI.2 : Répartition de quelques formes de sujet grammatical selon la position de la phrase dans l'unité paragraphe

Nous voyons immédiatement une variation de ces proportions selon les différentes positions paragraphiques. En effet, si l'on observe les proportions d'apparition, nous voyons qu'il y a à peu près trois fois moins de PRO3 en position P1 que dans le modèle théorique (*i.e.* dans toutes les phrases confondues) et proportionnellement plus de PRO3 dans les positions P2&+ et DP. Les NP et les SNindef qui affichent un proportion moyenne à peu près égale (6,9% et 6,2%) montrent des variations différentes : moins de NP en dernière phrase de paragraphe et plus ailleurs ; moins de SNindef en initiale de paragraphe et un peu plus ailleurs. Ce que nous faisons ici, c'est simplement une comparaison des proportions d'apparition. Cette comparaison ne nous permet pas de mesurer si la variation observée est significative ou pas.

VI.4.3. Constitution de données théoriques et test de signifiance : test de l'écart réduit

Pour mesurer l'écart entre la fréquence réelle d'apparition des PRO3 et leur fréquence théorique (selon l'hypothèse nulle), il faut calculer les données théoriques. Ce calcul consiste, pour chaque mesure (chaque cellule du tableau), à multiplier le nombre de configurations à prendre en compte (e.g. le nombre de phrases composant le sous-ensemble considéré) par la proportion moyenne d'apparition de la variable. Le tableau VI.3 permet une comparaison des fréquences réelles et des fréquences théoriques. La partie de droite indique la différence entre fréquence réelle et fréquence théorique. Par exemple, selon l'hypothèse nulle, il y aurait 505 PRO3 en P1 ((2 221 / 23 217) x 5 283). Les effectifs réels étant de 185, nous constatons un écart de -320 (185-505).

	Fréquence réelle			Proportion moyenne	Fréquence théorique			Écart entre fréquence réelle et théorique		
	P1	P2&+	DP		P1	P2&+	DP	P1	P2&+	DP
PRO3	185	1 541	495	*0,096*	505	1 290	425	-320	+251	+70
NP	407	963	222	*0,069*	362	925	305	+45	+38	-83
SNindef	293	871	283	*0,062*	329	841	277	-36	+30	+6
Autres sujets	4 398	10 112	3 447	*0,773*	4 086	10 431	3 439	+312	-319	+8
Total de phrases	5 283	13 487	4 447	*1,000*	5 283	13 487	4 447	0	0	0

Tableau VI.3 : Répartition de quelques formes de sujet grammatical selon la position de la phrase dans l'unité paragraphe

La mesure de ces écarts ne nous permet pas encore d'estimer si les écarts observés sont significatifs c'est-à-dire si les écarts sont le fruit du seul jeu du hasard ou non. Si l'on se concentre sur les écarts indiqués dans le tableau VI.3, nous remarquons par exemple que, en position P1, les PRO3 et les autres sujets montrent un écart de même valeur absolue (-320 et +312). Cependant, il y a beaucoup plus d'autres sujets que de PRO3 si l'on observe la proportion moyenne de ces deux variables. Nous avons besoin d'une mise à plat des écarts afin de pouvoir comparer les écarts entre eux et définir leur degré de signifiance.

Plusieurs tests existent pour mesurer le degré de signifiance d'un écart. Pour notre étude, nous utilisons le test de l'écart réduit – *z* (*z-score* en anglais). Nous aurions pu également utiliser le test du Khi², plus répandu dans la communauté de statistique linguistique. Cependant, le test de l'écart réduit nous est apparu plus simple à comprendre, à mettre en oeuvre et à interpréter. Le test de l'écart réduit nous permet d'avoir directement pour chaque variable une mesure de la signifiance de sa déviation par rapport au modèle théorique. À l'inverse, le test du Khi² donne en premier lieu un degré de corrélation global entre deux ensembles de variables (e.g. la forme du sujet grammatical et la position paragraphique de la phrase) qui permet de rejeter ou non, de façon globale, l'hypothèse nulle. Il faut ensuite rentrer dans les détails du calcul pour trouver la contribution (positive ou négative) de chaque variable à ce degré de corrélation, chose qui nous semblait assez délicate au début de notre initiation à la statistique.

La mesure l'écart réduit est égale au rapport entre l'écart réel à l'écart type d'une variable dans un sous-ensemble défini.

$$z = \frac{(\textit{fréquence rélle} - \textit{fréquence théorique})}{(\textit{écart type})}$$

Figure VI.1 : Calcul de l'écart réduit (*z*)

L'écart réel correspond à la différence observée entre la fréquence réelle et la fréquence théorique (telle que indiquée dans le tableau VI.3). L'écart-type correspond à la moyenne des écarts par rapport la moyenne ou, en d'autres termes, à l'écart moyen que l'on peut mesurer par rapport à la (proportion) moyenne. Son calcul correspond généralement à la racine carrée de la variance, *i.e.* de la moyenne des carrés[120] des écarts observés pour chaque entrée entre sa valeur et la moyenne. Dans nos analyses, nous utilisons une formule simplifiée de l'écart-type telle que fournie par Müller (1968:79) ou Manning & Schütze (1999:51). Cette formule de l'écart-type correspond à celle utilisée dans les situations de loi binomiale où chaque variable est considérée comme étant indépendante et où l'on veut comparer des probabilités d'apparition (vs. d'absence) d'un événement. Par exemple, dans le cas des PRO3, on considère que la probabilité de voir apparaître un PRO3 en position sujet est aléatoire. Quelle que soit le contexte de la phrase, il y a 9,6 chances sur 100 que son sujet soit un PRO3. Nous savons bien évidemment que, d'un point de vue linguistique, le contexte de la phrase est un facteur décisif de l'emploi d'un PRO3 en sujet grammatical (il faut entre autres choses, que le référent sujet ait été précédemment activé, voir partie V.4.3.a). Nous avons choisi de faire table rase de ces contraintes linguistiques afin, justement, de mettre au jours les contextes d'apparition de tel ou tel trait linguistique.

La figure VI.2 nous donne la formule de cet écart-type :

Soit *n* le nombre de mesures effectuées (pour notre étude, *n* correspond au nombre de phrases prises en compte) et *p* la proportion moyenne d'avoir *x* (*i.e.* la proportion moyenne estimée par $p = \frac{\text{nombre de x}}{\text{nombre de phrase}}$),

$$\sigma = \sqrt{n \times p \times (1 - p)}$$

Figure VI.2 : Formule simplifiée de l'écart-type théorique (σ) selon Müller (1968:79)

En appliquant cette formule simplifiée au cas des PRO3 en position P1, nous obtenons un écart-type de :

120 Les carrés sont toujours positifs, ce qui empêche aux écarts négatifs de compenser les écart positifs ou inversement.

$$21{,}41 = \sqrt{5283 \times 0{,}096 \times (1 - 0{,}096)}$$

Ce qui signifie que, d'après le modèle théorique, pour un échantillon de 5 283 phrases et concernant uniquement les PRO3, un écart d'environ 21,4 phrases par rapport à la fréquence théorique (505) est en moyenne observé. En d'autres termes, nous pouvons dire que selon le modèle théorique, il y aurait 505 phrases en position P1 dont le sujet est de forme PRO3 plus ou moins 21,5 phrases.

Il nous reste maintenant à diviser l'écart réel mesuré par rapport à l'écart-type calculé pour chaque fréquence, afin de connaître l'écart réduit associé à telle variable dans tel sous-ensemble. Dans le cas des PRO3 en P1, nous obtenons un écart réduit de : $z = \frac{(185 - 505)}{21{,}41} = \frac{-320}{21{,}41} = -15$.

Le tableau VI.4 nous montre les résultats de ces trois étapes de calcul (écart réel, écart-type et écart réduit) pour toutes les variables.

	Écart réel			*Écart-type (σ)*			*Écart réduit (z)*		
	P1	***P2&+***	***DP***	***P1***	***P2&+***	***DP***	***P1***	***P2&+***	***DP***
PRO3	-320	+251	+70	21,4	34,2	19,6	-15,0	7,3	3,5
NP	+45	+38	-83	18,4	29,3	16,9	2,4	1,3	-4,9
SNindef	-36	+30	+6	17,6	28,1	16,1	-2,1	1,1	0,4
Autres sujets	+312	-319	+8	30,4	48,6	27,9	10,3	-6,6	0,3

Tableau VI.4 : Répartition de quelques formes de sujet grammatical selon la position de la phrase dans l'unité paragraphe

La valeur *z* est ensuite projetée dans une table (tableau VI.5) qui indique la probabilité d'atteindre ou de dépasser un tel écart réduit. Par exemple, il y a une probabilité de 0,01 (*i.e.* une chance sur 100) d'atteindre un écart réduit (positif ou négatif) de 2,576.

p	1	0,1	0,05	0,01	0,001	0,000 1	0,000 01	0,000 001	0,000 000 1	0,000 000 01	0,000 000 001
z	0	1,6	2	2,5	3,3	3,9	4,4	4,9	5,3	5,7	6,1

Tableau VI.5 : table pour les écarts réduits, i.e. probabilité *p* d'atteindre un écart réduit *z*

Pour nos analyses, nous considérons que tout écart réduit supérieur à 2,5 est significatif. En d'autres termes, nous considérons qu'un écart qui a 1 chance sur 100 (ou moins) d'être le simple jeu du hasard est certainement lié à une contrainte linguistique[121]. Ainsi, nous considérons que toutes les variations de la fréquence de PRO3 selon les différentes positions paragraphiques sont significatives. À l'inverse, les variations concernant les phrases présentant un sujet grammatical de forme SNindef ne sont pas interprétées comme relevant de quelque contrainte discursive, mais comme relevant essentiellement du hasard.

L'analyse des variations de la composition de la position initiale s'effectuera essentiellement grâce à ce test qui nous permet de caractériser les écarts observés par leur polarité (au signe – correspond le fait qu'il y a moins de *x* que théoriquement attendu) et par leur degré de signifiance. Il est ensuite de notre seul ressort de faire un lien entre ces résultats statistiques et la validation d'une hypothèse linguistique, concernant par exemple la fonction discursive des initiales de paragraphes (P1), *i.e.* des changements de paragraphe.

La difficulté majeure des analyses quantitatives relève de la lecture des données et de leur mesure. Les phénomènes discursifs sont si complexes et la statistique linguistique encore si peu développée à ce niveau que peu de modèles peuvent nous servir d'appui. En nous basant sur

121 Les sciences humaines considèrent généralement qu'une probabilité de 0,05 constitue un seuil suffisant. Nous avons préféré un seuil plus bas afin de palier deux choses : les erreurs d'annotations que notre programme peut générer et le fait que l'on se situe dans une loi binomiale qui ne fait état d'aucune contrainte linguistique. De plus, nos résultats avec un seuil de 0,01 sont largement suffisants pour mettre au jour des variations pertinentes quant à la position initiale dans l'organisation du discours.

des calculs relativement simples et surtout non opaques, c'est plus une façon de comprendre les quantités qu'une utilisation d'outils statistiques précis que nous avons présentée ici. Nous aurions pu utiliser des calculs plus complexes et plus communs en statistique linguistique, comme le test du Khi² ou les analyses multi-factorielles. Cependant, la grande hétérogénéité des données et surtout notre difficulté à comprendre la portée de telles mesures nous ont poussée vers des mesures statistiques simples. Le test de l'écart réduit est apparu comme le seul test adapté à notre compétence nous permettant de mettre en relation nos résultats et la réalité qu'ils mesurent, *i.e.* d'interpréter nos résultats et de nous sentir libre de 'jouer' avec les facteurs de variation.

Chapitre VII
À la recherche des configurations d'indices : méthodologie d'investigation

Sommaire

Notre méthodologie consiste à chercher des variations pertinentes par rapport à l'étude de l'organisation du discours. Les variations observées dans notre corpus naissent du jeu entre trois facteurs textuels différents : le type de texte (le sous-corpus), la position textuelle (S1, P1 ou P2 ; voir IV.5) et la cooccurrence des formes en position initiale. Les variations observées nous instruisent sur les configurations d'indices caractéristiques d'un type de texte ou d'une position textuelle dans un type de texte. Ainsi, nous confrontons nos données aux associations théoriques établies entre (1) modes organisationnels et types de texte, (2) stratégies de séquentialité et position textuelle et (3) signalement de la séquentialité et formes de surface.

Nos analyses mesurent les variations significatives (dans notre étude, l'écart réduit – *z*) selon différents points de vue, afin de ne pas associer trop rapidement une forme à une fonction. En effet, comme nous l'avons précisé au

chapitre V, il n'y a pas de corrélations absolues entre un trait et une fonction. Ainsi, pour faire émerger des marqueurs discursifs, il faut s'assurer de l'influence conjointe des indices qui constituent ce marquage.

Par exemple, observer une concentration significative de circonstants temporels en initiale de sections ne suffit pas pour statuer sur la fonction de marquage de discontinuité des circonstants temporels. Il faut se demander si la position textuelle ne joue pas à elle seule ce rôle*, i.e.* il faut vérifier si la présence d'un circonstant temporel en initial correspond effectivement à une marque de discontinuité, quelle que soit la position textuelle. Il faut également mesurer si ce rôle n'est pas restreint à un seul type de texte. De même, nous pouvons tester la fonction de rupture d'un changement de section (qui nous semble *a priori* indéniable) en observant les éléments qui composent la position initiale des S1. Ainsi, chaque trait linguistique est à mettre en corrélation avec un autre trait relatif au sous-corpus, à la position textuelle – PosTxt – et à ce que nous avons appelé la « cohabitation en position initiale ».

Pour mesurer les variations en position initiale, la technique de base est d'observer :

- les écarts réduits des distributions des différentes formes selon :
  - le niveau organisationnel : étude de trois positions textuelles
  - la variation textuelle : étude de trois sous-corpus
- les cohabitations entre les différents composants de la position initiale (éléments détachés *vs.* intégrés)*, i.e.* les écarts réduits qu'entraîne la présence d'un composant sur la nature des autres composants
- les écarts réduits de ces cohabitations selon les mêmes facteurs qu'en premier point :
  - la position textuelle
  - le sous-corpus

Notre travail d'investigation comprend plusieurs étapes :

1. définir l'objet d'étude, c'est à dire le phénomène linguistique à étudier : ici le marquage de la segmentation textuelle en position initiale,
2. poser les hypothèses quant au fonctionnement de cet objet,
3. réunir un corpus dans lequel il est pertinent d'observer ce phénomène,
4. mettre en place des méthodes de repérage et de caractérisation des données à étudier, ce qui peut nécessiter l'utilisation d'outils de TAL,
5. définir des méthodes d'analyse qualitative et/ou quantitative permettant de répondre aux hypothèses posées,
6. confronter les résultats de l'analyse aux hypothèses de départ et
7. faire émerger, le cas échéant, de nouvelles hypothèses.

Ce chapitre a pour objet de mettre au clair les décisions d'ordre pratique par rapport aux choix théoriques effectués pour répondre aux hypothèses posées précédemment. Nous présentons en premier lieu le corpus construit pour l'étude. Pour effectuer nos analyses, une phase d'automatisation et donc d'informatisation a été nécessaire, ce que présente en partie la seconde partie de ce chapitre. Enfin, nous dressons un panorama des analyses effectuées et proposons un manuel de lecture des résultats.

VII.1. Caractéristiques du corpus d'étude

Notre corpus rassemble des textes de français écrit ayant pour visée discursive globale la présentation d'information. Le premier critère de constitution de notre corpus a donc été de choisir des textes expositifs (I.4.2). Ce critère relève d'une typologie essentiellement théorique. Toujours dans une typologie théorique, nous distinguons trois sous-corpus. Cette distinction s'appuie sur nos intuitions quant aux stratégies textuelles utilisées dans les textes de ces différents sous-corpus. Les trois sous-corpus peuvent être différenciés sur trois points :

- une fonction sociale différente,

- une mono-référentialité *vs.* une pluri-référentialité,
- une organisation par une localisation spatiale et temporelle forte *vs.* modérée.

Le premier sous-corpus utilisé est composé de trois textes de géographie sociale accompagnés de cartes[122]. Nous l'appelons ATLAS. Les textes d'ATLAS ont pour sujet la répartition d'un ou plusieurs phénomènes sociaux dans une zone géographique déterminée et dans une période temporelle plus ou moins déterminée. Ces textes peuvent être comparés à des documents de travail qui cherchent à faire un état des lieux d'un phénomène social, en prenant en compte tous les facteurs de variation socialement possible, afin de rassembler toutes les données d'un phénomène dans un ouvrage qui se doit d'être le plus lisible possible. « La France scolaire : de la maternelle au lycée » (Hérin & Rouaut 1994) constitue un atlas qui fait état du système scolaire en France au cours des cinquante dernières années. « Quarante années d'évolution politique de l'Ouest de la France » (Buléon, non publié) décrit les tendances politiques de l'Ouest de la France de 1960 à 2000. Le dernier texte intitulé « ATLAS transmanche » consiste en une collection de documents géographiques ayant pour sujet commun les phénomènes humains et sociaux recensés dans la zone transmanche ces dernières années[123]. Ces textes montrent une proportion très importante d'adverbiaux spatiaux et, dans de moindres mesures, d'adverbiaux temporels. Leur contenu semble se structurer principalement autour d'hyperthèmes et de localisations spatiales et temporelles. Ils présentent une forte organisation en cadres spatiaux et temporels, à l'intérieur desquels évoluent différentes progressions thématiques à thèmes dérivés.

Parallèlement à ce sous-corpus d'un type assez particulier, deux autres sous-corpus ont été constitués. Ces deux sous-corpus nous permettent d'explorer des textes présentant une organisation plus construite autour de continuités référentielles et moins dépendante des dimensions spatiale et temporelle.

Le sous-corpus PEOPL se compose de 32 portraits de personnages célèbres issus de l'Encyclopædia Universalis (quatrième édition, 1995). Ce sous-corpus est très homogène, étant issu d'un même ouvrage suivant une même ligne éditoriale. Nous avons retenu les plus longs portraits présentant des titres de sections. Naturellement, les textes composant ce sous-corpus sont mono-référentiels : ils concernent la même entité (dans notre cas, un être humain) du début à la fin. Ils peuvent également présenter une forte organisation temporelle.

Le sous-corpus GEOPO rassemble des textes publiés par l'Institut Francais des Relations Internationales – IFRI[124] concernant des problèmes de géopolitique actuelle. Ce sous-corpus est le plus 'tout-terrain' de notre corpus. Les textes appartiennent au registre journalistique et développent davantage une argumentation que les deux autres. Les entités principales sont des personnages contemporains et des phénomènes de société. Ce sont donc des textes pluri-référentiels qui peuvent, par endroits seulement, présenter une organisation spatiale ou temporelle.

---

122 En général, le contenu des textes peut être compréhensible sans les cartes. Il faut cependant noter l'interaction entre le texte et les cartes (Enjalbert & Gaio 2004). Ainsi, de nombreux énoncés renvoient et s'appuient sur les cartes :
Il peut être intéressant juxtaposer cartes des mouvements aériens avec celles des dessertes autoroutières et ferroviaires. On observe alors que la première est le négatif de la seconde. En effet, les mouvements aériens entre Paris et l'Est de la Zone Transmanche sont peu nombreux mais cette dernière possède un réseau autoroutier dense et un réseau ferroviaire à grande vitesse en fort développement, alors qu'il se produit l'inverse dans la partie Ouest.[ATLAS_1]
Ces deux cartes traduisent la permanence des structures spatiales du recrutement, même si quelques glissements sud-nord sont observés vers les régions intermédiaires (Poitou, Franche-Comté, Centre...), où l'on se tourne de plus en plus vers l'enseignement.[ATLAS_2]

123 L'ATLAS TransManche est une collection de documents géographiques en perpétuelle construction (la version utilisée date du 15 décembre 2003). Cette collection est de forme assez hétérogène puisqu'il n'y a pas de ligne éditoriale, les géographes pouvant insérer de leur propre chef une nouvelle page dans ce recueil. Pour l'étude, seuls les textes présentant des titres de sections ont été conservés ; les textes constitués à 90% de tableaux et/ou graphiques n'ont pas été utilisés (Les titres des articles retenus sont indiqués en annexe B).

124 L'Institut Français des Relations Internationales – IFRI est un centre indépendant européen de recherche, de rencontres et de débat sur les questions internationales. Il a été créé en 1979 par Thierry de Montbrial et se distingue par son indépendance (sans tutelle administrative et sans affiliation à un parti politique).

Dans une optique de classification par type, ces trois sous-corpus remplissent une fonction différente, même si celle-ci concerne toujours la communication d'informations. Les textes d'ATLAS constituent plus des documents de travail pour les géographes ou les historiens. Ils présentent des données et les décrivent sans spécialement développer d'argumentation sur le sujet. Les textes de PEOPL relèvent de l'encyclopédie. Ils cherchent à décrire et raconter la vie d'un homme[125] sous tous ses aspects, afin d'être le plus complet possible. GEOPO rassemble des textes qui proposent plus un échange d'idées qu'un simple exposé d'information. Le lecteur va y trouver un point de vue sur un fait, fait qui sera accompagné de données, de descriptions, de portraits, etc. mais toujours avec un aspect argumentatif ou polémique.

Le tableau VII.1 présente certaines caractéristiques externes aux textes des trois sous-corpus, caractéristiques relatives aux conditions de production, de réception, à la fonction sociale et aux domaines touchés par les textes des différents sous-corpus.

	ATLAS	***GEOPO***	***PEOPL***
Canal	Écrit		
Format	Publié ou sur site Web	Publié (revues papiers et électroniques)	Publié (encyclopédie papier et CD-ROM)
Cadre	Institutionnel		
Destinataire	Non compté Absent Interaction nulle Connaissances partagées non spécifiques		
Emetteur	Géographe(s) dont l'identité est pour certains connue	Journalistes, économistes, historiens, membres de l'IFRI ou collaborateurs	Comité de rédaction de l'Encyclopædia Universalis©
Fonction sociale	Document de travail	Article d'Information	Encyclopédie
Domaines	Géographie humaine	Actualité géopolitique	Personnages célèbres

Tableau VII.1 : Paramètres extralinguistiques des sous-corpus d'étude

Face à ces caractéristiques externes et en relation aux caractéristiques internes théoriques présentées jusqu'ici, certains traits formels ont été observés qui permettent de définir grossièrement et de façon induite, les trois sous-corpus utilisés. Il y a d'un côté des informations matérielles concernant la longueur des textes et des unités textuelles et de l'autre, des informations soutenant nos intuitions quant à l'aspect mono-pluri référentiel des textes ainsi qu'à l'importance de l'expression de localisations spatiales et temporelles.

VII.1.1. Caractéristiques observées des sous-corpus

VII.1.1.a) Taille, granularité et format du corpus

Comme il a été expliqué précédemment, notre analyse nécessite la prise en compte du texte dans son entier. Un autre critère de choix est lié à la volonté de prendre en compte le rôle du titre dans la structure discursive. Notre corpus rassemble donc des textes entiers relativement longs (présentant des titres de sections) pour lesquels l'information

125 Il n'y a pas de portraits de femme dans PEOPL. Non pas par opposition au principe de parité mais parce que les textes longs trouvés dans l'Encyclopædia Universalis© (quatrième édition, 1995) semblent 'réservés' aux hommes.

concernant la segmentation en sections, paragraphes et phrases est conservée, conformément aux unités textuelles définies typo-dispositionnellement.

Nous rappelons que les sections sont délimitées par deux titres, quel que soit leur niveau. Les paragraphes sont délimités par deux sauts de ligne (plus précisémnt appelés « alinéas ») sauf s'il y a structure énumérative auquel cas l'ensemble de l'énumération, *i.e.* l'ensemble des items[126], est considérée comme un paragraphe. Les unités phrases sont délimitées automatiquement par Syntex qui prend grosso-modo tout ce qui se trouve entre deux ponctuations finales fortes : un changement de paragraphe, la présence d'un point d'interrogation, d'exclamation ou simplement d'un point final (avec des exceptions, comme après une initiale). Le tableau suivant illustre notre délimitation et notre décompte des sections et des paragraphes.

section	paragraphe	texte
1	1	***LES PORTS DE PLAISANCE DU LITTORAL FRANÇAIS DES MERS DE LA MANCHE ET DU NORD*** [titre niveau 1]
	2	*La navigation de plaisance accessible au plus grand nombre est une activité récente comparée à d'autres qu'elle côtoie dans les espaces maritime et portuaire. A l'origine [...].*
	3	*Le champ géographique considéré regroupe le littoral français des mers de la Manche et du Nord. Cependant il est apparu indispensable d'intégrer les données relatives aux îles [...].*
	4	*Les critères et les éléments retenus appellent quelques commentaires préalables :*
	5	** Pour des raisons de lisibilité cartographique et de disponibilité des données mobilisables, il a semblé souhaitable de circonscrire l'enquête aux ports de plaisance à flot de 300 places et plus. Ce choix [...] ;* ** Certains ports, essentiellement en Bretagne, n'ont pu fournir d'informations concernant leurs visiteurs. Ceci est regrettable mais [...] ;* ** Dans les ports bretons à l'Ouest de St-Malo, le nombre de visiteurs est, en l'absence de données précises, un montant estimé par les responsables des ports eux-mêmes ;* ** Enfin, il faut signaler que les données sont globalisées pour un certain nombre de sites. Ainsi [...].*
2	6	***L'importance des facteurs insulaire et frontalier*** [titre niveau 2]
	7	***Les têtes de pont pour les relations transmanche*** [titre niveau 3]
	8	*Les données fournies doivent être manipulées avec précaution tant la diversité des modalités de décompte, précis ou approché, en valeur absolue ou en pourcentage, distinguant les Anglo-normands des Britanniques ou non, pourrait entraîner d'interprétations hasardeuses. Cependant, [...].*
À partir d'ici, les paragraphes ne sont plus représentés dans leur totalité. Il y a généralement plus d'un seul paragraphe par section.		
3	9	***Les ports du nord de la France*** [titre niveau 3]
	10	*Les ports de plaisance de la région Nord-Pas de Calais représentent [...]*
4	11	***Bretagne occidentale et Baie de Seine : des secteurs à développement contrasté*** [titre niveau 2]
	12	*De part et d'autre du golfe normand-breton, on trouve [...]*
5	13	***ANALYSE DES RELATIONS PAR LES FLUX ROUTIERS ENTRE CHERBOURG, CAEN, ROUEN ET LE HAVRE EN 1996*** [titre niveau 1] *[...] [ATLAS_1]*

Tableau VII.2: Délimitation et décompte des sections et paragraphes

En linguistique de corpus, il est d'usage de caractériser les corpus par leur nombre de mots. De fait, si l'on souhaite effectuer des études comparatives entre sous-corpus, l'unité de comparaison est le mot. Cependant, dans ce

126 Les amorces sont laissées à l'extérieur des paragraphes de type « énumération » car elles n'apparaissent pas forcément dans un paragraphe isolé. Elles peuvent tout à fait constituer la dernière phrase d'un paragraphe qui, dans son entier, ne peut être considéré comme l'amorce de l'énumération. Ce choix est appuyé par un soucis de régularité et par le fait que, d'un point de vue théorique, l'énumération est le seul objet nécessaire aux structures énumératives.

travail, l'unité pertinente n'est pas le mot mais la phrase puisque nos observables sont les éléments en position préverbale (peu importe le nombre de mots que contient la phrase) et que notre unité textuelle minimale est la phrase. Les trois sous-corpus sont donc comparables en nombre de phrases. Nous remarquons cependant que le nombre de mots est également comparable. Il est également indispensable de mettre en parallèle ce nombre et celui des autres unités pertinentes ici : les paragraphes et les sections.

Corpus	Textes	Titres[127]	Sections[128]	Paragraphes	Phrases	Mots
ATLAS	3	531	472	2 222	7 592	204 505
GEOPO	32	379	376	1 534	7 901	247 217
PEOPL	30	389	381	1 332	7 724	219 705

Tableau VII.3 : Description quantitative générale des sous-corpus d'étude

	Paragraphes par texte		Phrases par texte	
Corpus	Moyenne	Écart-type	Moyenne	Écart-type
ATLAS	770	226	2531	721
GEOPO	50	38	247	138
PEOPL	46	14	257	93
	Paragraphes par section		Phrases par paragraphe	
Corpus	Moyenne	Écart-type	Moyenne	Écart-type
ATLAS	5	5	3	2,5
GEOPO	4	3	5	3,5
PEOPL	4	3	6	4,0

Tableau VII.4 : Description quantitative moyenne des sous-corpus d'étude et écarts-types correspondant[129]

Cette description quantitative nous permet de distinguer dans notre corpus deux tendances formelles relatives à la longueur des textes et à leur découpage en paragraphes. Le découpage des sections en paragraphes semble beaucoup moins sujet aux variations entre sous-corpus. Dans la première catégorie, nous observons des textes assez longs aux paragraphes un peu plus courts (ATLAS) ; dans la seconde, des textes plus courts aux paragraphes relativement plus longs (GEOPO et PEOPL). Dans la première catégorie, ATLAS est composé de textes d'environ 2 500 phrases, soit dix fois plus que les textes de GEOPO et PEOPL. Par contre, les textes d'ATLAS sont découpés en paragraphes présentant, en moyenne, moins de 4 phrases, alors que les paragraphes dans GEOPO ou PEOPL sont constitués d'au moins 5 phrases. Cette différence de granularité n'est pas à négliger, surtout si des marqueurs de segmentation tels que les changements de paragraphe font partie des observables pris en compte dans l'analyse.

Nous trouvons également des distinctions entre les sous-corpus au niveau des écarts par rapport à la moyenne du nombre de phrases ou de paragraphes par texte. Un faible écart-type indique une certaine régularité dans le découpage typo-dispositionnel des textes d'un sous-corpus alors qu'un écart-type élevé indique une grande hétérogénéité. PEOPL est le sous-corpus qui semble le plus homogène. Les textes de PEOPL sont composés en moyenne de 46 paragraphes avec un écart-type relativement faible comparé à celui de ATLAS ou GEOPO (l'écart-type de 14 signifie qu'en moyenne, les textes montrent 46 paragraphes plus ou moins 14). GEOPO montre une composition beaucoup plus hétérogène. La taille des textes est très variable dans ce sous-corpus, que ce soit en nombre de

127 Les titres de sections ne sont comptabilisés ni comme des paragraphes, ni comme des phrases.

128 Le nombre de sections ne correspond pas au nombre de titres car nous comptabilisons les sections qui possèdent un texte propre. Or il peut y avoir des sections englobantes sans texte propre, comme la section titrée « *L'importance des facteurs insulaire et frontalier* » dans le tableau VII.2 (p.167). Cette section ne possède pas de texte propre mais englobe deux sous-sections qui, elles, possèdent un texte propre.

129 Le calcul de l'écart-type est expliqué dans la partie VI.4.1.

paragraphes ou de phrases. Par exemple, le texte le plus court est composé de 61 phrases et le texte le plus long de 704 phrases[130]. Par contre, à considérer la partie inférieure du tableau VII.4, l'organisation en dessous du niveau section montre, dans GEOPO, une grande régularité ce qui signifie que les sections et les paragraphes ont souvent une même taille dans GEOPO. ATLAS montre un autre schéma : de la variabilité tant dans la taille des textes que dans la taille des sections ou des paragraphes. La variabilité de la longueur des textes est moindre que celle observée dans GEOPO (le texte 2 présente 1709 phrases contre 3 056 phrases dans le texte 3). Par contre, ATLAS présente des sections de taille très variable pouvant aller jusqu'à 37 paragraphes et des paragraphes relativement petits (3 phrases) montrant des variations importantes vu la petitesse des paragraphes (écart-type = 2,5). D'un autre côté, les deux autres sous-corpus présentent des sections de taille peu variable (écart-type = 3) composées de paragraphes plus grands (en moyenne 5 ou 6 phrases).

Ces différentes variations peuvent faire émerger des correspondances entre certaines stratégies d'organisation et le découpage interne du texte. Par exemple, dans un texte présentant un découpage en paragraphes serrés (beaucoup de petits paragraphes), la marque de changement de paragraphe peut avoir un effet différent de celui dans un texte présentant un découpage en paragraphe plus distendu, de même pour le découpage en sections.

VII.1.1.b) Récurrence nominale et caractère mono/pluri-référentiel des textes

Nous définissons ici les noms récurrents par leur taux de recouvrement. Un nom récurrent est un nom (catégorisé comme tel par un étiqueteur automatique, dans notre cas le TreeTagger) dont le nombre d'occurrences recouvre au minimum 1% de toutes les occurrences nominales d'un texte (ce seuil est le fruit de notre unique volonté). Par exemple, dans un article de géopolitique intitulé "Le Liban et le couple syro-libanais dans le processus de paix", 1 952 occurrences nominales sont repérées. Sur ces 1 952 occurrences, 6 noms récurrents apparaissent : "Liban" (70 occurrences => 3,6% des occurrences nominales), "Syrie" (30 occurrences), "paix" (28 occurrences), "négociation" (26 occurrences), "forces" (22 occurrences) et "armée" (22 occurrences). L'annexe C présente les noms les plus récurrents pour chaque texte des différents sous-corpus. Nous observons trois comportements différents selon les trois sous-corpus.

Nos intuitions quant à l'homogénéité référentielle de PEOPL se trouvent en partie confortées par la récurrence lexicale que ce sous-corpus présente. PEOPL se distingue des autres par la faible diversité des noms récurrents. Généralement, le nom du personnage dont on fait le portrait ressort en premier, puis viennent quelques-uns de ses traits caractéristiques dont la dénomination de son domaine d'activité. Les textes de PEOPL présentent en moyenne 5.5 noms récurrents et un écart-type de 2,7 (les textes comportent entre 1 et 14 noms récurrents)[131].

Pour GEOPO en revanche, il y a soit quelques noms qui se disputent la première place, soit un nom qui sort du lot. Dans les deux cas, un certain nombre de noms récurrents suivent. Ces noms 'secondaires' ne sont pas des caractéristiques spécifiques des noms de tête (à la différence de PEOPL). Les textes de GEOPO affichent en moyenne 9,6 noms récurrents et un écart-type de 3, à peine plus important que les textes de PEOPL (les textes comportent entre 5 et 17 noms récurrents).

Enfin, en ce qui concerne ATLAS, le premier texte (l' « Atlas TransManche ») montre un comportement très différent des deux autres textes. En effet, seuls trois noms recouvrent plus d'1% des noms présents dans ce texte. Ce texte semble donc être particulièrement disparate, comparé aux textes 2 et 3 qui présentent 14 et 8 noms récurrents. Cette disparité est également soulignée par la faible fréquence qu'affichent les trois mots les plus récurrents (moins de

130 Le nombre exact de mots, phrases, paragraphes et sections pour chaque texte est donné dans l'annexe A.

131 Le calcul de l'écart-type est expliqué dans la partie VI.4.1.

1,3%). Il ne s'agit donc pas d'un usage en masse de termes génériques, mais d'un usage hétérogène de noms spécifiques[132]. Cette hétérogénéité des résultats pour ATLAS nous amène à une moyenne de 8,3 avec un important écart-type de 5,5. Cette moyenne est donc très peu pertinente, surtout en rapport à celle de GEOPO.

VII.1.1.c) Distribution de certains SP spatiaux et temporels

Pour avoir un aperçu de l'expression de la localisation spatiale et temporelle dans nos trois sous-corpus, nous avons calculé la fréquence d'apparition de certains syntagmes prépositionnels – SP – pouvant exprimer de façon quasi-certaine la spatialité et la temporalité. En premier lieu, nous avons repéré tous les SP composés d'une des prépositions {*à, dans, en, sur, depuis, au cours de, lors de*} suivie d'un SN : SP = Prep{...} + SN. Ces SP n'expriment pas spécialement le temps ou l'espace : *dans le droit pénal fédéral, dans un bar, sur ces mesures adoptées dans l'urgence, dans l' urgence, depuis la Seconde Guerre Mondiale* (exemples issus du corpus GEOPO).

Parmi ces SP, nous avons identifié plus précisément des SP exprimant de façon quasi-certaine une localisation spatiale ou temporelle. Pour ce faire, nous avons précisé la composition du SN régi par la préposition. Ainsi, les SP spatiaux sont définis comme des SP commençant par une des prépositions {à, dans, en, sur} suivie d'un SN dont la tête correspond soit à un nom propre, soit à l'expression d'un point cardinal soit à un nom de lieu (région, pays, ville, etc.) Les SP temporels sont définis comme des SP commençant par une des prépositions retenues suivie d'un SN dont la tête correspond soit à une date (1900, 1er janvier 1527), soit à une localisation de type début, milieu, fin, soit à l'expression d'une unité temporelle telle que année, siècle...[133] Le tableau VII.5 indique le nombre d'occurrences correspondant à ces patrons dans les différents sous-corpus et y associe leur proportion à apparaître en position initiale.

	Nombre d'occurrences				*Proportion à apparaître en position initiale (%)*			
	Corpus entier	*ATLAS*	*GEOPO*	*PEOPL*	*Corpus entier*	*ATLAS*	*GEOPO*	*PEOPL*
SP spatiaux	5 284	2 565	1 556	1 163	*21,1*	*24*	*17,9*	*19,1*
SP temporels	2 960	1 397	868	695	*29,0*	*27*	*30,3*	*31,5*
SP autres	21 275	6 273	8 480	6 522	*17,5*	*17,5*	*18,1*	*16,7*
Total des SP repérés	29 519	10 235	10 904	8 380	*19,3*	*20,4*	*19,1*	*18,3*

Tableau VII.5: Répartition de quelques SP dans les trois sous-corpus et leur proportion à apparaître en position initiale

Dans ATLAS, un quart des SP composés d'une des prépositions {*à, dans, en, sur, depuis, au cours de, lors de*} sont des SP spatiaux et 14% des SP temporels. GEOPO et PEOPL montrent presque deux fois moins de SP spatiaux ou temporels répondant aux patrons utilisés. Ces données confortent nos intuitions quant à l'expression plus importante dans ATLAS des localisations spatiales et temporelles. Si l'on observe maintenant la propension de ces SP à apparaître en position initiale, la différence d'organisation discursive entre les sous-corpus apparaît encore plus clairement.

29% des SP temporels apparaissent en position initiale contre 21% des SP spatiaux. Ces proportions varient selon le sous-corpus et plus précisément selon que l'on se situe ou non dans ATLAS. Dans ATLAS, près d'un quart des SP spatiaux sont en position initiale contre moins de 20% dans GEOPO ou PEOPL, ce qui va dans le sens de notre typologie *a priori* qui supposait pour les textes d'ATLAS une organisation autour de la dimension spatiale. La dimension temporelle montre un comportement différent. Dans ATLAS, les SP temporels montrent une proportion à apparaître en

132 Nous rappelons qu'à l'inverse des deux autres textes d'ATLAS, l'Atlas Transmanche est un ouvrage en ligne rassemblant plusieurs chapitres d'auteurs différents n'ayant en commun que la thématique, sans ligne éditoriale, ce qui explique certainement l'hétérogénéité lexicale.

133 La liste des SP spatiaux et temporels les plus fréquemment repérés avec ces patrons est donnée en annexe E.

position initiale à peu près égale à celle des SP spatiaux, ce qui suggère que les SP temporels et les SP spatiaux présentent la même capacité à orienter le discours. Par contre, dans GEOPO comme dans PEOPL, les SP temporels apparaissent beaucoup plus en position initiale que les SP spatiaux : près d'un tiers des SP temporels repérés sont en position initiale. ATLAS montre donc moins de SP temporels en position initiale que les deux autres sous-corpus, comme si la fonction d'orientation des SP spatiaux dans ce sous-corpus enlevait du 'travail' aux SP temporels. Nos analyses fouilleront davantage ces pistes de recherche.

VII.1.2. Récapitulatif des caractéristiques du corpus

Notre étude se base donc sur trois sous-corpus choisis pour leur potentialité à montrer des structurations spatiales/temporelles cohabitant avec des structurations référentielles. Les analyses qui leur seront appliquées auront pour but de mettre au jour des différences et des ressemblances dans leur structuration, en observant ce qui s'y passe en position initiale de phrases, de paragraphes et de sections.

En plus de cette caractéristique liée au type de contenu, les trois sous-corpus présentent des caractéristiques de forme et de visée discursive différentes, ce que résume le tableau.

	ATLAS	GEOPO	PEOPL
localisation temporelle	moyenne		
localisation spatiale	forte	?	
contexte référentiel	site pluri-référentiel	– ----- site mono-référentiel ------> +	
textes	longs et de taille régulière	longs et de taille très variable	longs et de taille relativement variable
sections	courtes et de taille variable	longues et de taille régulière	
paragraphes	courts et de taille régulière	plutôt longs et de taille variable	
type procédural dominant	descriptif	argumentatif	narratif
type textuel	textes expositifs		

Tableau VII.6 : Caractéristiques théoriques et observées des sous-corpus

Nous avons présenté jusqu'ici les difficultés d'une approche *data-driven* en EOD. D'un côté, nous défendons la subtilité du marquage de la structuration discursive, d'un autre, nous voulons utiliser des méthodes automatiques pour opérationnaliser cette subtilité. Les parties suivantes précisent toutes les étapes de notre méthodologie.

VII.2. Constitution des observables

Une fois notre corpus construit, l'étape de constitution des observables consiste à annoter les différents éléments composant la position initiale. Comme précisé dans la partie VI.1.1 nous nous situons dans une approche *data-driven* qui se base sur une analyse exhaustive de tout ce qui compose la position initiale. La 'liste' des indices de séquentialité potentiels présentée dans le chapitre V est là pour asseoir notre interprétation des données. Nous ne nous sommes pas spécialement focalisée sur ces formes. Cependant, c'est bien en rapport à ces indices potentiels que nous avons, par exemple, annoté les cas où le Thème topical présente une entité déjà exprimée dans le contexte précédent. Nous illustrons ici la limite labile entre approche *hypothesis-driven* et approche *data-driven* (cf. partie VI.1.1).

L'extraction et la caractérisation de nos observables correspond à une chaîne de traitement composée de quatre modules présentés succinctement dans la partie VII.4. Cette chaîne de traitement est à l'heure actuelle principalement

procédurale et de nombreuses lignes de programme ne correspondent pas réellement à la définition de nos observables, ce qui rend sa présentation relativement complexe. La partie VII.4 explique comment nous avons pensé cette chaîne. L'annexe H présente l'ossature des différents modules et l'annexe J le résultat de deux mini-évaluations effectuées à son sujet.

VII.2.1. Repérage et extraction

Avant de caractériser nos observables (les éléments constitutifs de la position initiale), un programme de **préparation du corpus** (module d'import, cf. VII.4.2) permet de caractériser toutes les phrases[134] présentes dans le corpus. Dans ce repérage, les paragraphes et les phrases sont définis d'une façon purement typodispositionnelle :

Unité textuelle typodispositionnelle	Borne gauche	Borne droite
Section	Titre (inclu)	Titre (exclu)
Paragraphe	Retour à la ligne {¶} (exclu)	Retour à la ligne {¶} (inclus)
Phrase	Ponctuation {¶.?!} (exclue)	Ponctuation {¶.?!} (incluse)

Tableau VII.7 : Caractéristiques typodispositionnelles des sections, paragraphes et phrases

Ainsi délimités, les différentes phrases et parfois paragraphes (les titres sont des paragraphes, typo-dispositionnellement parlant) sont caractérisées par un statut et une localisation, comme le montre le tableau VII.7.

Les trois sous-corpus sont entrés dans le programme sous la forme de trois fichiers au format texte brut. Ce format ne conserve pas la mise en forme matérielle (physique) , exceptée le découpage en paragraphes. Concernant le **repérage des titres**, qui est réalisé automatiquement puis vérifié manuellement, d'autres critères que la mise en forme matérielle sont donc à utiliser. Pour les titres d'articles, un balisage manuel est opéré par l'utilisateur (en l'occurrence moi-même) avant tout traitement. Pour les titres de sections, les critères utilisés sont la présence d'un label (introduction, conclusion, §, chapitre, partie, section) ou d'un système de numérotation, la longueur du paragraphe et l'absence/présence de ponctuation finale (voir l'annexe D). Le repérage automatique des titres de section a été vérifié manuellement, ce qui nous a permis d'ajouter l'information concernant le niveau du titre dans la titraille. Ainsi, un titre de niveau 2 ayant une numérotation est étiqueté NUM_TITRE2_x, où x correspond à la numérotation utilisée. Par exemple, le titre de cette partie « *VII.2.1. Repérage et extraction* » sera étiqueté NUM_TITRE3_VII.2.1. (*i.e.* c'est un titre numéroté – NUM – de niveau trois – TITRE3 – dont la numérotation a la forme *VII.2.1.*)

Cette phase de préparation s'achève par le comptage du nombre de textes, de titres, de sections, de paragraphes et de phrases des différents sous-corpus, ce qui fournit la description quantitative des sous-corpus présentée précédemment. Une fois les sous-corpus préparés, ils sont soumis à un programme conçu pour **le repérage des différents éléments présents en position préverbale**, des constructions spéciales et des éléments ainsi mis en focus. Ce module de segmentation (module 2, cf. VII.4.2) parcourt toutes les phrases des textes (titres compris), et extrait successivement pour chacune d'elles :

Dénomination	Description
Ponct	Ponctuation finale de la phrase précédente { ? . ! : … }.
Puce	Signe typographique présent en ouverture de phrase {* - •}
Pred	Prédicat de la phrase : le programme cherche le verbe principal puis divise la phrase en une partie prédicat (noté ici) et une partie position initiale (qui va faire l'objet des caractérisations suivantes)
INIT : éléments en position initiale précédant le sujet grammatical (et suivant le connecteur)	

134 Le découpage en phrases est réalisé par un module intégré à Syntex.

Dénomination		Description
	Connect	Connecteur « pur »[135]
	INIT1	Premier élément de la position initiale précédant le sujet (hors connecteur). Ce premier élément peut être accompagné d'un second élément explicatif du premier (ex. VII.3)
	t_INIT1	Tête de l'INIT1*
	e_INIT1	Suite d'étiquettes morphosyntaxiques de l'INIT1
	INIT2	Tous les éléments situés entre INIT1 et le sujet grammatical
	t_INIT2	Tête du 2e élément détaché*
	e_INIT2	Suite d'étiquettes morphosyntaxiques de l'INIT2
ThTop : thème topical ~ sujet grammatical d'une construction non spéciale		
	ThTop	élément répondant au patron d'un sujet grammatical, sauf si une construction spéciale a été repérée, et alors cet élément est vide
	t_ThTop	Tête du ThTop*
	e_ThTop	Suite d'étiquettes morphosyntaxiques du ThTop
ThSpe = Thème spécifique ou phrase à construction spéciale		
	TypTs	Type de la construction à Thème spécifique repérée : (pseudo)clivée, présentationnelle, impersonelle, commentaire, sujet inversé, topicalisation ou autre.
	Focus	Éléments mis en focus par les constructions spéciales
	t_Focus	Tête de l'élément focus*
	e_Focus	Suite d'étiquettes morphosyntaxiques de l'élément focus
* le repérage des têtes se base sur l'analyse de Syntex et ne prend en compte que les têtes nominales ou pronominales. Pour les SN, la tête correspond au nom recteur du déterminant. Pour les SP, la tête correspond au nom régi par la préposition ; pour toutes les propositions infinitives, il s'agit du premier nom rencontré. Pour les syntagmes adjectivaux, même s'il ne s'agit pas de la tête syntaxiquement parlant, le premier nom rencontré est mémorisé pour calculer de possibles reprises. Enfin, en cas de coordination, toutes les têtes nominales sont enregistrées et présentées : tête1 & tête2 & etc.		

Tableau VII.8 : Annotations des éléments repérés pour chaque phrase

Pour illustrer le repérage des différents éléments, nous référons aux trois exemples suivants :

(VII.1) *Ainsi, aux États-Unis, parmi les personnes interrogées et souhaitant que le processus de mondialisation soit arrêté ou inversé, 49% affirment que le gouvernement américain n'a pas les capacités de le faire.*

(VII.2) *Dans l'immédiat, ni aux États-Unis, ni au Brésil, on ne semble craindre la propagation d'une crise considérée comme très spécifique.*

(VII.3) *Quant à la drôlerie, qui est parfois dérision, elle est aussi éloignée que possible du "grotesque triste" d'un Flaubert.*

qui sont alors annotés de la façon suivante :

(VII.1') **<Connect>***Ainsi,***</Connect>**
<INIT1>*aux***<t_INIT1>***États-Unis***</t_INIT1>**,**</INIT1>**
<INIT2>*parmi les* **<t_INIT2>***personnes***</t_INIT2>** *interrogées et souhaitant que le processus de mondialisation soit arrêté ou inversé,***</INIT2>**
<ThTop>*49***<t_ThTop>**% **</t_ThTop></ThTop>**
<pred>*affirment que le gouvernement américain n'a pas les capacités de le faire.***</pred>**

(VII.2') **<INIT1>***Dans l'***<t_INIT1>***immédiat***<t_INIT1>**,**</INIT1>**
<INIT2>*ni aux* **<t_INIT2>***États-Unis***</t_INIT2>**, *ni au* **<t_INIT2>***Brésil***</t_INIT2>**,**</INIT2>**
<ThSpe>*on* **<pred>***ne semble craindre la propagation d'une crise considérée comme très spécifique.***</pred></ThSpe>**

(VII.3') **<INIT1>***Quant à la* **<t_INIT1>***drôlerie***<t_INIT1>**, *qui est parfois dérision,***</INIT1>**
<ThTop>*elle***</ThTop>**
<pred>*est aussi éloignée que possible du "grotesque triste" d'un Flaubert.***</pred>**

135 Par « pur », nous faisons référence à la définition restreinte de cette classe de mots, telle que présentée précédemment dans la partie V.3.4. Une liste des connecteurs repérés est donnée en annexe I.

Le **repérage des connecteurs « purs » – Connect –** consiste à repérer les mots simples qui précèdent ce que nous considérons comme les éléments détachés en initiale. Pour ce faire, une liste fermée de patrons a été constituée (voir V.3.4 et annexe H). Cette liste englobe inévitablement des éléments annotés Connect mais n'étant pas des connecteurs (*i.e.* du bruit) et omet des éléments qu'il aurait fallu annoter Connect (*i.e.* du silence). Le choix a été ici de fermer au maximum la classe des connecteurs et donc d'avoir plus de silence que de bruit, *i.e.* plus de connecteurs non annotés Connect que d'éléments 'non connecteurs' annotés Connect. Lorsqu'un élément pouvant être annoté Connect n'est pas annoté comme tel, il se retrouve alors annoté INIT1. Ce choix est motivé en grande partie par le flou qui existe dans la définition des connecteurs (comme nous l'avons vu en V.3.4). En effet, la limite entre connecteurs et introducteurs de cadres est parfois délicate à définir, surtout si l'on considère les cadres organisationnels (*troisièmement*, *d'une part*) qui forment eux-même une classe aux limites floues. La décision prise est de préférer une caractérisation en cadre (et donc INIT1) à une caractérisation en connecteur, afin que tout ce qui n'est pas assurément un connecteur puisse être plus finement analysé ultérieurement.

Le critère formel retenu pour **délimiter les éléments détachés en initiale – INIT–** du reste de la phrase qu'elle soit avec Thème topical ou Thème spécifique est principalement lié à la ponctuation. Les ponctuations retenues comme signes pertinents de détachement sont : {, : -}, ce qui signifie que le programme s'appuie sur ces signes pour distinguer les éléments détachés des éléments intégrés. Un élément détaché en initiale consiste alors en une suite de caractères entourés de marques de détachement tel que le montre les deux schémas suivants :

début de phrase (Connecteur(,)) INIT1 { , - : } (INIT2 { , - : })

Cependant, notre programme repère également les éléments détachés où le détachement n'est pas marqué par une virgule. Deux cas sont repérés :

(1) quand il y a sujet inversé comme en (VII.4),

(VII.4) **<INIT1>***Dans la zone transmanche* **</INIT1>***existent ainsi la Communauté urbaine de Lille, de Dunkerque, de Cherbourg, les districts de Rennes, de Caen .*

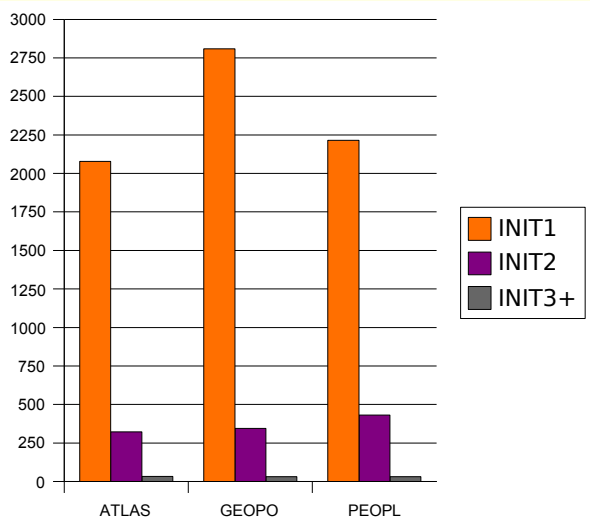
(2) quand il y a non ambiguïté comme en (VII.5) car il y a deux blocs Syntex dont le premier de forme subordonnée, adverbiale, prépositionnelle, adjectivale, etc. et le second de forme nominale ou pronominale.

(VII.5) *Cependant,* **<INIT1>***si les volumes sont croissants* **</INIT1>***les disparités ne vont pas forcément en se renforçant en Manche-Ouest.*

Le choix de **distinguer plusieurs éléments détachés en initiale** (**INIT1 et INIT2**) répond à la possibilité d'avoir des thèmes multiples, qui présentent plusieurs unités indépendantes placées en position initiale. Comme il a été montré en partie IV.4, ces unités peuvent correspondre à deux (ou plus) introducteurs de cadres ; on a alors des cadres multiples qu'il faut prendre en compte comme tels et non comme un seul cadre. Cette distinction peut valoir pour des cadres de types différents (un cadre organisationnel plus un cadre véricconditionnel : *Par contre*, *Sur le plan musical* ou *En France*, *au contraire*) ou des cadres multiples de même type (*En 1986, dans une soixantaine de départements*, ou *En primaire, dans les écoles publiques,* ou *En trois années, de 1992 à 1995*,).

Parallèlement, le choix de ne classer les éléments détachés qu'en INIT1 et INIT2 (et non INIT3, INIT4, etc.) se justifie de trois façons : premièrement, d'un point de vue théorique, la conception de l'initiale comme répondant au principe de l'information cruciale en premier (CIF, Enkvist 1989, voir partie IV.1) donne un poids décroissant aux éléments plus on s'éloigne du début de l'unité textuelle en question. Cette prégnance du premier élément détaché est de la plus grande importance lorsqu'on se place dans l'organisation du texte selon des continuités texto-stratégiques (III.3.4). D'un autre point de vue, si l'on tient compte du principe d'ordonnancement des mots selon la distinction

donné/nouveau, le premier élément détaché constitue l'élément par lequel le lien avec le contexte précédent est préférentiellement établi ; ce qui rejoint alors notre prise en compte du mode de segmentation en chaînes. Les résultats d'extraction montrent à ce sujet que 13% des INIT1 (909/7 021) comportent des indices de relation anaphorique[136] contre seulement 4,5% des INIT2 (49/1 085).



Graphique VII.1 : Les thèmes multiples en position initiale détachée selon les sous-corpus (en nombre de phrase)

Enfin, en s'appuyant sur les réalités langagières présentes dans le corpus d'étude, on voit clairement que le fait d'avoir plus de deux éléments détachés en position initiale (INIT3+) reste très marginal (94/23 217 => 0,4% des phrases étudiées), ce que montre le graphique VII.1. Cette observation est renforcée par les données présentes dans le corpus intitulé « 1 001 circonstants »[137], où parmi 1 118 circonstants relevés, 87,2% ne présentent qu'un seul élément, contre 10,2% à deux éléments, 1,8% à trois et 0,6% à quatre et plus. Les observations d'Hasselgård (1996, 2004b) vont dans le même sens, puisqu'elle remarque que sur les 1 665 séquences[138] d'adverbiaux spatiaux/temporels repérés, 82% comportent deux éléments, 9,5% trois éléments et 8,5% quatre ou plus.

Un autre point concernant la distinction entre INIT1 (le premier élément détaché) et INIT2 (les éléments détachés suivants) concerne la multifonctionnalité des signes de ponctuation utilisés comme signes de détachement. Comme nous l'avons dit, INIT1 et INIT2 sont deux éléments indépendants. Or, la virgule et encore plus le tiret peuvent marquer des éléments en incise qui précisent un aspect de l'information exprimée dans le premier élément. Il faut alors éviter de catégoriser cette incise en INIT2. Cependant, peu de marques formelles peuvent nous indiquer que l'on a affaire à une incise et non à un élément indépendant. Certaines expressions comme *c'est-à-dire* ou *tel(s-le(s)) que* sont fiables pour rattacher le deuxième élément détaché au précédent, de même lorsque ce deuxième élément correspond à une proposition relative ou à un syntagme adjectival (sous condition dans ce cas que le premier élément ne soit pas lui aussi adjectival). Mais bien d'autres configurations existent et nécessitent alors, pour être reconnues, une analyse

136 Les différents indices de relation anaphorique sont donnés dans la partie suivante.

137 Ce corpus a été constitué en 1990-1991 par le laboratoire de linguistique ELSAP (Étude Linguistique de la Signification à travers l'Ambiguïté et la Paraphrase, URA 1234 du CNRS). Il a été établi à partir de 12 articles extraits d'un numéro du journal *Le Monde* qui ont été systématiquement dépouillés afin d'en extraire tous les circonstants présents. Ces circonstants ont ensuite été annotés semi-automatiquement. Les annotations effectuées sont : identifiant de la phrase d'accueil, position dans la phrase et par rapport aux autres circonstants dans le cas de circonstants multiples, catégorie syntaxique, ponctuation, intégration syntaxique (complément facultatif ou essentiel).

138 Dans le travail d'Hasselgård (1996), les séquences correspondent aux éléments détachés composés d'au moins deux éléments.

manuelle. On peut cependant noter que ces constructions sont généralement assez complexes comme l'illustre l'exemple VII.6. Une étude syntaxique approfondie pourrait nous aider, mais là encore ce n'est pas le but de ce travail.

(VII.6) ***Si l'on peut, en simplifiant grossièrement - mais Hegel le fait souvent -, assigner ces trois thèses respectivement à Spinoza, Descartes et Leibniz,*** *on percevra la modernité de la critique hégélienne en remarquant que la conjugaison des deux premières commande un humanisme existentialiste de type sartrien, et la troisième un structuralisme idéologique.[PEOPL_21]*

Le dernier point encore dépendant de décisions d'ordre pratique est la distinction faite entre **phrases avec Thème topical ou Thème spécifique**. Tout comme pour la question des connecteurs, le programme se base sur une liste plus ou moins fermée des patrons caractéristiques de constructions spéciales (la liste de ces patrons est donnée dans l'annexe H), ce qui laisse là encore une certaine place pour le bruit et le silence.

Dans les cas ambigus (pour le programme), nous avons préféré avoir plus de bruit en Thème topical qu'en Thème spécifique. Cela revient donc à accepter plus de silence en Thème spécifique. Ce choix répond expressément au programme élaboré qui considère la construction avec Thème topical comme la construction par défaut. Ce n'est donc que si un patron de construction spéciale[139] est repéré que la phrase sera considérée comme construite avec un Thème spécifique. Si aucun patron n'est repéré, par défaut, la phrase comporte un Thème topical. Avec la liste de patrons établie, le taux de précision pour le repérage des constructions spéciales égale 98,8 et celui de rappel 99,4 (voir l'annexe J pour les détails de la validation), ce qui est largement raisonnable pour avoir des caractérisations fiables pour des analyses quantitatives. Les erreurs de traitement sont essentiellement du bruit au niveau des Thèmes topicaux de forme *il* qui constituent de forts indices de continuation, contrairement aux constructions impersonnelles.

Une fois ce repérage effectué, un autre module (module 3, cf. VII.4.2) aide à la caractérisation des éléments extraits. Lors de cette caractérisation, une vérification du repérage s'effectue également, accompagnée s'il le faut de quelques modifications.

VII.2.2. Caractérisation des éléments annotés : INIT, ThTop et ThSpe

L'objectif principal de cette caractérisation est de pouvoir décrire de façon quantitative la composition de la position initiale et ses variations selon différents facteurs discursifs exposés dans le chapitre suivant. Les éléments extraits du module précédent subissent ainsi un traitement basé sur leur forme et parfois leur contenu, permettant de retrouver les grandes catégories définies dans la partie V.1.4.

À ces caractérisations s'ajoutent diverses informations concernant les occurrences repérées :

- Si un élément du titre de section en cours est repris par un élément de la phrase en cours de traitement, l'élément est annoté comme présentant une reprise du titre. Si la phrase en cours de traitement est un titre de section, le nombre de reprises qu'il présente dans 'sa section' est enregistré.
- Si un ou plusieurs mots fréquents sont repérés dans la phrase en cours de traitement, les éléments dans lesquels ils apparaissent ((t)INIT, (t)ThTop ou pred) ainsi que leur nombre d'occurrence sont enregistrés.

VII.2.2.a) Caractérisation des éléments détachés en initiale – INIT

La caractérisation des INIT nécessite des informations complémentaires à l'étiquetage morpho-syntaxique donné par Syntex. Nous avons dû construire des listes d'expressions régulières permettant de définir automatiquement les différents types d'INIT observés. L'annotation résultante présente trois catégories : la catégorie morpho-syntaxique, la

139 Le programme de repérage des constructions spéciales est résumé en annexe H.

fonction sémantico-discursive et le rôle sémantique. Les INIT peuvent remplir la fonction d'apposition, d'adverbial circonstanciel, d'adverbial modalisateur, d'adverbial textuel (dont les marqueurs d'intégration linéaire – MIL). Ils peuvent également correspondre aux éléments détachés des constructions spéciales (topicalisations, pseudo-clivées, sujet inversés). Le rôle sémantique n'est assigné qu'aux INIT fonctionnant au niveau idéationnel (voir V.4). Ces rôles sont déterminés relativement à la dimension sémantique exprimée (le temps, le lieu, le domaine, etc.).

Catégorie morpho-syntaxique	***Annotation***
syntagme prépositionnel	SP
syntagme nominal	SN
proposition infinitive	INF
proposition finie	FIN
proposition avec participe passé	PPA
gérondif ou proposition avec participe présent	PPR
syntagme adjectival	ADJ
proposition relative	REL
syntagme adverbial	ADV
forme indéfinie	000
Fonction sémantico-discursive	***Annotation***
circonstant	CIRC
marqueur d'intégration linéaire	MIL#
adverbial modalisateur	MODA
adverbial textuel	TEXT
élément détaché d'une dislocation	TOP
séquence thématique d'une construction pseudo-clivée	ST
argument d'une construction à sujet inversé	ARGU
apposition	APPO
fonction indéfinie	0000
Rôle sémantique	***Annotation***
temps	tps
lieu	spa
domaine de connaissance ou d'activité	not*
thématique	
source d'énonciation particulière	
autres circonstants ou rôle sémantique indéfini	0
* « not » est l'abréviation relative au terme « cadre notionnel » (cf. V.4.1)	

Tableau VII.9 : Caractérisation des INIT

Les étiquettes utilisées compilent les trois catégories d'information sous le format XXX_XXXX_xxx (par exemple, SP__CIRC_tps ou INF_TOP__0).

Certaines annotations sont laissées incomplètes (caractérisations indéfinies : 0), ce qui correspond soit à un choix justifié soit à une impossibilité de caractériser. Une annotation incomplète se rencontre lorsque l'INIT présente une forme qui ne correspond pas à une des expressions régulières définies. La catégorie des circonstants est certainement la plus difficile à définir formellement. Une information peut également ne pas être annotée car elle n'est ni pertinente ni appropriée, ce qui est le cas de la catégorisation par rôle sémantique pour toutes les fonctions sémantico-discursives autres que CIRC (*En premier lieu* = SP_TEXT_0, *Malheureusement* = ADV_MODA_0). Les appositions et les arguments de constructions inversées peuvent de façon raisonnée être associés à une telle catégorisation :

(VII.7) **<INIT type='PPA_APPO_tps'>** *Rendu public le 18 octobre,***</INIT>**[140] *l'accord a été signé le 5 octobre pour une durée de un mois. [GEOPO_2]*

(VII.8) **<INIT type='SP__ARGU_tps'>***À la fin du XVIIIe siècle***</INIT>** *apparaît une "manière", d'audience quasi nationale, le verbunkos, issu des danses de recrutement amalgamées à des éléments orientaux et viennois, manière qui eut un rayonnement international assez important à travers le style hongrois des compositeurs du XIXe siècle (Franz Liszt, Ferenc Erkel, Stephen Heller, Mihály Mosonyi). [PEOPL_15]*

Mais il ne s'agit pas ici de décrire tout ce qui peut exprimer du temps, de l'espace, etc. Pour les appositions et les arguments inversés, la catégorie sémantique peut donc être annotée, mais cette annotation n'est pas utilisée dans nos analyses. Les éléments répondant pertinemment à une catégorisation sémantique restent pour notre étude les circonstants qui seuls semblent pouvoir poser un cadre de discours.

La dernière étape d'annotation des INIT consiste à indiquer si l'INIT présente une relation anaphorique. Différents indices peuvent nous informer sur la présence d'une relation anaphorique en INIT : la présence du pronom *il* (VII.9), d'un SN démonstratif (VII.10), d'un SN de forme : Det + *même* + NomXXDate (VII.11) ou encore des locutions adverbiales *plus tard* ou *avant*, *auparavant* (VII.12).

(VII.9) *En effet,* **<INIT type="SP__CIRC_tps_ana">***après un grand nombre d'années passées à la tête de commissions où ils ont réussi à s'imposer comme de véritables piliers du Congrès***</INIT>***, les présidents des commissions les plus importantes négligent volontiers le jeu des partis. [GEOPO_20]*

(VII.10) **<INIT type="FIN_CIRC_ssi_ana">***Si cette science technocratique existait***</INIT>***, elle permettrait de départager dans tous les cas savoir et non-savoir, hommes compétents et charlatans : elle contrôlerait le fonctionnement et le bon usage de l' ensemble des sciences. [PEOPL_19]*

(VII.11) **<INIT type="SN__CIRC_tps_ana">***La même année***</INIT>,** *il compose la Suite n°1 pour orchestre et les trois premiers mouvements de la Suite n°2, qu'il ne terminera que deux ans plus tard. [PEOPL_15]*

(VII.12) **<INIT type="SN__CIRC_tps_ana">***Vingt-cinq années plus tard***</INIT>,** *le rapport Droite/Gauche s'établit approximativement à 50/50. [ATLAS_3]*

Enfin, il faut préciser que seuls les deux premiers éléments détachés sont caractérisés : si le deuxième élément comprend plusieurs éléments indépendants, l'annotation ne s'applique qu'au premier de ces éléments, les raisons de ce choix ont déjà été exposées dans la partie précédente.

VII.2.2.b) Caractérisation des thèmes topicaux – ThTop

L'opération de caractérisation est relativement facile pour ce qui est des Thèmes topicaux (qui correspondent aux sujets grammaticaux des constructions canoniques), puisqu'elle ne requiert pas d'autres informations que celles fournies par Syntex, étant uniquement de nature morpho-syntaxique. En regroupant certaines formes peu significatives par rapport à nos sous-corpus et notre objectif de recherche, nous obtenons huit catégories telles que les présente le tableau VII.10. Ce tableau montre également les éléments pour lesquels le paramètre « reprise » (_R) est pertinent, paramètre qui correspond au fait que la tête du Thème topical a déjà été mentionnée dans la section en cours. Dans le chapitre suivant présentant les résultats de l'analyse, les diverses formes seront désignées par leur nom de catégorie, suivie le cas échéant de la mention « _R ».

140 Cette balise indique la fin de l'élément détaché.

Catégories morpho-syntaxique	Annotation
pronom personnel de 3e personne : *il*	PRO3
pronom démonstratif	PROdem
syntagme nominal possessif	SNposs
syntagme nominal démonstratif	SNdem(_R)
syntagme nominal défini	SNdef(_R)
syntagme nominal indéfini (avec déterminant indéfini, quantifieur ou déterminant numéral)	SNindef(_R)
nom propre	NP(_R)
pronoms indéfinis, SN sans déterminant (hors NP), constructions infinitives et éléments non caractérisés	autres(_R)

Tableau VII.10 : Caractérisation des ThTop

VII.2.2.c) Caractérisation des constructions à Thème spécifique – ThSpe

Cinq constructions à Thème spécifique sont distinguées dans cette étude. Leur caractérisation se base sur des patrons plus ou moins complexes[141].

Brève Description	Annotation
Constructions clivées ou **pseudoclivées** construites selon le schéma : *C'est ... qu_ ...* ou, pour les pseudoclivées : *Ce qu_ .... c'est ....*	Cliv
Constructions qui constituent des **commentaires du locuteur** sur les propos exprimés. Ces phrases suivent le schéma : *On/Nous Vmodalisateur qu_...*	On...
Constructions impersonnelles construites autour des schémas : Il Vmodalisateur qu_... ou Il est Adjectif_modalisateur qu_...	ILimp
Sujets inversés caractérisés par le fait que le verbe précède le sujet. Un argument inversé peut alors apparaître en INIT (voir exemple (VII.7))	SujInv
Constructions présentationnelles, encore appelées constructions existentielles, qui suivent le schéma : *C'est....* ou *Il s'agit d_ ...* ou *Il y* [forme conjuguée du verbe *avoir*]	Present
Dislocations à gauche : *le x, c'est ...*	Disloc
Autres constructions spéciales	autres

Tableau VII.11: Caractérisation des ThSpe

Pour la plupart des constructions à Thème spécifique, le focus a été repéré*, i.e.* le syntagme mis en emphase par la construction.

(VII.13) **<PHR nat="Cliv_SN">** *Ce sont* **<FOCUS morph="SNdef_MP" redeno="0">** *les acheteurs individuels (traders, raffineurs)* **</FOCUS>** *qui sont en concurrence pour s' approvisionner, et non les États ou les économies nationales.***</PHR>** [*GEOPO_14]*

Nous n'étudierons pas spécialement les focus ainsi repérés. Leur caractérisation nous est principalement utile pour distinguer les constructions spéciales mettant le focus sur un SN (VII.13) de celles mettant le focus sur un SP (VII.14).

(VII.14) **<PHR nat="Cliv_SP">** *C'est* **<FOCUS morph="SP" redeno="0">** *au cours des années 70* **</FOCUS>, [...],** *que le développement des lignes maritimes traversant la Manche a été le plus fort.***</PHR>** [*ATLAS_1]*

Cette caractérisation peut être pertinente au niveau des clivées, afin de distinguer les clivées qui mettent le focus sur un adverbial de forme SP. En effet, ainsi que le remarque Hasselgård (2004a) et Bilhaut (2006), les adverbiaux focalisés afficheraient un rôle bien similaire à celui des adverbiaux introducteurs de cadres, notamment pour les

141 La totalité des patrons utilisés est donnée dans l'annexe H.

adverbiaux temporels ou spatiaux. L'exemple suivant montre deux adverbiaux temporels en focus (en gras) qui introduisent un même cadre temporel.

(VII.15) *Le développement des lignes maritimes est au tout premier rang de ces facteurs de rapprochement. Avant-guerre il existait deux traversées quotidiennes entre Douvres et Calais, aujourd'hui on atteint les quatre-vingt liaisons par jour (...). C'est **au cours des années 70**, la décennie où les grands traits de la matrice historique et spatiale dans laquelle nous vivons se sont affirmés, que le développement des lignes maritimes traversant la Manche a été le plus fort. Chaque année la croissance de trafic sur les trajets courts était de l'ordre de 10 %. La Brittany Ferries, dont l'histoire même a ses racines dans des liens antérieurs entre la Bretagne et la Cornouailles (la Brittany-Ferries est née en 1972 sous le nom de "Bretagne-Angleterre-Irlande" (BAI). Elle a pris son nom actuel en 1974 en élargissant son activité au transport de passagers. Basée à Roscoff, elle était initialement centrée sur l'exportation de marchandises agricoles. Elle s'est développée sur les ports de Bretagne et de Normandie, souvent au travers de la création de SEM avec les collectivités territoriales), ouvre une ligne de ferry régulière Roscoff - Plymouth en 1973, puis Saint-Malo - Portsmouth en 1976.► C'est **dans la même décennie** que les lignes se développent à partir de Cherbourg et Le Havre en direction de Southampton, Poole, puis Portsmouth, et en 1986 qu'ouvre la ligne Caen - Portsmouth. Ces développements sur la Manche Ouest font alors de toute la façade des régions bordières une zone de contact régulier, ce qui était jusqu'alors l'apanage du détroit.► Le trafic du détroit continue de croître en volume, [...]* [ATLAS_1]

VII.2.2.d) Degré d'accessibilité et « descriptions longues » vs. « courtes »

Les phrases sont également caractérisées selon l'échelle d'accessibilité d'Ariel (1990) présentée en V.4.3. Ainsi, nous avons sept degrés d'accessibilité – DegAccess – associés à différents regroupements de formes en sujet grammatical :

DegAccess_n	*Formes correspondantes*
0	Descriptions indéfinies, sujet de forme autres (e.g. SN sans déterminants) ou de forme indéfinie et toutes les construction à Thème spécifique
1	Nom propre nouveau
2	Description définie complète
3	Description définie réduite et/ou avec reprise
4	Nom propre repris
5	Description démonstrative avec modifieur
6	Description démonstrative réduite et/ou avec reprise
7	Pronoms et SN possessifs

Tableau VII.12 : Formes attribuées aux différents degrés d'accessibilité

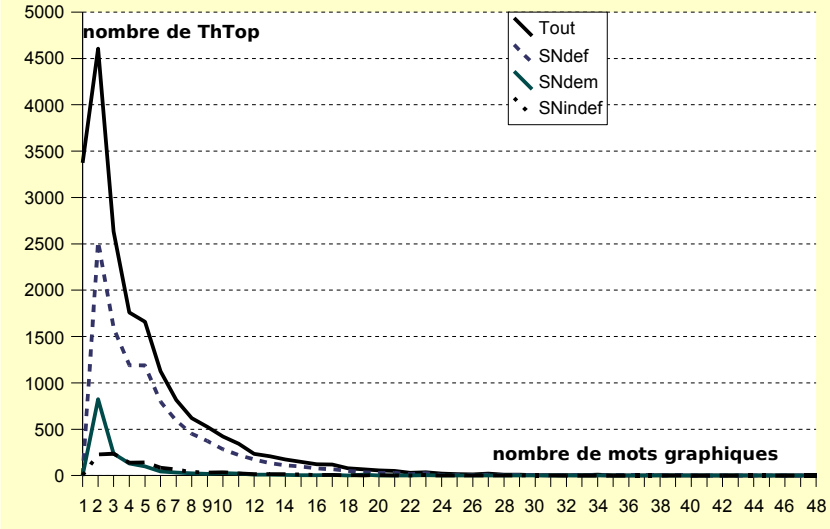
L'avantage de représenter les phrases par leur degré d'accessibilité en position initiale est de permettre un accès selon un point de vue plus cognitif. Ainsi, notre exploration de la position initiale gagnera, dans certains cas, en lisibilité. Si l'on observe les formes correspondant aux différents degrés d'accessibilité, nous remarquons que la seule information qui n'a pas encore été annotée concerne la distinction entre descriptions courtes et descriptions complètes.

Pour permettre la distinction automatique entre description complète et description réduite, nous avons opté pour une méthode des plus simpliste qui consiste à compter le nombre de mots (définis graphiquement comme une suite de caractères délimitée par des espaces typographiques) présents dans les descriptions repérées[142]. Cette méthode ne repère pas les descriptions complètes ou réduites au sens linguistique du terme, c'est pourquoi nous préférons utiliser les termes « descriptions longues » et « **descriptions courtes ».**

Pour notre étude, les SN présentant plus de quatre mots sont considérés comme des descriptions longues et les SN présentant au maximum quatre mots sont considérés comme des descriptions courtes. Cette méthode s'applique uniquement aux SN définis et démonstratifs. Pour soutenir cette caractérisation, nous avons observé la composition en

142 Des études françaises proposent des analyses plus fines permettant l'élaboration d'un outil de repérage automatique des descriptions réduites (Jacques 2003). Malheureusement, aucun de ces travaux n'a abouti à la création effective d'un tel outil.

nombre d'espaces typographiques des SNdef et SNdem présents dans notre corpus. Le graphique VII.2 indique le nombre de mots dans tous les Thèmes topicaux de notre corpus[143]. Nous y avons distingué les SN définis, les SN démonstratifs et les SN indéfinis.



Graphique VII.2: Nombre de mots graphiques dans les Thèmes Topicaux

Comme on le voit, Les SN définis – qui sont les types de Thème topical majoritaires – suivent la tendance générale : les expressions présentant deux mots sont presque deux fois plus nombreuses que celles comportant trois mots (2 525 -> 1 586 occ. pour les SNdef et 4 604 -> 2 634 pour tous les types de Thème topical). Les SN démonstratifs exagèrent cette tendance puisque les SNdem comprenant deux mots (826 occ.) représentent plus de trois fois ceux en comprenant trois (237 occ.) Concernant les SN indéfinis qui sont généralement associés à l'arrivée d'un nouveau référent et donc peuvent difficilement être réduits, le schéma diffère du mouvement général. La dégringolade mesurée lorsque l'on passe à cinq mots disparaît. Les SN composés de deux mots (228 occ.) sont aussi nombreux que ceux composés de trois mots (236 occ.). Les SN indéfinis avec deux mots ne représentent même pas le double de ceux en comportant cinq. Dans le sous-corpus ATLAS, nous observons une quasi égalité entre tous les SN indéfinis comportant jusqu'à 5 mots graphiques (59 -> 56 -> 51 -> 58).

Dans tous les cas, nous observons un nombre relativement faible d'occurrences (moins de 10% des SN) dès lors que l'on comptabilise au moins quatre mots.

Nous voyons d'après cette figure que les SNdef montrent une composition en termes de nombre de mots graphiques différente de celle des SNdem. Pour les SNdef, il faut considérer des SN ayant cinq mots pour dépasser les 50%, alors que pour les **SNdem**, en ne prenant que les SN ayant au maximum trois mots, nous atteignons déjà les 60% de SNdem. Ces résultats peuvent expliquer l'hypothèse selon laquelle le SNdem sert généralement à une reprise coréférentielle (directe ou indirecte) ou à une encapsulation (*i.e.* une anaphore résumante comme l'expression *« ces trois cas »* ci-après). Dans ces trois cas, il s'agit de référer à une entité en cours d'activation, ce qui se réalise principalement par des descriptions réduites.

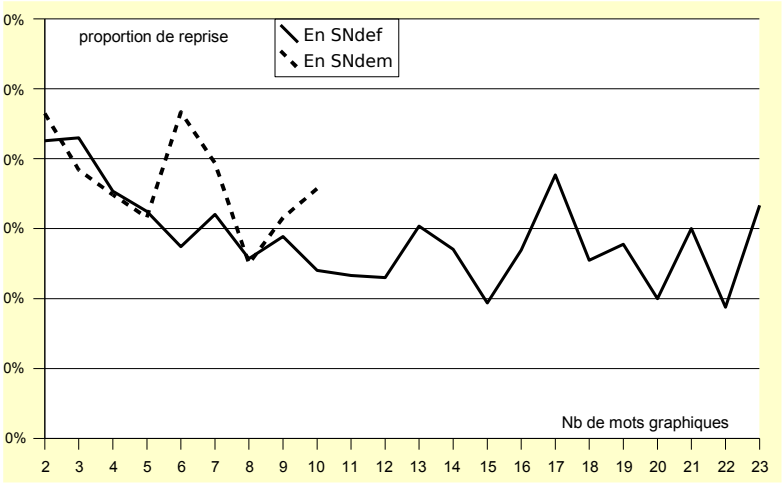
Du côté des SNdef, la situation est plus complexe. Un **SNdef** peut réaliser une reprise anaphorique, tout comme il peut servir à introduire un référent nouveau ou semi-actif (*i.e.* identifiable parmi un ensemble déjà actif). Dans ce

143 Les Thèmes topicaux ne comportant qu'un seul mot correspondent principalement aux pronoms et aux noms propres. (3 228 occurrences)

dernier cas, une description longue est nécessaire afin de préciser l'élément et sa relation avec l'ensemble actif. L'exemple VII.16 (infra.) montre plusieurs descriptions définies longues.

Pour asseoir davantage notre décision, nous avons mesuré la corrélation entre le nombre de mots contenus dans les descriptions et le fait que la tête du SN présente ou non une reprise (annoté _R dans la définition des caractérisation des Thèmes topicaux). Nous nous attendons alors à observer une certaine corrélation entre la propension à constituer une reprise et le nombre de mots (une description courte).

Le graphique VII.3 indique que le comportement des SNdem et des SNdef est très différent et relativement inattendu. Alors que nous nous attendions à des courbes décroissantes de façon relativement régulière pour les deux types de SN, nous avons des courbes en dents de scie, présentant des pics à des endroits différents. La courbe des SNdem s'arrête à 10 mots, seuil à partir duquel la faiblesse des occurrences rend les corrélations entre proportions de reprises et nombre de mots impossibles à généraliser (voir le graphique VII.2). Par exemple, le fait de ne trouver que 5 SNdem de 16 mots dont 4 présentant une reprise lexicale, résulte sur une corrélation très forte entre reprise et descriptions comportant 16 mots, corrélation dont on peut mettre en doute la représentativité. Pour les SNdef, nous nous sommes arrêtée au nombre de 23 mots, seuil à partir duquel nous recensons moins de 10 occurrences.



Graphique VII.3 : Corrélation entre longueur des descriptions en Thème topical et reprises en tête de description

Jusqu'à un nombre de 12 mots, la corrélation entre proportion de reprise en SNdef et nombre de mots se rapproche de notre idée préconçue d'une courbe décroissante. Les effets dents de scie y sont beaucoup plus faibles que pour les SNdem. Au delà des 12 mots, les dents de scies augmentent et on observe une quasi égalité en proportion de reprise entre des **descriptions comportant 4 ou 17 mots. En fait, moins on a d'occurrences, plus les résultats sont déroutants.**

La courbe des **SNdem** ne montre pas de corrélation entre reprise et nombre de mots. Que les descriptions démonstratives soient composés de deux ou six mots, la probabilité d'avoir une reprise en tête de syntagme reste égale (autour de 50%). Les exemples (VII.16) et (VII.17) présentent quelques exemples d'expressions référentielles calculées 'courtes' ou 'longues'. Nous sommes consciente que cette classification est imparfaite. Cependant, elle nous permet de dégager des régularités acceptables concernant les descriptions 'courtes' qui nous intéressent particulièrement ici, pour leur potentialité à marquer une continuité référentielle.

(VII.16) ***L'Etat fédéral [SNdef court]****, lui, se concentre dans des domaines qui ont toujours été les siens : l'amélioration des services de sécurité, la protection des frontières et la lutte contre toute puissance extérieure qui soutiendrait d'une façon*

ou d'une autre les organisations terroristes. ***La récente réorganisation annoncée du FBI [SNdef long]*** *va dans ce sens. Dorénavant,* ***le Bureau [SNdef court]*** *ne devrait plus se pencher autant sur les attaques à main armée dans n'importe quelle banque du pays, mais concentrer ses énergies sur une lutte nationale contre le terrorisme.* ***Les autorités locales [SNdef court]*** *sont donc placées en première ligne, et ressentent d'ailleurs le coût des nouvelles attentes à leur endroit. En effet, une des conséquences des attentats du 11 septembre a été une forte réduction des services offerts par les Etats, les villes et les comtés, à tel point que certains Etats envisageraient maintenant de supprimer toute aide aux villes, renforçant par là-même le coût de la sécurité pour les gouvernements non-étatiques.* ***Le transfert des ressources fédérées vers le poste de la protection anti-terroriste [SNdef long]*** *a donc été massif. Un article du Los Angeles Time soulignait ainsi que "Les autorités locales n'avaient pas eu une responsabilité de cette envergure en matière de défense depuis l'époque des Indiens et de la Frontière". Ainsi, lorsque l'Etat fédéral a tenté, à l'automne, d'arrêter près de 5 000 personnes pour les interroger,* ***les polices locales [SNdef court]****ont été les premières concernées. Dès les attentats, ce sont bien les forces de sécurité locales, police et pompiers, qui ont été immédiatement chargées des opérations.* ***La suite des événements [SNdef long]*** *a encore renforcé le poids des responsabilités sur les premiers secours (first responders), principalement du ressort des autorités locales. [GEOPO_1]*

(VII.17) *De même,* ***le projet de création d'un Presidential Space Advisory Group [SNdef long]*** *n'est pas repris.* ***Ce groupe de conseillers de très haut vol (SNdem long)*** *aurait compté des membres de la communauté scientifique civile, des chefs d'entreprises aéronautiques et spatiales, et des conseillers militaires. Il se serait réuni à la demande du président pour se prononcer sur les politiques spatiales aussi bien civiles que militaires, commerciales que scientifiques.* ***Ce groupe [SNdem court]*** *n'aurait pas fait partie du gouvernement, à proprement parler. En cela, il différait du National Space Council, qui a existé de 1958 à 1973, puis de 1988 à 1993.* ***Ce conseil [SNdem court]****, qui était placé au même niveau de l'exécutif, avait une existence plus formelle et un large pouvoir de décision. [GEOPO_3]*

VII.2.3. Récapitulatif des annotations générées

À cette étape de l'étude, nous disposons d'un grand nombre de données analysées automatiquement. Les annotations générées ne seront pas toutes utilisées dans nos analyses quantitatives. Nous avons opté pour une annotation maximale des informations contenues en position initiale de phrases. Notre approche étant exploratoire, rien ne permettait de prédire quel type d'information serait utile ou inutile. De plus, dans un souci de 'recyclabilité', nous avons jugé plus économique de prendre en compte le maximum d'informations. Au fur et à mesure de notre analyse, nous nous sommes vite rendue compte que la quantité d'informations annotées était telle qu'un travail de thèse ne pouvait toutes les prendre en compte. Nous avons toutefois tenu à présenter ces informations afin de montrer l'étendue de nos données, le point de départ de notre analyse *data-driven*. Le tableau VII.13 présente ainsi toutes les informations annotées automatiquement par notre programme, même celles dont nous ne nous servirons pas dans nos analyses. Par exemple, les mentions « MotFreq » ou « Maillon » ne sont présentes que dans ce tableau. Ces annotations peuvent être soumises à différentes transformations. Un exemple de transformation est donné dans l'annexe K qui présente un extrait du corpus GEOPO annoté au format XML accompagné des définitions des balises XML générées.

	Champ	Description et annotations
LOCALISATION	Corpus	Nom du sous-corpus
	NumPhr	Identifiant de la phrase
	PosTxt	Position de la phrase dans le texte (initiale de sections = S1, initiale de paragraphes = P1, initiale de phrases intraparagraphiques = P2)
	Texte	Identifiant du texte
	NumSec	Identifiant de la section
	NumPara	Identifiant du paragraphe
	DerdeSec	Si cette phrase est la dernière de la section : oui/non
	DerdePara	Si cette phrase est la dernière du paragraphe : oui/non
Titre		Dernier titre de section repéré
Quoi		Caractérisation de la phrase : "TITRE", "AMORCE", "ITEM", "TEXTE"

Champ		Description et annotations
Puce		Numérotation du titre ou puces {"-", "*", ...}
Ponct		Ponctuation finale de la phrase précédente ("S" "§" "." ";" ":" "?" "!" "...")
	Connect	S'il y a connecteur, occurrence de ce connecteur
	e_Connect	Suite d'étiquettes TreeTagger du connecteur (séparateur = " ")
INIT1	TypINIT1	Nature, fonction et rôle sémantique de l'INIT1 (ex : "SP_CIRC_tps")
	INIT1	Texte de l'INIT1
	t_INIT1	Tête de l'INIT1
	e_INIT1	Suite d'étiquettes TreeTagger de l'INIT1 (séparateur = " ")
	Pos_INIT1	Localisation par rapport à la phrase la plus proche comportant un INIT1_CIRC annoté. Cette localisation commence à -1 (phrase précédant celle dans laquelle il y a un INIT1_CIRC) et s'arrête à la fin d'un paragraphe ou lorsqu'un INIT1_CIRC de même rôle sémantique est repéré. L'annotation -1(SP__CIRC_tps) signifie que cette phrase succède à une phrase présentant un INIT1 de type SP_CIRC_tps. Note : si la phrase dans laquelle il y a un INIT1_CIRC commence un nouveau paragraphe, la localisation -1 (TypInint1) n'est pas annotée car jugée non pertinente.
INIT2	TypINIT2	Nature, fonction et rôle sémantique de l'INIT2
	INIT2	Texte de l'INIT2
	t_INIT2	Tête de l'INIT2
	e_INIT2	Suite d'étiquettes TreeTagger de l'INIT2 (séparateur = " ")
ThTop	TypThTop	Nature de l'élément sujet ThTop
	ThTop	Texte du ThTop
	t_ThTop	Tête du ThTop
	e_ThTop	Suite d'étiquettes TreeTagger du ThTop (séparateur = " ")
ThSpe	TypThSpe	Type de ThSpe
	TypFocus	Nature de l'élément focalisé par le ThSpe (cette nature est identique à celle du ThTop)
	Focus	Texte de l'élément focus
	t_Focus	Tête de l'élément focus
	e_Focus	Suite d'étiquettes TreeTagger du Focus (séparateur = " ")
e_Verb		Suite d'étiquettes TreeTagger du syntagme verbal tel que délimité par Syntex
Pred		Texte du prédicat. Dans le cas des ThTop, le prédicat comprend le verbe principal et tout ce qui suit. Dans le cas des ThSpe, le prédicat est considéré comme toute la construction spéciale (tout ce qui suit les INIT).
(Co)Référence et Accessibilité	Reprise_Tit	Si un élément du titre de la section est présent dans cette phrase, ce champ enregistre par quel élément est effectuée cette reprise : "INIT1", "INIT2", "ThTop" ou "Pred".
	Titre_Repris	Si la phrase en cours est un titre de section, ce champ contient le nombre de reprises d'un élément de ce titre dans la section.
	Redenomin ThTop	Si la tête du ThTop ou du Focus a déjà été mentionnée dans la section en cours, localiser cette précédente mention.
	Redenomin INIT	Si la tête de l'INIT1 a déjà été mentionnée dans la section en cours, localiser cette précédente mention (voir le champ précédent)
	MotFreq	Si la phrase présente une ou plusieurs occurrence(s) de mots les plus récurrents (voir annexe C), localisation de cette ou ces occurrence(s). Par exemple, "x:tTT;Pred" signifie que le mot récurrent *x* apparaît en tête de ThTop et dans le prédicat
	Maillon	Si le ThTop de cette phrase est un PRO3, rechercher s'il est de même genre et nombre que le ThTop de la phrase précédente (que ce soit un pronom ou non). Si oui, alors cette phrase constitue le deuxième maillon d'une TPconstante ("2"), la phrase précédente en constitue le premier maillon ("1"). Selon la même règle, la phrase suivante peut constituer la troisième ("3"), et ainsi de suite.
	DegAccess	Degré d'accessibilité de la phrase en prenant pour référent son sujet grammatical

Tableau VII.13 : Récapitulatif des annotations appliquées au corpus d'étude

VII.3. Analyses quantitatives effectuées sur le corpus

Nos méthodes d'analyses ont pour base la variation. Pour observer si des différences significatives au niveau de la position initiale distinguent nos trois sous-corpus, les mesures effectuées sont considérées pour chaque sous-corpus séparément. Les variations observées sont confrontées à nos intuitions sur les stratégies textuelles à l'œuvre dans les sous-corpus, ainsi qu'à la différence de structuration visuelle des trois sous-corpus[144] (voir VII.1.1.a).

Nos analyses peuvent être divisée en quatre étapes :

1. mesure de la composition théorique de la position initiale (sur le corpus entier) ;
2. mesure des spécificités de chaque sous-corpus ;
3. mesure des spécificités des différentes positions textuelles pour le corpus entier et chaque sous-corpus;
4. mesure des corrélations entre certains éléments ressortissants des groupes de mesure précédents.

La composition générale de la position initiale telle qu'observée dans tout le corpus constitue notre modèle théorique. Ainsi, l'hypothèse nulle consiste à associer à chaque élément caractérisé la même distribution que celle observée dans le corpus entier. Cette hypothèse nulle est d'abord testée pour chaque sous-corpus.

Un élément qui montre des variations selon les sous-corpus est un élément qui peut permettre de distinguer un sous-corpus et donc ses particularités telles que présentées en VII.1.1. Cette observation nous permet de mesurer les écarts à l'intérieur d'un sous-corpus (distinguer des comportements particuliers à tel sous-corpus. Par exemple, il se peut que dans ATLAS, les adverbiaux temporels soient associés à des Thèmes topicaux de nature différente de ceux dans PEOPL). En même temps que cette première analyse met au jour les spécificités de chaque sous-corpus, elle nous permet de poser des sous-modèles théoriques, associés à chaque sous-corpus. La mesure des spécificités des différentes positions textuelles peut alors s'effectuer sur la base d'hypothèses nulles spécifiques à chaque sous-corpus.

Un des grands axes de ce travail repose sur le rôle crucial attribué à la position initiale, et ce non seulement au niveau de la phrase mais aussi au niveau des autres unités textuelles que sont les sections et les paragraphes. Pour étudier ce rôle sur plusieurs niveaux, il convient d'observer les écarts significatifs de la fréquence de nos observables entre les trois positions textuelles distinguées : première phrase de section (S1), première phrase de paragraphe (P1), ou phrase intraparagraphique (P2). Cela est bien entendu à comparer entre sous-corpus, le rôle discursif de tel indice n'étant pas considéré comme fixe à travers différents types de textes.

Nos hypothèses concernant ce facteur de variation sont les suivantes : un élément qui apparaît préférentiellement en P2 est candidat à la fonction d'indice de continuité et de connexion. S'il apparaît préférentiellement en S1 ou P1, il est candidat à la fonction d'indice de discontinuité et d'orientation. Bien qu'au départ, nous ne faisions pas de différence entre initiale de sections et initiale de paragraphes (nous mesurions les résultats en confondant ces deux positions textuelles), la distinction entre S1 et P1 s'est rapidement imposée. En effet, la composition de la position initiale en S1 et en P1 est réellement différente. Cela nous fait entrevoir une forte corrélation entre S1 et rupture d'une part, et P1 et déplacement d'autre part.

La recherche des configurations d'indice de séquentialité s'effectue en plusieurs étapes. La première consiste à mesurer les variations significatives selon les différentes positions textuelles. Nous obtenons ainsi des associations 'préférées' entre les trois positions et des formes particulières. Une forme qui apparaît significativement plus dans une position particulière, indifféremment du sous-corpus envisagé, a de fortes chances de constituer un indice du niveau de segmentation associé à cette position. Une forme qui apparaît significativement plus dans une position particulière

144 Pour rappel, deux types de textes se distinguent : des textes longs aux paragraphes courts (ATLAS) et des textes courts aux paragraphes longs (GEOPO et PEOPL).

mais seulement dans un sous-corpus, a de fortes chances de constituer un indice du niveau de segmentation associé à cette position pour un type de texte particulier. Cette forme peut alors être révélatrice d'un certain mode organisationnel. Ainsi, nous pouvons observer des modes organisationnels qui restent stables à travers les différents types de textes et des modes organisationnels propres à un certain sous-corpus. Nous verrons par exemple que les adverbiaux temporels montrent des comportement similaires dans les trois sous-corpus, alors que les adverbiaux spatiaux ne constituent des indices de séquentialité que dans ATLAS.

À la lumière des variations observées au niveau de la composition générale, des formes particulières ont émergé. Ces formes sont remarquables pour leur grande sensibilité aux facteurs de type de texte et de position textuelle. Elles deviennent ainsi le départ de la dernière étape de notre analyse. Cette étape consiste à vérifier si ce n'est pas justement la position textuelle qui attribue à telle ou telle forme une fonction d'indice de segmentation. Pour ce faire, nous avons mesuré l'influence de ces formes sur leur environnement, *i.e.* sur les autres éléments en position initiale de la phrase d'accueil et de la phrase suivante, et cela en dehors des positions textuelles vraisemblablement associées à ces formes.

Pour tous les éléments détachés pertinents, nous avons comparé la distribution des degrés d'accessibilité qui le suivent au modèle théorique. Les éléments qui présentent des écarts significatifs sont des indices potentiels de (dis)continuité touchant les progressions thématiques. Cette comparaison est effectuée en distinguant les différents positions textuelles et sous-corpus (comparaison avec le modèle théorique). Si un élément montre un comportement identique à celui observé dans le modèle théorique, c'est que sa présence ou son absence n'entraîne pas de comportement spécifique au niveau des progressions thématiques et donc qu'il ne marque pas une TSC différente.

Par exemple, nous associons les adverbiaux temporels à des phénomènes de rupture ou de déplacement. Cette hypothèse est soutenue par leur fréquence d'apparition en initiale de sections ou de paragraphes. Pour valider cette hypothèse, nous devons vérifier si, en dehors de ces positions textuelles, la présence de ce type d'adverbial entraîne effectivement des ruptures et des déplacements. Pour cela, nous nous appuyons sur l'observation de ce qui se passe au niveau des degrés d'accessibilité. Il s'agit alors de voir si une distribution particulière des différents degrés d'accessibilité se dessine lorsqu'il y a un adverbial temporel en INIT1. Le même type d'analyse est réalisé pour tester la valeur d'indice des Thèmes topicaux et spécifiques. Par exemple, nous associons les pronoms aux phénomènes de continuité. Mais que se passe-t-il en initiale détachée des phrases présentant un pronom ? Nous avons donc observé, pour tous les Thèmes topicaux et spécifiques pertinents, la distribution des INIT qui le précèdent. La mesure des variations significatives par rapport au modèle théorique nous informe alors sur la cohabitation entre un ThTop/ThSpe et un INIT ou l'absence d'INIT.

La fonction discursive d'un élément en position initiale ne se résume pas à la phrase dans laquelle il apparaît. Notre processus d'annotation nous informe sur la position des phrases relativement au dernier adverbial circonstanciel rencontré. Ce processus s'interrompt dès lors qu'il y a changement de paragraphe ou l'apparition d'un autre adverbial circonstanciel de même rôle sémantique. Nous pouvons donc mesurer les variations observées au niveau des Thèmes topicaux ou spécifiques dans la phrase se situant juste après celle dans laquelle un adverbial circonstanciel a été rencontré, et cela selon le type sémantique du circonstant, les sous-corpus et les positions textuelles.

Toutes ces analyses de cas particuliers nous informent sur la capacité des différents éléments à constituer un indice de séquentialité. Nous pouvons, d'après ces analyses, définir le comportement des différents éléments repérés selon leur affiliation ou non à un type de texte particulier, à une position textuelle particulière et à un environnement spécifique.

VII.4. Mise en oeuvre informatique du repérage et de la caractérisation des observables[145]

Cette partie expose comment nous avons construit le traitement automatique du repérage et de la caractérisation des éléments décrits jusqu'ici et comment fonctionne ce programme d'analyse. Le recours à des traitements automatiques fait partie intégrante de notre thèse. Nous avons déjà expliqué notre volonté de travailler sur un grand volume de données et de prendre en compte tous les éléments apparaissant en position initiale (chapitre VI). Ces besoins ne peuvent être satisfaits sans la réalisation d'un programme adapté à notre objet d'étude. La réalisation de ce programme représente une partie importante du travail réalisé pour cette thèse. Cette réalisation s'est fait progressivement et de façon exploratoire. Nous ne savions pas vraiment quels objets notre programme allait devoir caractériser. Nous savions simplement que ces objets se situaient en position préverbale de phrase.

Ce programme se base sur la sortie d'un analyseur syntaxique opérationnel : SYNTEX (Bourigault 2007). Nous n'avions pas d'hypothèses de départ quant à l'utilité d'un tel analyseur. Au début, notre programme se basait uniquement sur les étiquettes morpho-syntaxiques fournies par le TreeTagger (nous avions fait également des tentatives avec les sorties de l'étiqueteur Cordial[146]). Au même moment se développait au sein de l'ERSS l'élaboration de l'outil SYNTEX, outil qui intégrait dans sa chaîne de traitement le TreeTagger. Cette conjonction d'événements[147] nous a amené à nous intéresser aux sorties de cet analyseur et à nous rendre compte progressivement des informations que nous pouvions (et ne pouvions pas) utiliser pour notre étude.

Nous avons dans un premier temps créé un programme permettant de délimiter la position préverbale des phrases et d'y repérer les éléments détachés en initiale et les sujets grammaticaux. C'est lors de cette première phase que nous avons pris conscience de la nécessité de traiter la présence de connecteurs en initiale de phrase et les cas de constructions spéciales. Nous nous sommes ensuite attaquée à la caractérisation des éléments repérés. Nous avions une certaine expérience de la caractérisation des introducteurs de cadre, expérience acquise lors de notre travail de maîtrise et de DEA. Par contre, nous n'avions que peu d'expérience quant aux autres types d'éléments présents en position préverbale. La caractérisation de ceux-ci s'est donc faite à tâtons, par observation des données elle-mêmes et confrontation de ces données aux grammaires existantes du français (principalement la grammaire méthodique de Riegel *et al.*), jusqu'à aboutir à un traitement qui nous a semblé fiable et d'assez large couverture. La dernière étape de réalisation du programme d'analyse a concerné le traitement des informations de redénomination, de récurrence lexicale, de degré d'accessibilité, de reprise des titres de section, etc. La prise en compte de ces informations pour l'analyse est apparue autant lors de l'avancée de notre réflexion linguistique que lors de l'avancée de notre programme informatique.

Le programme que nous avons créé est réalisé dans l'environnement du logiciel MSAccess© développé par la société Microsoft. Ce logiciel, dédié à la gestion de bases de données et la création d'interfaces de type formulaire, est associé au langage de programmation Visual Basic (langage orienté objet). De façon générale, le programme génère une base de donnée dans laquelle chaque sous-corpus est représenté sous forme de quatre tables : deux tables sont le résultat de l'importation des fichiers sources (une table contenant le texte d'origine avec quelques modifications et une table contenant les résultats de l'analyse Syntex de chaque phrase du sous-corpus) et une table contient les

145 Nous remercions Didier Bourigault pour nous avoir motivée et guidée dans la rédaction post-soutenance de cette partie et nous avoir permis de mettre un peu plus en avant l'aspect TAL de notre travail.

146 Cordial est développé par l'entreprise 'Synapse développement' http://www.synapse-fr.com/.

147 À ces événements, nous ajoutons notre collaboration avec Cécile Frérot (Hồ-Đắc & Frérot 2004) qui a joué un grand rôle dans notre attirance pour SYNTEX.

résultats de notre traitement automatique de la position préverbale de chaque phrase. Enfin, une table rassemble les données des trois sous-corpus pour réaliser les analyses sur le corpus en entier.

VII.4.1. Fichiers sources

VII.4.1.a) Les textes sources

Le traitement nécessite deux fichiers sources : un fichier contenant les textes à analyser et un fichier contenant l'analyse syntaxique des phrases des textes. Les textes ont préalablement subi un certain 'nettoyage'. Les éléments non textuels (tableaux, illustrations, figures, etc.) ainsi que les éléments situés 'en dehors' de la linéarité du texte ou en annexes (notes de bas de page et numéro de renvoi inséré dans le texte même, commentaires mis en marge, sommaires, bibliographies, glossaires, etc.) sont enlevés. Les changements de paragraphe ou de section sont uniformisés à un alinéa, les puces de liste énumérative sont transformées en un caractère reconnaissable par les codages iso ou UTF. Enfin, tous les texte d'un même sous-corpus sont intégrés dans un même fichier texte délimités par un identifiant de type TXT1, TXT2, TXT3, etc.

Le travail de pré-traitement a été plus fastidieux pour le sous-corpus GEOPO étant donné le format pdf des fichiers d'origine. La copie (ou l'exportation) du texte d'un fichier pdf ne prend pas en compte la mise en forme matérielle d'un document. Ainsi, chaque fin de ligne est collée (ou exportée) comme un renvoi à la ligne (qui correspond au caractère non imprimable ↵). Il faut donc remplacer ces caractères par le caractère non imprimable ¶ dans les cas de changements de paragraphe ou par un espace typographique dans les cas de continuité intraparagraphique. Dans les cas de textes en colonnes, l'opération est plus lourde. Le découpage en colonnes n'étant pas pris en compte lors de l'export, le texte d'une ligne de la colonne de gauche se retrouve concaténé au texte de la ligne de la colonne de droite correspondante. Il faut donc copier colonne par colonne (avec l'outil de sélection livré sous Acrobat Reader©) avant de transformer les fins de lignes. Ces opérations minutieuses de copier-coller et de remplacement sélective des sauts de ligne en changement de paragraphe ou pas ont représenté un certain temps de travail (entre dix et vingt heures pour le sous-corpus GEOPO), sans compter le temps passé à la sélection des fichiers pdf (*i.e.* des fichiers non protégés contre la copie et présentant des textes assez longs).

Le but de ce pré-traitement est de conserver au maximum la linéarité du texte d'origine. Ces textes 'nettoyés' sont ensuite analysés syntaxiquement par le logiciel SYNTEX.

VII.4.1.b) Les fichiers anasynt produits par de SYNTEX

L'analyse syntaxique des textes est effectuée par le logiciel SYNTEX développé par Didier Bourigault (Bourigault 2007). SYNTEX a constitué pour nous un outil nécessaire afin de (1) distinguer la zone préverbale du reste de la phrase, (2) délimiter des syntagmes complets afin d'isoler les différents éléments présents en position préverbale, et (3) repérer les têtes nominales des syntagmes, utiles pour identifier les reprises lexicales. Sa capacité à analyser rapidement et efficacement les données sans nécessiter d'autres ressources que le texte brut à analyser constitue un avantage indéniable lorsque l'on travaille sur de grandes quantités.

SYNTEX est un analyseur syntaxique opérationnel et 'écologique' qui se base sur les productions réelles d'un corpus pour se constituer, par apprentissage endogène, les ressources lexicales dont il a besoin pour désambiguïser

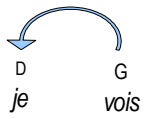
les cas de dépendances syntaxiques équivoques[148]. Ces ressources permettent par exemple de savoir la probabilité d'association de chaque verbe à telle ou telle préposition. Grâce à cet apprentissage endogène, l'analyse s'adapte au corpus (on peut supposer que les probabilités d'association varient selon le type de contenu présent dans le corpus).

> « L'objectif du projet SYNTEX [Bourigault & Fabre 2000] était la construction d'un analyseur opérationnel, précis et efficace, qui produise des analyses aussi correctes et complètes que possible, sur des textes de genres variés, avec des temps de traitement raisonnables pour être compatibles avec la nécessité d'aborder des volumes de plus en plus importants ; un analyseur qui soit utilisable dans une large gamme d'applications, que ce soit du côté de la recherche académique, en linguistique, sciences humaines, intelligence artificielle, ou de celui des applications industrielles, pour la construction d'ontologies, le traitement de l'information, la recherche d'information, etc. » (Bourigault 2007:67)

L'analyse réalisée par SYNTEX est le fruit d'une chaîne de traitement dans laquelle SYNTEX ne représente 'que' la fonction d'analyseur en dépendances. Cette chaîne de traitement est organisée en quatre phrases :

1. préétiquetage et segmentation du fichier d'entrée (fichier texte) : segmentation en phrases, tokénisation en mots et préétiquetage des mots ;
2. étiquetage morpho-syntaxique effectué par l'outil Treetagger[149] ;
3. conversion des sorties du Treetagger pour les adapter à l'analyse Syntex ;
4. analyse en dépendances effectué par SYNTEX.

Au sortir de cette chaîne de traitement, à chaque mot sont associés sa position dans la phrase, son lemme, sa catégorie morpho-syntaxique et ses dépendances syntaxiques. Les dépendances syntaxiques sont orientées *i.e.* elles vont d'un mot gouverneur (G) vers un mot dépendant (D) :



Ainsi, chaque mot peut gouverner d'autres mots et être gouverné par un mot (et un seul). Ces dépendances sont caractérisées syntaxiquement par un jeu de relations : relation sujet (SUJ), relation déterminant (DET), relation de coordination (CC), etc.

Par exemple, la phrase :

> *Lancée au printemps 2001, cette étude devait s'achever en début de année 2002.* *[GEOPO_2]*

donne en sortie de SYNTEX (dans un fichier à extension .anasynt)[150] :

> *<SEQ id=xxxxxxxxxxxxx>*
> *<TXT> Lancée au printemps 2001 , cette étude devait s ' achever en début de année 2002.*
> *<ETIQ> PpaFS|lancer|Lancée|1|0|PREP;2 Prep|à|à|2|PREP;1|NOMPREP;4 DetMS|le|le|3|DET;4|0 NomMS|printemps|printemps|4|NOMPREP;2|DET;3,EPI;5 NomXXDate|2001|2001|5|EPI;4|0 Typo|,|,|6|0|0 DetFS|ce|cette|7|DET;8|0 NomFS|étude|étude|8|SUJ;9|DET;7 VCONJS|se achever|devait s' achever|9|0| SUJ;8,PREP;10 Prep|en début de|en début de|10|PREP;9|NOMPREP;11 NomXXDate|année|année|11| NOMPREP;10|EPI;12 NomXXDate|2002|2002|12|EPI;11|0 Typo|.|.|13|0|0*

La ligne commençant par « <SEQ... » donne l'identifiant de la phrase, la ligne commençant par <TEXT> donne le texte de la phrase et la ligne commençant par <ETIQ> donne le résultat de l'analyse. Chaque information syntaxique relative à un mot est séparée par un | et chaque analyse de mot est séparée par une tabulation. Par exemple, l'analyse du mot *étude* se lit de la façon suivante :

148 L'analyse ne se base cependant pas uniquement sur de l'apprentissage endogène, SYNTEX utilisant, dans certains cas seulement, des ressources lexicales extérieures, pour le repérage des locutions adverbiales par exemple.

149 Le Treetagger a été conçu à l'université de Stuttgart (http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger). Les autres modules ont été créés par les mêmes concepteurs que Syntex *i.e.* Didier Bourigault, l'ERSS et la société Synomia.

150 Ce résultat est celui fournit par la version de SYNTEX datant de 2004. De nombreuses évolutions ont été apportées depuis comme par exemple le rattachement des appositions (cf. Bourigault 2007).

- le nom commun féminin singulier (NomFS)
- de lemme *étude*
- et d'occurrence *étude*
- constitue le huitième mot de la phrase (8)
- est gouverné par le neuvième mot de la phrase (la forme conjuguée du verbe *se achever*) par une relation sujet (SUJ;9)
- et est gouverneur du septième mot de la phrase (le déterminant féminin singulier correspondant à l'occurrence *cette*).

Il faut ici noter que SYNTEX ne rattache jamais les éléments détachés en initiale et donc n'analyse pas leur relation avec le reste de la phrase (ce que réalise notre programme). Ainsi, dans l'analyse proposée en exemple ci-dessus, l'apposition détachée en initiale n'est rattaché à rien. SYNTEX nous permet de reconstituer le bloc appositif dans son entier, mais pas de prédire sa fonction d'apposition.

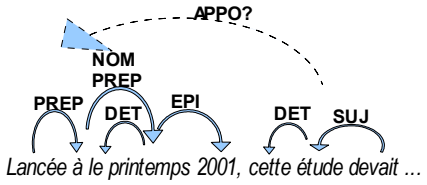


Figure VII.1: Exemple de sortie de l'analyseur SYNTEX

VII.4.2. Traitement et analyse des données

Nous avons réaliser une chaîne de traitement qui comprends quatre modules successifs correspondant chacun à une étape de l'analyse. Lors de ces différentes étapes, les résultats issus des modules précédents sont parfois modifiés.

1. Un **module d'import** (environ 1200 lignes de programme) acquiert les données textuelles pour chaque sous-corpus dans deux tables (deux tables par sous-corpus) :
   - La table « *corpus*_PARA » rassemble les textes non 'syntexisés'. Chaque ligne représente un paragraphe tel que délimité typodispositionnellement. Chaque paragraphe est associé (automatiquement) à un statut : TITRE, TEXTE, ITEM, AMORCE.
   - La table « *corpus*_SYNTEX » rassemble les résultats de l'analyse Syntex. Chaque ligne représente une phrase (telle que délimitée par le module de segmentation de SYNTEX) et met en regard le texte de la phrase et son analyse syntaxique.

TEXTE	Analyse Syntex[151]	Quoi[152]
Lancée au printemps 2001, cette étude devait s' achever en début de année 2002 .	PpaFS_lancer_Lancée_1_0_PREP;2_Z7Z_Prep_à_à_2_PREP;1_NOMPREP;4_Z7Z_DetMS_le_le_3_DET;4_0_Z7Z_NomMS_printemps_printemps_4_NOMPREP;2_DET;3,EPI;5_Z7Z_NomXXDate_2001_2001_5_EPI;4_0_Z7Z_Typo_._,_6_0_0_Z7Z_DetFS_ce_cette_7_DET;8_0_Z7Z_NomFS_étude_étude_8_SUJ;9_DET;7_Z7Z_VCONJS_se achever_devait s' achever_9_0_SUJ;8,PREP;10_Z7Z_Prep_en début de_en début de_10_PREP;9_NOMPREP;11_Z7Z_NomXXDate_année_année_11_NOMPREP;10_EPI;12_Z7Z_NomXXDate_2002_2002_12_EPI;11_0_Z7Z_Typo_._._13_0_0_Z7Z_	TEXTE

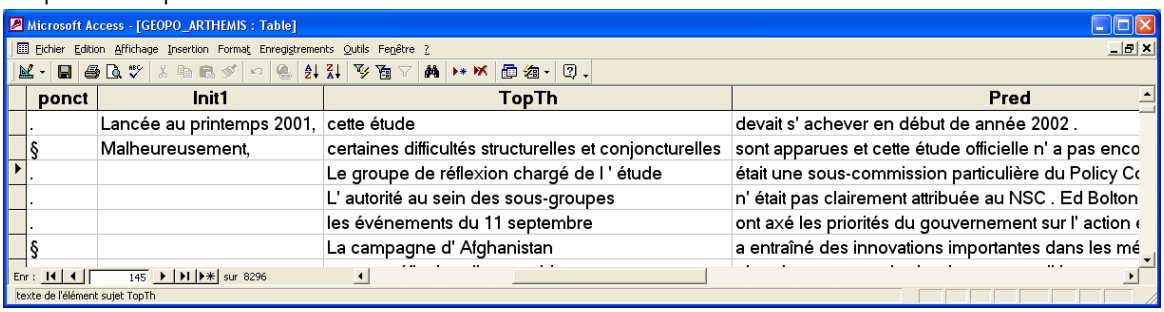
151 Pour des raisons de traitement par un programme en visual basic, nous transformons légèrement le fichier anasynt. Les | sont remplacées par des _, les tabulations sont remplacées par la suite de caractères _Z7Z_, les espaces typographiques précédents les virgules et les apostrophes dans la ligne <TXT> sont supprimés.

152 Cette colonne reprend le statut du paragraphe dans lequel apparaît la phrase. Dans les cas d'amorce d'énumération, on peut avoir un paragraphe TEXTE dont la dernière phrase a la statut d'AMORCE.

2. Un **module de segmentation** (environ 3000 lignes de programme) délimite les éléments en position préverbale de chaque phrase : Connect, INIT1, INIT2, ThTop et ThSpe. Les étapes de ce module sont les suivantes :

- repérage des puces et numérotations,
- repérage des connecteurs 'purs',
- repérage du (premier) verbe principal de la phrase,
- délimitation de la zone préverbale,
- repérage des différents blocs Syntex (par récursivité, on part d'un mot et on suit les relations de dépendance jusqu'à arriver au syntagme le plus complet[153]),
- analyse de ces blocs pour distinguer les éléments détachés en initiale des sujets grammaticaux,
- analyse des phrases pour détecter les constructions spéciales et typage de ces constructions.
  (de plus amples précisions sont données dans l'annexe H).

Cette délimitation aboutit à la création d'une nouvelle table (*corpus*_ARTHEMIS) qui contient les éléments de chaque sous-corpus.


Microsoft Access - [GEOPO_ARTHEMIS : Table]

Fichier Edition Affichage Insertion Format Enregistrements Outils Fenêtre ?

ponct	Init1	TopTh	Pred
.	Lancée au printemps 2001,	cette étude	devait s' achever en début de année 2002 .
§	Malheureusement,	certaines difficultés structurelles et conjoncturelles	sont apparues et cette étude officielle n' a pas enco
.		Le groupe de réflexion chargé de l ' étude	était une sous-commission particulière du Policy Co
.		L' autorité au sein des sous-groupes	n' était pas clairement attribuée au NSC . Ed Bolton
.		les événements du 11 septembre	ont axé les priorités du gouvernement sur l' action e
§		La campagne d' Afghanistan	a entraîné des innovations importantes dans les mé

Enr : 145 sur 8296

texte de l'élément sujet TopTh

Figure VII.2: Extrait de la table GEOPO_ARTHEMIS résultant du module de segmentation du traitement automatique de la position initiale

Lors de cette étape, certaines modifications sont effectuées dans la table *corpus*_SYNTEX, modifications relatives à certaines erreurs d'analyse ou d'étiquetage morpho-syntaxique. Par exemple, la phrase suivante montre une erreur de l'étiquetage réalisé par le TreeTagger :

L' analyse de Carroll voyait l' esprit occupé d' une chose, préoccupé d' une autre, d' où l' interférence, et le lapsus .

est analysée par SYNTEX de la façon suivante (nous ne mettons que le début de l'analyse) :

Pro?S|le|L'|1|OBJ;2|0 VCONJ|analyser|analyse|2|0|OBJ;1,PRDE;3 Prep|de|de|3|PRDE;2|NOMPREP;4 NomPrXXPrenom|Carroll|Carroll|4|NOMPREP;3|0 VCONJS|voir|voyait|5|0|OBJ;7 DetMS|le|l'|6|DET;7|0 NomMS|esprit|esprit|7|OBJ;5|DET;6,ADJ;8,PRDE;9 PpaMS|occuper|occupé|8|ADJ;7|0 Prep|de|d'|9|PRDE;7|NOMPREP;11

La phase de segmentation nous permet d'identifier *L'analyse* comme étant le sujet (bien qu'aucune relation de la sorte n'ait été analysée par Syntex) et transforme ainsi l'analyse Syntex en[154] :

Det??|le|L'|1|OBJ;2|0 Nom??|analyser|analyse|2|0|OBJ;1,PRDE;3 Prep|de|de|3|PRDE;2|NOMPREP;4 NomPrXXPrenom|Carroll|Carroll|4|NOMPREP;3|0 VCONJS|voir|voyait|5|0|OBJ;7 DetMS|le|l'|6|DET;7|0 NomMS|esprit|esprit|7|OBJ;5|DET;6,ADJ;8,PRDE;9 PpaMS|occuper|occupé|8|ADJ;7|0 Prep|de|d'|9|PRDE;7|NOMPREP;11

3. Un **module de caractérisation** (environ 1300 lignes de programme), qui se base essentiellement sur des expressions régulières, associe à chaque élément de la table *corpus*_ARTHEMIS les informations telles que définies dans le tableau VII.13 p. . Chaque élément subit plusieurs passes. Pour les éléments détachés en initiale, la première passe consiste en une catégorisation morpho-syntaxique effectuée à l'aide des étiquettes morpho-syntaxiques issues

153 Notre objet d'étude étant la position initiale, la relation sujet (lorsqu'elle était gouvernée par le verbe principale de la phrase) constitue pour nous une relation marquant une fin de bloc. Si l'on suivait les relations Syntex jusqu'au bout, on obtiendrait la proposition entière (avec verbe et compléments rattachés).

154 Nous ne modifions que les informations nécessaires à notre traitement, ce qui explique que nous ne modifions ni les dépendances (OBJ -> DET) ni les lemmes des mots à l'analyse modifiée.

du Treetagger. Ensuite, sa fonction et son rôle sémantique sont déterminés à l'aide d'une liste d'expressions régulières (liste données dans l'annexe H). Cette liste s'est enrichie progressivement au cours de la construction et de l'affinage du programme. Pour les Thèmes topicaux, l'utilisation des étiquettes morpho-syntaxiques a généralement suffit sauf pour la détection des redénominations (une étape du programme y est consacré). Les autres informations sont le résultat de différents programmes intégrés à ce module.

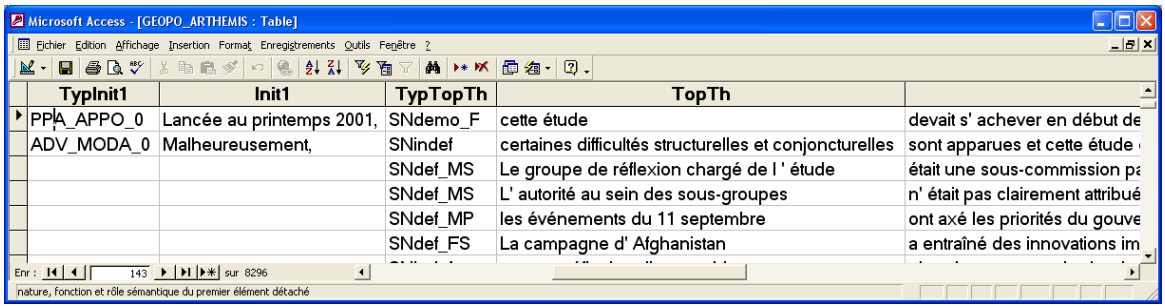


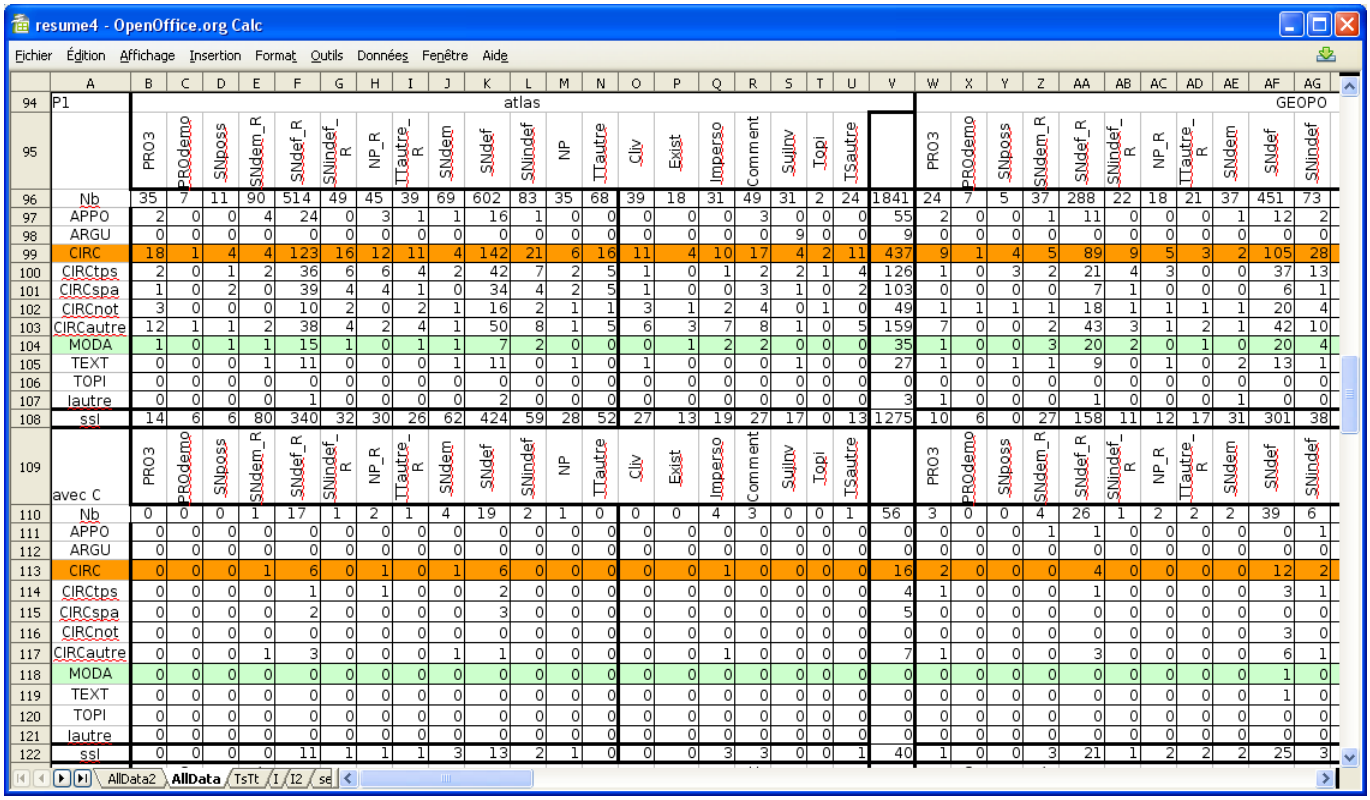
Figure VII.3: Extrait de la table GEOPO_ARTHEMIS résultant du module de caractérisation du traitement automatique de la position initiale

4. Un **module de comptage** (environ 800 lignes de programme) génère à partir des tables *corpus*_ARTHEMIS de chacun des sous-corpus une table générale (COMPTA). Chaque ligne correspond à une phrase pour laquelle sont notés : son sous-corpus d'origine, sa position textuelle, ses identifiants (de phrase, de paragraphe, de texte), la présence d'un Connect, l'étiquette morpho-fonctiono-sémantique de l'INIT si il y a INIT, l'étiquette morpho-syntaxique de son ThTop ou le type de ThSpe, le fait qu'il y a ou non reprise d'un élément du titre, le fait qu'il y ait ou non reprise en ThTop, en INIT, etc...

Un dernier module (environ 600 lignes de programme), dont les résultats ne sont pas utilisés dans cette thèse, permet la génération d'un fichier xml où la plupart des analyses réalisées sont traduites sous forme de balises (ce fichier est généré par l'export de la colonne ParaVizuXML des tables _PARA). Un extrait de fichier xml généré est donné dans l'annexe K.

VII.4.3. Réalisation des analyses quantitatives

Les possibilités de requête sous MSAccess© permettent de construire des tableaux qui donnent le nombre d'occurrences de chaque combinaison d'INIT/ssINIT - ThTop/ThSpe, pour chaque sous-corpus, dans chaque position textuelle, selon qu'il y ait ou non présence d'un Connect. Ces tableaux sont ensuite exportés sous un tableur (OpenOffice.org), cf. figure VII.4.



resume4 - OpenOffice.org Calc

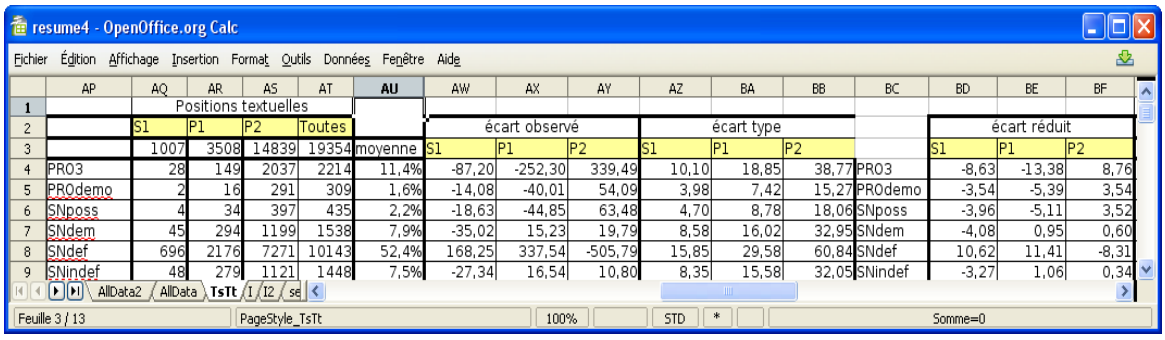
Fichier Édition Affichage Insertion Format Outils Données Fenêtre Aide

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
94	P1									atlas																							GEOPO	
95		PRO3	PROdemo	SNposs	SNdem_R	SNdef_R	SNindef_R	NP_R	TTautre_R	SNdem	SNdef	SNindef	NP	TTautre	Cliv	Exist	Imperso	Comment	Sujinv	Topi	TSautre		PRO3	PROdemo	SNposs	SNdem_R	SNdef_R	SNindef_R	NP_R	TTautre_R	SNdem	SNdef	SNindef
96	Nb	35	7	11	90	514	49	45	39	69	602	83	35	68	39	18	31	49	31	2	24	1841	24	7	5	37	288	22	18	21	37	451	73
97	APPO	2	0	0	4	24	0	3	1	1	16	1	0	0	0	0	0	3	0	0	0	55	2	0	0	1	11	0	0	0	1	12	2
98	ARGU	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	9	0	0	0	0	0	0	0	0	0	0	0
99	CIRC	18	1	4	4	123	16	12	11	4	142	21	6	16	11	4	10	17	4	2	11	437	9	1	4	5	89	9	5	3	2	105	28
100	CIRCtps	2	0	1	2	36	6	6	4	2	42	7	2	5	1	0	1	2	2	1	4	126	1	0	3	2	21	4	3	0	0	37	13
101	CIRCspa	1	0	2	0	39	4	4	1	0	34	4	2	5	1	0	0	3	1	0	2	103	0	0	0	0	7	1	0	0	0	6	1
102	CIRCnot	3	0	0	0	10	2	0	2	1	16	2	1	1	3	1	2	4	0	1	0	49	1	1	1	1	18	1	1	1	1	20	4
103	CIRCautre	12	1	1	2	38	4	2	4	1	50	8	1	5	6	3	7	8	1	0	5	159	7	0	0	2	43	3	1	2	1	42	10
104	MODA	1	0	1	1	15	1	0	1	1	7	2	0	0	0	1	2	2	0	0	0	35	1	0	0	3	20	2	0	1	0	20	4
105	TEXT	0	0	0	1	11	0	0	0	1	11	0	1	0	1	0	0	0	1	0	0	27	1	0	1	1	9	0	1	0	2	13	1
106	TOPI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
107	lautre	0	0	0	0	1	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	3	1	0	0	0	1	0	0	0	1	0	0
108	ssl	14	6	6	80	340	32	30	26	62	424	59	28	52	27	13	19	27	17	0	13	1275	10	6	0	27	158	11	12	17	31	301	38
109	avec C	PRO3	PROdemo	SNposs	SNdem_R	SNdef_R	SNindef_R	NP_R	TTautre_R	SNdem	SNdef	SNindef	NP	TTautre	Cliv	Exist	Imperso	Comment	Sujinv	Topi	TSautre		PRO3	PROdemo	SNposs	SNdem_R	SNdef_R	SNindef_R	NP_R	TTautre_R	SNdem	SNdef	SNindef
110	Nb	0	0	0	1	17	1	2	1	4	19	2	1	0	0	0	4	3	0	0	1	56	3	0	0	4	26	1	2	2	2	39	6
111	APPO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1
112	ARGU	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
113	CIRC	0	0	0	1	6	0	1	0	1	6	0	0	0	0	0	1	0	0	0	0	16	2	0	0	0	4	0	0	0	0	12	2
114	CIRCtps	0	0	0	0	1	0	1	0	0	2	0	0	0	0	0	0	0	0	0	0	4	1	0	0	0	1	0	0	0	0	3	1
115	CIRCspa	0	0	0	0	2	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0
116	CIRCnot	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0
117	CIRCautre	0	0	0	1	3	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	7	1	0	0	0	3	0	0	0	0	6	1
118	MODA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
119	TEXT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
120	TOPI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
121	lautre	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
122	ssl	0	0	0	0	11	1	1	1	3	13	2	1	0	0	0	3	3	0	0	1	40	1	0	0	3	21	1	2	2	2	25	3

AllData2 | AllData | TsTt | I | I2 | se

Figure VII.4: Feuille de calcul rassemblant la mesure des données collectées

Cela permet ensuite de calculer l'écart réduit de chaque élément dans les différentes configurations prises en compte pour l'étude (figure VII.5).



resume4 - OpenOffice.org Calc

Fichier Édition Affichage Insertion Format Outils Données Fenêtre Aide

	AP	AQ	AR	AS	AT	AU	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BF
1		Positions textuelles														
2		S1	P1	P2	Toutes		écart observé			écart type				écart réduit		
3		1007	3508	14839	19354	moyenne	S1	P1	P2	S1	P1	P2		S1	P1	P2
4	PRO3	28	149	2037	2214	11,4%	-87,20	-252,30	339,49	10,10	18,85	38,77	PRO3	-8,63	-13,38	8,76
5	PROdemo	2	16	291	309	1,6%	-14,08	-40,01	54,09	3,98	7,42	15,27	PROdemo	-3,54	-5,39	3,54
6	SNposs	4	34	397	435	2,2%	-18,63	-44,85	63,48	4,70	8,78	18,06	SNposs	-3,96	-5,11	3,52
7	SNdem	45	294	1199	1538	7,9%	-35,02	15,23	19,79	8,58	16,02	32,95	SNdem	-4,08	0,95	0,60
8	SNdef	696	2176	7271	10143	52,4%	168,25	337,54	-505,79	15,85	29,58	60,84	SNdef	10,62	11,41	-8,31
9	SNindef	48	279	1121	1448	7,5%	-27,34	16,54	10,80	8,35	15,58	32,05	SNindef	-3,27	1,06	0,34

AllData2 | AllData | TsTt | I | I2 | se

Feuille 3 / 13 | PageStyle_TsTt | 100% | STD | * | Somme=0

Figure VII.5: Feuille de calcul type pour le calcul de l'écart réduit

La mesure de toutes les autres données (DegAccess, Phr+1) sont effectuées de cette façon. Les écarts réduits sont ensuite représentés sous forme de graphiques qui représentent les écarts réduits de nos indices. La lecture de ces représentations graphiques est expliquée dans la partie suivante.

VII.5. Petit manuel pour la lecture des résultats

Toutes les caractérisations effectuées et les différents points de vue adoptés génèrent un gigantesque volume de données et un risque élevé de s'y noyer, tant pour l'analyste que pour le lecteur. Les pistes pour observer ces données

nous sont apparues progressivement à mesure que nous nous familiarisions avec elles. La difficulté est alors de faire comprendre ces pistes au lecteur qui n'a pas acquis cette familiarisation. Cette petite partie est là pour expliquer comment nous avons choisi de présenter les résultats. Cette explication se base sur un cas simple : la présence d'un INIT (cet exemple est repris en VIII.3).

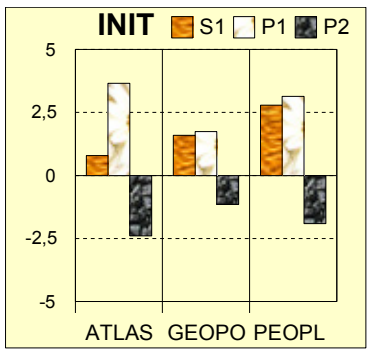
La première méthode consiste à présenter sous forme de tableau le nombre d'occurrences repérées et la proportion que cela représente sur toutes les phrases du corpus. Cette première colonne s'accompagne généralement du détail par sous-corpus. Pour des raisons ergonomiques, seuls les pourcentages par sous-corpus, *i.e.* les profils-colonnes sont indiqués. Le pourcentage représente chaque sous-corpus. Le profil-ligne n'est généralement pas mentionné. Pour comparer les sous-corpus entre eux, nous avons recours au calcul de l'écart réduit tel que présenté en VI.4.3.

	corpus		%			z		
	Nb phrases	%	ATLAS	GEOPO	PEOPL	ATLAS	GEOPO	PEOPL
Avec INIT	7 022	*30*	*27*	*35*	*28*	-6,3	+9,8	-3,7
Sans INIT	16 195	*70*	*73*	*65*	*72*	+6,3	-9,8	+3,7
Nb phrases total	23 217		7 592	7 901	7 724			

Tableau VII.14 : Représentation des données par tableau

Les écarts réduits correspondent à un calcul de la distance entre le nombre de données réellement mesurées et le nombre théorique mesuré sur le corpus entier, le sous-corpus ou une certaine PosTxt dans un certain sous-corpus pour le dernier groupe d'analyse. Plus leur valeur est grande, plus l'écart est dit significatif. L'écart réduit (représenté par la lettre *z*) nous permet de comparer l'influence des facteurs de variation envisagés en mettant à une même échelle les écarts observés.

Pour plus de lisibilité, les écarts réduits peuvent également être représentés sous forme d'histogrammes. Le graphique VII.4 réunit les variations de la répartition des phrases présentant un élément détaché en initiale selon les différentes positions textuelles, pour chaque sous-corpus. Nous voyons ici que ATLAS, GEOPO et PEOPL affichent un comportement différent.

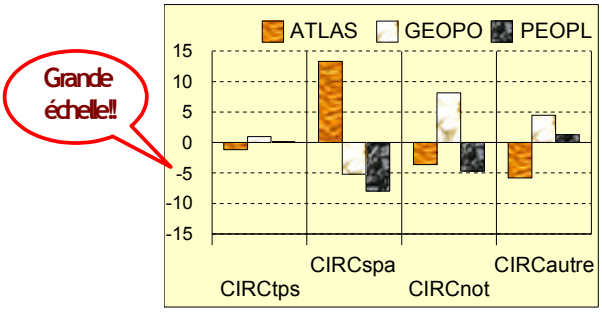


Graphique VII.4: Représentation des écarts réduits

ATLAS et PEOPL présentent significativement plus d'INIT en initiale de paragraphes (P1) et significativement moins à l'intérieur des paragraphes (P2). PEOPL montre également un écart réduit significatif en initiale de sections (S1). GEOPO, lui, ne montre aucun écart significatif, ce qui signifie que la présence ou l'absence d'INIT en initiale de phrases n'est pas influencé par la position textuelle dans ce sous-corpus. Il est cependant celui qui montre le plus d'INIT et sa différence avec les autres est significative vu l'écart *z*=+9,8 mesuré (voir tableau VII.14)

Pour tous nos facteurs de variation, nous avons trois variables. ATLAS, GEOPO et PEOPL sont les variables du facteur type textuel et S1, P1 et P2 sont les variables du facteur position textuelle. Différents schémas d'association entre un type d'élément et un facteur de variation apparaissent à travers nos données. Pour illustrer les différents schémas, nous prenons les variations observées selon les sous-corpus au niveau de la répartition des différents rôles sémantiques des circonstants (graphique VII.5).

Cette bulle informe le lecteur d'un changement important d'échelle au niveau de l'axe indiquant les écarts réduits. Cela permet de souligner la distinction entre des schémas d'écart de forme semblable mais dont les mesures diffèrent fortement. Observez la différence d'échelle entre le graphique précédent et celui-ci.



Graphique VII.5 : Quatre schémas d'association entre un facteur de variation et un type d'élément

Premièrement, nous avons les cas du schéma « neutre » où aucun écart significatif n'est mesuré. Ce schéma signifie que la présence de le comportement de cet élément n'est pas dépendante du facteur discursif considéré. En d'autres termes, aucune variable de ce facteur discursif n'influe sur la distribution du dit élément. Dans notre exemple, les circonstants temporels (CIRCtps) apparaissent insensibles au facteur « type de texte ».

Un second schéma correspond à une « association exclusive » entre une variable et un élément. Ce schéma présente un écart positif fort pour une variable *x* qui entraîne un écart opposé pour les autres variables du même facteur de variation. Au niveau des sous-corpus, l'écart positif élevé signifie que la majorité des éléments de ce type provient d'un sous-corpus spécifique et que l'on peut donc caractériser ce sous-corpus par la forte présence de cet élément. Sans ce sous-corpus, il y aurait significativement moins de ce type d'élément dans nos données. Les variations des adverbiaux spatiaux (CIRCspa) et notionnels (CIRCnot) correspondent typiquement à ce type de schéma. Au niveau des positions textuelles, l'écart positif signifie que l'élément observé est utilisé dans cette position textuelle en particulier. Nous pouvons alors supposer que cet élément accompagne la stratégie de séquentialité qu'indique la position textuelle (S1 et rupture, P1 et déplacement, P2 et continuité).

Un troisième schéma constitue le reflet du schéma de l'association exclusive : la « dissociation exclusive ». Il s'agit d'un écart négatif élevé pour une variable qui entraîne deux écarts positifs chez les autres sous-corpus. L'écart négatif signifie que la faible fréquence de ce type d'élément est caractéristique de cette variable, *i.e.* en dehors de cette variable, on aurait significativement plus de ce type d'élément.

Pour ces deux derniers schémas, plus les écarts sont élevés, plus l'association ou la dissociation entre la variable et l'élément est forte. Dans notre exemple, l'association entre ATLAS et les adverbiaux spatiaux est plus forte que celle observée entre GEOPO et les adverbiaux notionnels.

Un quatrième schéma apparaît également où toutes les variables affichent un comportement différent : un écart neutre, un écart significatif positif et un écart négatif significatif, c'est le schéma de « dispersion » que nous observons pour les adverbiaux circonstanciels « autre » (CIRCautre), *i.e.* ni temporels, ni spatiaux, ni notionnels. Dans les schémas d'association et de dissociation exclusive, les écarts opposés sont une réponse logique à un écart élevé et non la signification que les variables en opposition sont associées à une absence de tel élément. Dans le schéma de dispersion, nous sommes confrontés à un élément très sensible aux différences entre les variables d'un même facteur discursif. Dans ce cas, la moyenne correspond à une distribution pour une variable (celui qui ne montre pas d'écart significatif). Il n'y a pas d'exclusivité pour une variable particulière, mais par contre, des degrés de présence de l'élément qui peuvent caractériser les différentes variables.

Partie 4.

Résultats et Discussion

« Interprétation »

Chapitre VIII
La position initiale : généralités et spécificités des sous-corpus

Sommaire

Ce chapitre a pour but la description générale de ce qui compose la position initiale. Cet état des lieux nous permet de poser le modèle théorique appliqué à notre étude. Ainsi, les variations constatées et exposées dans le chapitre suivant se feront par rapport aux données exposées ici. Cependant, nous observons d'ores et déjà les variations entre sous-corpus. Comme nous l'avons plusieurs fois répété au cours des chapitres précédents, nous ne concevons pas de marquage de la séquentialité sans prise en compte du type de texte analysé. Nous posons d'ailleurs l'hypothèse générale qu'il existe un marquage commun aux textes expositifs et des marquages spécifiques à chaque sous-corpus.

Dans un premier temps, nous présentons comment se répartissent chaque type d'observables : connecteurs – Connect, éléments détachés en initiale – INIT, Thèmes topicaux – ThTop – et constructions avec Thème spécifique – ThSpe. Ensuite, nous présentons comment se répartissent les collocations entre les différents types d'INIT et ce qui suit, que ce soit un Thème topical ou spécifique. Cette présentation nous permet de mettre au jour les 'collocations préférées'.

Les mesures de ces répartitions au niveau du corpus entier sont indiquées en nombre d'occurrences et en pourcentage. Les variations entre sous-corpus sont essentiellement présentées sous forme de graphiques des écarts réduits (voir VII.4).

Ce chapitre nous permet de dresser de façon générale les types d'éléments les plus présents en position initiale et de mettre au jour les grandes particularités des différents sous-corpus. Ces particularités constituent le modèle théorique pour chaque sous-corpus. Ainsi, lorsque nous observons les variations selon les positions textuelles (chapitre IX) et selon certaines configurations particulières (chapitre X), nous posons l'hypothèse nulle que ces particularités sont homogènes au sein d'un sous-corpus, quelle que soit la position textuelle de la phrase et en dépit de l'influence des éléments entre eux.

VIII.1. Patrons de position initiale : Connect*INIT*ThTop/ThSpe

Selon la caractérisation présentée dans le chapitre précédent, la position initiale peut présenter un connecteur, un (ou plusieurs) élément(s) détaché(s) et un Thème topical ou spécifique[155]. Le tableau VIII.1 montre la fréquence de chacun de ces éléments dans notre corpus. La composition la plus fréquente est finalement la plus 'simple' : ni connecteur ni INIT et un ThTop (plus de 50% des phrases, voir tableau VIII.2).

	Nb de phrase	%
Avec ThTop	19 360	*83,4*
Avec ThSpe	3 857	*16,6*
Avec INIT	7 022	*30,2*
Sans INIT	16 196	*69,8*
Avec Connect	2 220	*9,6*
Sans Connect	20 997	*90,4*

Tableau VIII.1 : Éléments en position initiale

Toutefois, la présence d'un élément détaché en initial ou d'une construction spéciale ne correspond pas à ce que l'on pourrait appeler un cas rare : presque une phrase sur trois commence par un INIT et une phrase sur six est une construction spéciale. La présence d'un connecteur en initiale de phrase est moins fréquente, autour des 10%. Les différents éléments délimités en position initiale ne sont donc pas des phénomènes 'rares' et nous disposons de suffisamment de données pour appliquer nos analyses statistiques.

L'association des trois types d'éléments (ThTop/ThSpe, INIT, Connect) aboutit à huit patrons de position initiale :

Connect+INIT+ThTop	*Ainsi, d'un point de vue législatif cette fois, l'administration Bush a fait présenter une loi de lutte contre le terrorisme.[GEOPO_1]*
Connect+ThTop	*De même, les autorités "gouvernementales" ont, [...], refusé d'attribuer aux centrales nucléaires la même protection fédérale que celle dont bénéficient dorénavant les aéroports.[GEOPO_1]*
INIT+ThTop	*Pendant ce temps, les craintes d'attentats contre d'autres types de cibles civiles se multipliaient.[GEOPO_1]*
ThTop seul	*La Manche est devenue le passage maritime le plus encombré du monde [ATLAS_1]*
Connect+INIT+ThSpe	*Néanmoins, malgré cette situation, il semble que le parti du Président doive modérer ses ambitions et gérer un grand nombre de contraintes.[GEOPO_8]*
Connect+ThSpe	*Et ce n'est pas un hasard si l'inventeur de ce personnage, de son nom et de sa légende, fut un religieux mercenaire de la très catholique Espagne : le dramaturge Tirso de Molina. [PEOPL_1]*
INIT+ThSpe	*Selon les conclusions de l'étude, il est possible que la NIMA ait utilisé les images Ikonos pour dresser des cartes précises et à jour des régions en cause.[GEOPO_2]*
ThSpe seul	*C'est d'autant plus vrai que ces mesures ne sont pas précisément des décisions sur lesquelles l'administration se serait engagée à revenir.[GEOPO_1]*

155 Nous rappelons que ThTop et ThSpe sont deux éléments exclusifs.

Ces huit patrons ne se distribuent pas équitablement au sein de notre corpus et de nos sous-corpus. Le patron ThTop seul peut être considéré comme le plus 'naturel'. Il correspond à la construction canonique des phrases en français {sujet+prédicat} où le sujet n'est pas un thème 'vide' comme dans les constructions spéciales (ThSpe).

	corpus		%		
Patrons (ordre décroissant)	**Nb de phrase**	**%**	**ATLAS**	**GEOPO**	**PEOPL**
ThTop	12 097	*52,1*	*60,8*	*47,7*	*48*
INIT+ThTop	5 445	*23,5*	*22,2*	*28,1*	*19,9*
ThSpe	2 450	*10,6*	*8,8*	*9*	*13,9*
Connect+ThTop	1 359	*5,9*	*2,9*	*6,8*	*7,8*
INIT+ThSpe	1 005	*4,3*	*3,5*	*4,5*	*5*
Connect+INIT+ThTop	459	*2*	*1,1*	*2,2*	*2,6*
Connect+ThSpe	290	*1,2*	*0,5*	*1,2*	*2,1*
Connect+INIT+ThSpe	112	*0,5*	0,2	*0,4*	*0,9*
	23 217	*100*	7 592	7 901	7 724

Tableau VIII.2 : Répartition des huit patrons en position initiale

D'une façon générale, les phrases présentant un ThTop sont majoritaires (83,4%) par rapport aux ThSpe (16,6%). Ce rapport reste relativement stable selon qu'il y a en début de phrase un INIT, un connecteur ou rien du tout. Ce n'est que dans les situations où il y a un connecteur et un INIT qu'il y a une légère remontée des ThSpe. Mais le test de l'écart réduit, nous montre que cette variation n'est pas significative (z=+1,79). Rappelons que la catégorie des ThSpe rassemble des constructions aux fonctions très diverses, ce qui peut souvent expliquer la faiblesse des variations observées.

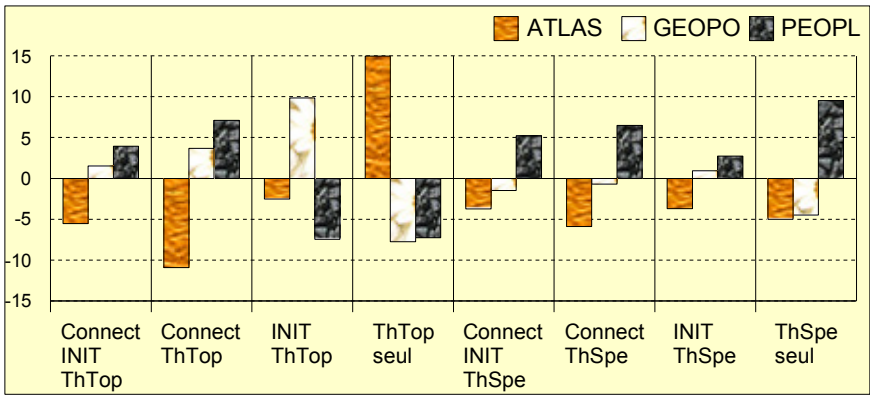
La relative stabilité du rapport ThTop/ThSpe se retrouve plus ou moins au travers des différents sous-corpus : ATLAS et GEOPO suivent le mouvement général, à l'inverse de PEOPL qui montre de plus grandes spécificités dans l'emploi des ThTop/ThSpe. Pour n'importe quel type de patron, PEOPL montre significativement plus de ThSpe (voir les écarts réduits du tableau VIII.3). Mais en plus de cette grande proportion de ThSpe, PEOPL présente une plus grande disparité dans la répartition des ThTop/ThSpe selon ce qui précède. Nous observons dans tous les sous-corpus une grande faiblesse de Thèmes topicaux précédés d'un connecteur et d'un élément détaché (Connect+INIT+ThTop). À l'inverse, les patrons avec connecteur et élément détaché sont ceux qui présentent le plus de Thème spécifique. Dans PEOPL, cette proportion est accentuée, ce qui signifie une plus grande association dans ce sous-corpus entre les ThSpe et la présence conjointe de Connect et d'INIT.

La forte majorité de phrases avec ThTop seul peut également s'expliquer par la fonction discursive généralement associée aux constructions spéciales. En effet, les Thème spécifiques sont généralement associés à des structures informationelles susceptibles d'indiquer un déplacement ou une rupture. De façon évidente, les stratégies de déplacement et de rupture sont moins fréquentes que celles de continuité. La proportion de Thèmes topicaux *vs.* spécifiques illustre donc l'idée de texture : pour qu'un texte soit cohérent, il faut qu'il y ait une forte continuité entre ses unités. Or cette texture est plus fortement construite autour des Thèmes topicaux qui permettent les progressions thématiques qu'autour des Thèmes spécifiques qui déplacent l'expression des éléments idéationnels en position Rhème. D'un point de vue cognitif, le moyen le plus économique pour établir cette continuité est de suivre le patron : position préverbale = thème simple = sujet grammatical = information donnée.

ATLAS est le corpus qui présente le plus de patron ThTop seul : six phrases sur dix présentent ce type de patron. En comparant les sous-corpus entre eux, nous obtenons un écart réduit *z*(ATLAS_ThTop)=+15,2[156]! – le plus fort écart réduit observé (voir tableau VIII.3).

	ATLAS	GEOPO	PEOPL	∑ des écarts
Connect+INIT+ThTop	-5,5	+1,5	+4	**11,0**
Connect+ThTop	-10,9	+3,7	+7,1	**21,7**
INIT+ThTop	-2,5	+9,9	-7,5	**19,9**
ThTop seul	+15,2	-7,8	-7,3	**30,3**
Connect+INIT+ThSpe	-3,8	-1,5	+5,2	**10,5**
Connect+ThSpe	-5,9	-0,7	+6,5	**13,1**
INIT+ThSpe	-3,7	+0,9	+2,7	**7,3**
ThSpe seul	-5,0	-4,5	+9,5	**19,0**
∑ des écarts	**52,5**	**30,5**	**49,8**	
(Connct)+(INIT)+ThSpe	**-8,6**	**-3,7**	**+12,3**	

Tableau VIII.3 : Variations selon les sous-corpus des patrons de position initiale



Graphique VIII.1 : Variations selon les sous-corpus des patrons de position initiale

La comparaison de la répartition des patrons dans les différents sous-corpus nous montre qu'aucun patron n'est neutre (*i.e.* insensible au type de texte), comme le montrent le tableau VIII.3 et le graphique VIII.1 équivalent[157]. Certains patrons montrent une sensibilité très forte à la variation entre sous-corpus. Les Thèmes topicaux et spécifiques semblent bien s'opposer avec une exclusivité pour Thèmes topicaux seuls chez ATLAS et pour Thèmes spécifiques seuls chez PEOPL. GEOPO semble particulièrement user des patrons INIT+ThTop. Les constructions présentant un Thème spécifique semblent moins dépendantes du sous-corpus que celles présentant un Thème topical.

La probabilité d'avoir un élément détaché en position initiale est de 30% (7 022 phrases présentent au moins un INIT1), c'est-à-dire deux fois plus grande que celle d'avoir une construction spéciale. Cette probabilité ne semble pas être influencée par la nature topicale ou spécifique du thème de la phrase puisqu'elle reste à peu près identique pour les phrases avec ThTop et ThSpe (30,5% et 29%, *cf.* tableau VIII.4). En revanche, la présence d'un connecteur diminue d'environ 5% la probabilité d'avoir un INIT, alors que sans connecteur en initiale, cette proportion se situe autour de 30%. L'écart observé lorsqu'il y a un INIT dans les phrases commençant par un connecteur affiche un score négatif *z*=-4,6, ce qui signifie que la diminution observée est significative.

156 Nous rappelons qu'un écart réduit est dit significatif au delà de +/- 2,5 (voir VI.4.2).

157 Pour cette première mention des écarts réduits, nous avons mis en parallèle la valeur des écarts mesurés et la représentation graphique de ces écarts. Dans la suite de nos analyses, seuls les graphiques seront présentés.

	ThTop	ThSpe	Sans Connect	Avec Connect
Phrases avec INIT	30,5	*29,0*	30,7	*25,7*
	ThTop	**ThSpe**	**Sans INIT**	**Avec INIT**
Phrases avec Connect	*9,4*	*10,4*	*10,2*	*8,1*

Tableau VIII.4 : Proportion d'INIT et de Connect selon le type d'élément présent en position initiale

La probabilité d'avoir un connecteur est de moins de 10% : 2 220 phrases présentent un connecteur en initiale. Les phrases présentant un ThSpe ou ne contenant pas d'INIT affichent la même probabilité (voir tableau VIII.4). Au contraire, les phrases avec ThTop ou avec INIT affichent moins de Connect. En appliquant le test de l'écart réduit, nous observons que seules les phrases avec INIT montrent significativement moins de Connect (z=-4,6)[158].

Il ressort de cette première série de mesures que la position initiale n'est absolument pas insensible au type de texte. Chaque sous-corpus présente une répartition particulière des patrons. ATLAS et PEOPL présentent des répartitions opposées et particulièrement significatives. À l'inverse, GEOPO est moins spécifique et donc plus proche du modèle théorique. Ses écarts significatifs montrent que les Thèmes topicaux dans GEOPO sont généralement introduits par quelque chose (un connecteur ou un élément détaché). Les patrons avec connecteurs se trouvent préférentiellement dans PEOPL, ainsi que ceux avec ThSpe.

Les parties suivantes s'intéressent en détail aux différents éléments de la position initiale, nous y retrouvons ces tendances générales auxquelles s'ajoutent des variations relatives à la nature de ces différents éléments. Nous nous intéressons plus particulièrement à la nature des éléments détachés et des Thèmes topicaux. Ces deux éléments jouent un rôle très important dans la construction de la représentation mentale. C'est au niveau des éléments détachés et des Thèmes topicaux que vont s'exprimer les *settings* constitutifs du *fond* et les entités *figures* du *text-world*.

À l'inverse de ces deux éléments, les connecteurs 'purs' et les constructions à Thème spécifique correspondent à des catégories construites par une définition négative. Les Connect ont été annotés de la sorte pour leur absence de contenu idéationnel et les ThSpe pour leur structure syntaxique 'non-canonique'. Cette définition négative implique une catégorie assez hétérogène : les différents connecteurs 'purs' expriment quantité de relation de discours différentes et les différentes constructions spéciales permettent des structures informationnelles très diverses voir parfois opposée. L'absence d'une définition positive et l'hétérogénéité constitutive alliées à la faible fréquence des ces deux éléments nous poussent à mettre un peu de côté leur analyse. Nous exposons toutefois quelques observations effectuées sur les variations entre sous-corpus dans les deux dernières sections de ce chapitre.

VIII.2. Nature des Thèmes Topicaux – ThTop

83,4% des phrases présentent un Thème topical et non une construction spéciale. Ces 19 360 phrases présentent dans 7% des cas un Connect (sans INIT), dans 28% un INIT (sans Connect) et dans 2,5% un Connect et un INIT. Les Thèmes topicaux apparaissent donc majoritairement seuls (dans 62,5% des cas). Deux catégories de mesures nous intéressent particulièrement au niveau des ThTop : leur forme et leur capacité à indiquer une co-référence.

158 Les phrases avec ThTop affichent un écart réduit de z=-1,9 lorsqu'il y a présence d'un Connect en initiale.

VIII.2.1. Répartition des différents types de Thème Topical

La forme prédominante des Thèmes topicaux est le SN défini (SNdef) qui représente plus de la moitié de tous les ThTop (52,4%) et plus de 40% de toutes les phrases (43,6%). La seconde position est occupée par les PRO3, loin derrière, autour de 10% (11,5% des ThTop ou 9,6% des phrases).

	Corpus entier			%					
	Nb	%'	%	ATLAS		GEOPO		PEOPL	
SNdef	10 143	43,7	52,4	59,7		59,5		36,5	
PRO3	2 221	9,6	11,5	8,4		8,2		18,5	
Ttautre	1 675	7,2	8,7	6,8		7,7		11,8	
NP	1 592	6,9	8,2	4,7		4,8		15,9	
SNdem	1 538	6,6	7,9	9,2		8,4		6,1	
SNindef	1 447	6,2	7,5	8,2		7,8		6,4	
SNposs	435	1,9	2,2	1,5		1,8		3,5	
PROdemo	309	1,3	1,6	1,5		1,9		1,3	
			100	100		100		100	
ThTop	19 360		83,4	6 607	87,1	6 710	84,9	6 037	78,2
Nb Phr.	23 217			7 592		7 901		7 724	

* la colonne % affiche la proportion pour les ThTop et la colonne %' affiche la proportion pour le corpus entier.

Tableau VIII.5 : Répartition des différents types de ThTop dans notre corpus

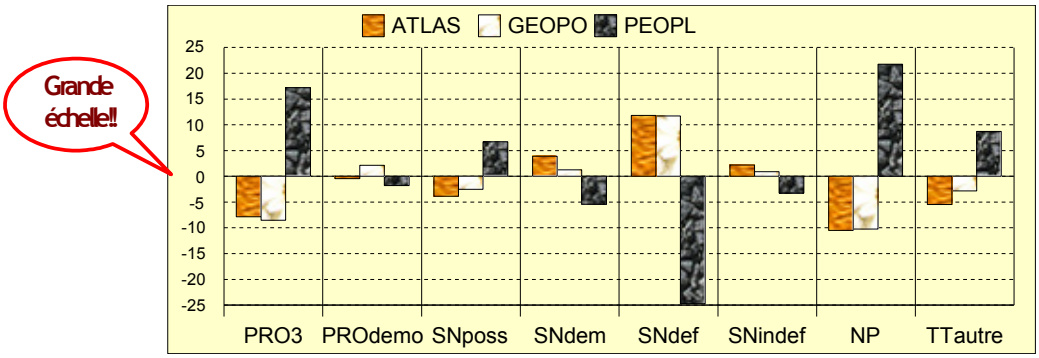
La prédominance du SNdef se retrouve dans tous les sous-corpus, même dans PEOPL qui pourtant affiche un taux élevé de noms propres – NP. Presque 16% des ThTop de PEOPL ont un sujet de forme NP (ce qui équivaut à 12,4% de toutes les phrases), contre moins de 5% chez ATLAS ou GEOPO. PEOPL montre un écart réduit *z(SNdef)* inférieur à -24 et un écart réduit *z(NP)* supérieur à +21 (voir graphique VIII.2). Nous assistons à une sorte de permutation : PEOPL utilise un nom propre là où ATLAS et GEOPO utilisent une description définie. Cette idée sera revue particulièrement lorsque l'on observe les variation selon la position textuelle (IX.2).

La thématique des textes de PEOPL (portrait d'une célébrité) explique assurément cette forte association PEOPL/NP. Ce sous-corpus présente le plus de particularités : toutes les formes excepté les pronoms démonstratifs – PROdemo – présentent un z > 2,5. PEOPL est également celui qui présente les écarts les plus grands, *i.e.* des particularités très fortes, ce qui se voit à la longueur des 'bâtons' relatifs à PEOPL (les plus sombres). Nous remarquons également que PEOPL est toujours significativement à l'opposé de ATLAS et GEOPO qui présentent des comportements similaires, il y a donc association ou dissociation exclusive avec PEOPL.

Les pronoms de 3e personnes – PRO3, qui représentent 11,5% des ThTop, passent quasiment du simple au double entre ATLAS ou GEOPO (environ 8% des ThTop) et PEOPL (18,5%). Cette déviance est significative et va de pair avec le caractère mono-référentiel de ce sous-corpus (VII.1.1.b). La force de l'écart positif chez PEOPL entraîne un écart négatif chez les deux autres sous-corpus. Cet écart négatif conforte nos intuitions quant au caractère pluri-référentiel de ces deux sous-corpus. Nous retrouvons dans de moindres mesures le même schéma d'association exclusive entre les SN possessifs et PEOPL.

La proportion de SN démonstratifs est assez faible. Ce qui se retrouve dans de nombreuses études sur l'alternance SNdem/SNdef dans les chaînes de référence (voir V.4.3.c). Les SNdem et SNdef ne présentent pas le même comportement. Alors que pour les SNdef, ATLAS et GEOPO sont à égalité, pour les SNdem, ATLAS et GEOPO

se distinguent. Chaque sous-corpus présente un écart différent, ce qui suppose une forte dépendance entre type de texte et fréquence des SNdem.



Les écarts réduits sont mesurés par rapport à la totalité des ThSpe (et non la totalité des phrases)

Graphique VIII.2 : Écarts significatifs selon les sous-corpus dans la répartition des ThTop/ThSpe

La graphique VIII.2 fait apparaître six types de Thèmes topicaux présentant des écarts significatifs. Seuls les PROdemo et les SNindef ne semblent pas très influencés par le sous-corpus. Les descriptions définies sujets apparaissent comme étant les plus sensibles au genre. Les noms propres affichent le schéma inverse dans les mêmes mesures. Viennent ensuite les PRO3. Nous nous intéressons particulièrement aux SNdef, PRO3 et NP qui semblent constituer des indices pertinents pour l'étude des relation de (dis)continuité. (cf. V.4.3). Nous verrons que les écarts observés ici se retrouvent de manière encore plus significative lorsque l'on considère le facteur de la position textuelle. Ces trois types de Thème topical font l'objet d'analyses particulières dans le chapitre X.

VIII.2.2. De la co-référence en Thème topical

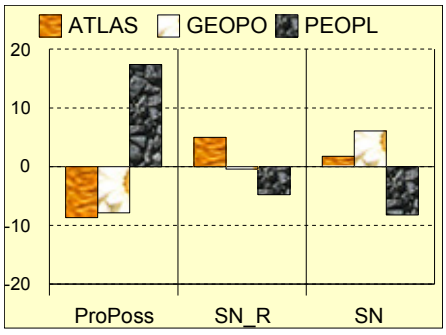
Trois indices nous informent sur la possibilité de co-référence en Thème topical : (1) la présence d'un pronom ou d'un SN possessif; (2) la présence d'une reprise lexicale[159] ou (3) la réduction d'une description. Les données présentées dans le tableau VIII.5 sont reprises ici, dans le tableau VIII.6, selon un classement différent et en distinguant pour chaque type de ThTop (ni pronominal ni possessif) le fait qu'il y a ou non reprise (représenté par un « _R » à la fin de la forme concernée). La mesure du nombre de descriptions courtes (voir VII.2.2.d) se fait dans un deuxième temps, une description courte pouvant également présenter une reprise.

159 Notre définition opérationnelle de la notion de reprise est donnée en VII.2.2.b.

		Corpus entier			ATLAS		GEOPO		PEOPL	
		Nb	%		Nb	%	Nb	%	Nb	%
ProPoss	PRO3	2 221	11,5		555	**	547		1 119	
	PROdemo	309	1,6		101	Expression erronée **	129		79	
	SNposs	435	2,2	15	102		120	12	213	23,5
SN_R	SNdem_R	641	3,3		280		234		127	
	SNdef_R	3 573	18,5		1 516		1 394		663	
	NP_R	792	4,1	31	126	34,5	101	31	565	27,5
	ThTop_R autre	603	3,1		177		200		226	
	SNindef_R	413	2,1		185		138		90	
SN	SNdem	897	4,6		331		327		239	
	SNdef	6 570	34		2 428		2 601		1 541	
	NP	800	4,1	54	183	52,5	221	57	396	49
	ThTop autre	1 072	5,5		270		315		487	
	SNindef	1 034	5,3		356		383		295	
ThTop		19 360		100	6 610	100	6 710	100	6 040	100

Tableau VIII.6 : Répartition des différents types de ThTop en prenant en compte les reprises (_R)

15,3% des ThTop sont des pronoms ou des SN possessifs – ProPoss. En ajoutant les 31,5% des ThTop qui constituent une reprise d'un nom déjà mentionné, nous obtenons 47% de ThTop présentant de façon relativement probable un lien avec le discours précédent. Ces trois proportions montrent de fortes variations selon les sous-corpus, et chaque sous-corpus semble préférer un des groupes ainsi délimités, ce qu'illustre le graphique VIII.3.



Graphique VIII.3: Co-référence pronominale et lexicale en ThTop selon les sous-corpus

PEOPL présente deux fois plus de pronoms et de SNposs en ThTop que les deux autres sous-corpus, ce qui se mesure par un écart réduit de +17 pour ce groupe : les ProPoss (PEOPL présente 1 411 ProPoss contre 758 dans ATLAS et 796 dans GEOPO). La force de cette association entraîne un usage moindre des descriptions (avec ou sans reprise). L'exemple VIII.3 p.208, utilisé pour montrer l'usage des noms propres répétés – NP_R – par PEOPL montre un exemple des stratégies co-référentielles dans PEOPL.

Les SN(_R), *i.e.* les SN avec ou sans reprise, affichent des schémas d'écart dispersés : ceux avec reprise chez ATLAS et ceux sans reprises chez GEOPO. L'usage important des reprises lexicales dans ATLAS peut être lié au fait que ce sous-corpus offre une palette relativement faible de noms récurrents (surtout comparé à GEOPO, voir partie VII.1.1.b). Les syntagmes de type *les départements X*, *les régions X*, etc. sont légion dans ATLAS, qui, de plus, recourt souvent à des progressions thématiques à thèmes dérivés[160]. Ces progressions peuvent se réaliser par l'utilisation de différentes descriptions où la tête lexicale reprend l'hyperthème (ou la tête de l'hyperthème) et l'expansion caractérise un sous-thème (exemple VIII.1).

(VIII.1) ***Les établissements*** *sous la tutelle du ministères des Affaires sociales, de la Santé et de la Ville hébergent des enfants présentant généralement des troubles psychiatriques, relationnels ou physiques moyens ou graves, ou encore*

160 Les modèles des progressions thématiques de Daneš est expliqué en III.3.1.

des enfants placés dans le cadre de l'aide sociale à l'enfance. ***Les établissements médicaux*** *accueillent près de 20 000 jeunes, dont un peu moins de la moitié sont scolarisés.* ***Les établissements médico-éducatifs*** *ont en charge surtout des déficients mentaux, environ 110 000, dont les trois quarts sont scolarisés.* ***Les établissements des services de la Santé*** *scolarisent ainsi 30% des enfants de l'enseignement spécial, avec des variations importantes d'un département à l'autre: plus de la moitié dans les départements toulousains, mais moins du quart dans le Nord et l'Île-de-France.[ATLAS_2]*

GEOPO offre au contraire des SN sémantiquement plus divers et des progressions thématiques de type très hétérogène (exemple VIII.2), offrant un mélange de progressions constantes (entre les phrases 2 et 3 de l'exemple VIII.2), de progressions linéaires (entre les phrases 4 et 5 de l'exemple VIII.2) et d'absence de progression thématique (la plupart des relations intraphrastiques de l'exemple ne se basent pas sur des progressions thématiques).

(VIII.2) ***La fragilité légale de la lutte contre le terrorisme****, en interne, est encore plus flagrante à l'extérieur du territoire.* ***Les prisonniers faits en Afghanistan*** *sont déclarés " ennemis combattants " (enemy combattant), une catégorie inconnue du droit international.* ***Ils*** *ont été transférés sur la base (américaine depuis 1903) de Guantanamo dès février 2002.* ***Le Secrétaire à la Défense, Donald Rumsfeld,*** *a officiellement reconnu que les prisonniers de Guantanamo sont là pour une durée illimitée.* ***Ils*** *n'ont bien sûr pas d'avocats, et, selon le décret présidentiel de novembre 2001, il serait possible de juger ces hommes par des tribunaux militaires d'exception. Jusqu'à présent,* ***seule la Croix Rouge Internationale (CICR)*** *a pu leur rendre visite.* ***Les conventions de Genève*** *ne sont donc que très partiellement appliquées, constat établi par un rapport d'Amnesty International en mars 2002.* ***Les libertés que s'autorisent les autorités américaines*** *sont encore plus claires au niveau de la collaboration internationale : certains observateurs soulignent que les Etats-Unis utiliseraient ainsi les règles d'extradition pour faciliter les interrogatoires. Ainsi,* ***une personne arrêtée en Indonésie*** *peut, sur demande des Etats-Unis, être transférée en Egypte, pour subir un interrogatoire plus " adapté".* ***Les chiffres sur cette pratique*** *ne sont pas connus. Sur ce terrain,* ***les oppositions à l'attitude gouvernementale*** *sont assez clairsemées : peu de gens semblent vouloir revenir sur le statut fait aux prisonniers capturés en Afghanistan. Jusqu'à présent,* ***seul un Juge fédéral de la Quatrième Cour d'Appel (Circuit Court)****, pourtant conservateur de réputation, John H. Wilkinson, a contesté la capacité légale du gouvernement à désigner de sa seule autorité les " ennemis combattants " . Mais par contre, dès que cette politique implique des citoyens américains,* ***la situation*** *devient plus délicate pour l'Etat fédéral. En d'autres termes, lorsque les conditions de détention des non-Américains s'étendent aux citoyens Américains eux-mêmes, alors il y a un réel débat, qui, sans forcément mobiliser l'opinion, pousse au moins les autorités à tenter de justifier leur attitude. Depuis la mise en oeuvre de ces textes, cela s'est produit à plusieurs reprises.[GEOPO_22]*

Rappelons que ces variations se calculent en comparant les sous-corpus les uns aux autres. Cela ne signifie aucunement que GEOPO ne présente pas de reprises. Il présente tout de même 43% de Thèmes topicaux pronominaux, possessifs ou avec reprise, ce qui représente presque 3000 phrases. Les variations observées signifient simplement que GEOPO présente une distribution importante de SN sans reprise lexicale par rapport à la moyenne générale, représentée ici par ATLAS. Cela peut également suggérer que ce sous-corpus utilise davantage des procédés de co-référence indirecte où il n'y a pas de reprise lexicale, mais des anaphores résumantes[161] (« *cette pratique* » et «la situation » dans l'exemple VIII.2) ou des liens de cohésion lexicale (entre « *la Croix Rouge Internationale* » et « *les conventions de Genève* »)

Seuls les ProPoss affichent un schéma d'association exclusive avec PEOPL. Les données du tableau VIII.6 montrent que près de 50% des pronoms et SNposs se situent dans ce sous-corpus. Dans cette catégorie, les pronoms démonstratifs affichent un comportement différent : une répartition des écarts inversée avec une fréquence plus importante chez GEOPO. Cependant, ces écarts ne sont pas significatifs, ce qui fait qu'avec ou sans les PROdemo PEOPL représente 50% des pronoms et possessifs du corpus.

De façon générale, nous voyons des techniques de continuité référentielles réellement différentes entre les trois sous-corpus, ce que les exemples (VIII.1), (VIII.2) et (VIII.3) montrent particulièrement bien. La suite de cette partie confirme cette différence.

161 Conte (1996) parle d'« anaphore encapsulante » ou « encapsulation ». Legallois (2006) s'intéresse également au fonctionnement des noms sous-spécifiés tels que *situation, cas, fait, problème,* etc.

Chaque type de SN affiche entre 30% et 50% de reprise. Nous rappelons que les SN à reprise correspondent aux SN dont la tête correspond à un mot déjà présent dans la section en cours. Comme le montre le tableau VIII.7, cette proportion est très dépendante du sous-corpus considéré.

	Proportion à présenter une reprise _R (%)			
	Corpus entier	ATLAS	GEOPO	PEOPL
NP	50	41	31	59
SNdem	42	46	42	35
autres	36	40	39	32
SNdef	35	38	35	30
SNindef	29	34	26	23

Tableau VIII.7 : Proportion des différents type de ThTop à présenter une reprise (_R)

Les noms propres apparaissent là encore comme les plus sensibles au sous-corpus (voir la comparaison des écarts par sous-corpus dans le graphique VIII.4 infra). Nous avons déjà remarqué l'association établie entre PEOPL et ce type de Thème topical, le caractère mono-référentiel et la thématique des textes de ce sous-corpus expliquant certainement cette association. Dans PEOPL, presque 60% des noms propres présentent une reprise (voir Schnedecker 2005 et partie III.2.3) alors qu'en moyenne, nous restons à 50%. L'exemple VIII.3 donne une idée de la manière dont les textes de ce sous-corpus utilisent les NP_R en alternance aux PRO3 (tous les ThTop de cet extrait ont été mis en gras).

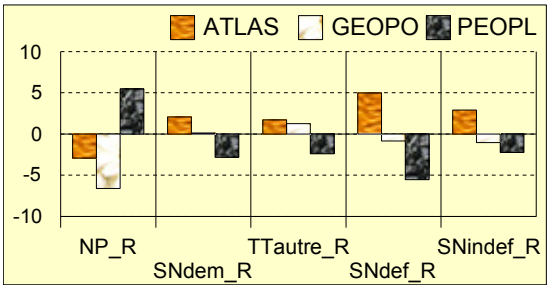
(VIII.3) ***Léonard [de Vinci]**, quittant l'atelier de Verrocchio, a pu être quelque temps au service de Laurent de Médicis ; [...]. **Une note du Codex atlanticus**, [...], contient peut-être, plutôt qu'une allusion aux médecins, une sorte de bilan opposant à la générosité de Laurent l'indifférence de son neveu Léon X (la note est de 1515).*
*En 1482-1483, **Léonard** est au service de Ludovic le More qui vient de s'emparer du duché de Milan (1480). **Il** devient le grand animateur de la cour. Après 1499, **il** cherche un autre protecteur princier: [...] ; revenu à Florence, il quitte sa ville natale à plusieurs reprises. **Il** intéresse César Borgia [...]. Aux yeux des princes français comme à ceux de César Borgia, **Léonard**, si célèbre qu'il soit comme peintre, compte pour ses autres capacités. On est frappé aussi par la facilité avec laquelle l'artiste-ingénieur passe du service d'un protecteur à celui de son adversaire. **Il** revient à Milan avec les princes français qui ont chassé Ludovic ; à la fin de 1504, il est à Piombino, auprès de Jacoppo IV d'Appiano, qui, l'année précédente, avait été chassé par César Borgia, le patron de Léonard. **Les grands esprits** n'ont pas de camp. **Léonard** appartient à qui se l'attache et lui laisse un loisir pour l'étude. **Paul Jove** a été frappé de ses capacités comme organisateur de fêtes, musicien, etc., et conclut que ces aptitudes "l'ont rendu cher à tous les princes qui l'ont connu...". En dehors des décors de théâtre ou de parade, **Léonard** a composé des rébus, constitué des recueils de devinettes et de fables, rédigé des devises, des imprese .[PEOPL_11]*

Cet exemple fait clairement apparaître l'importance des noms propres répétés dans PEOPL. Nous verrons dans les chapitres suivantes dans quelle position textuelle et après quel type d'INIT se situe préférentiellement ces NP_R. Selon l'idée qu'une redénomination intervient pour accompagner un déplacement (voir V.4.3.b), la position textuelle préférée sera certainement les premières phrases de paragraphe ou après un adverbial circonstanciel (ce qui est observé plusieurs fois dans l'exemple VIII.3)[162].

Sans prendre en compte la forte proportion des NP_R propre à PEOPL, les SNdem affichent le plus haut taux de reprise avec 42%. Ce sont en effet ce type de Thème topical qui présentent le plus de reprises dans ATLAS et GEOPO. Cette proportion est cohérente avec le rôle fortement co-référentiel du déterminant démonstratif (voir V.4.3.c). Cependant, le nombre d'occurrences de SNdem reste relativement faible (1 538 occurrences en tout, tableau VIII.5 p.204).

162 Voir à ce sujet les parties IX.2.3.b et X.4.1.b.

De façon générale, les Thèmes topicaux présentent plus de reprises dans ATLAS (excepté les pour les noms propres évidemment). Cependant, seules les variations observées au niveau des SN définis et indéfinis sont significatives, comme l'indique le graphique VIII.4. À l'inverse, PEOPL présente un écart négatif pour tous les SN autres que NP.



Graphique VIII.4 : Variations de la proportion de reprise selon le type de SN et le sous-corpus

De nombreux SNindef montrent une reprise, ce qui est contraire au sens instructionnel généralement associé aux déterminants indéfinis (voir V.4.3.c). En observant les 413 cas de SNindef_R (dont près de la moitié se situent dans ATLAS), on remarque rapidement que les têtes des SNindef_R correspondent à des noms génériques (*situation*, *cas*, etc.), des noms de localisation géographique générale (*région, département*, *zone*, *circonscription,* etc.) ou encore aux noms récurrents. Ces noms constituent généralement le premier nom dans des constructions N de N (moins généralement dans PEOPL). Par exemple :

> *Un **nombre** non négligeable de communes [ATLAS_1]; 11 **%** des enfants de CP [ATLAS_2]; Un second **cas** de figure [ATLAS_3]; cinq **départements** de l' Ouest [ATLAS_3]; Certains **faits** [GEOPO_6]; Aucun **événement** de sa vie privée [PEOPL_16]; Chaque **chose** [PEOPL_19].*

Nous voyons à la lecture de ces exemples qu'il ne s'agit pas réellement de reprises au sens coréférentiel du terme. Nous ne donnons donc pas de poids coréférentiel à ces reprises en SN indéfini. Le taux élevé de SNindef_R dans ATLAS ne fausse pas les variations observées dans le graphique VIII.4. Si l'on refait les calculs en ignorant les SNindef_R, aucun changement significatif n'est observé.

	Proportion à constituer une description courte				
	Corpus entier **Nb.**	**%**	***ATLAS***	***GEOPO***	***PEOPL***
SNdem courts	1 063	*69*	*67*	*72,5*	*70*
Tout SNdem (Nb.)	1 538		611	561	366
SNdef courts	4 111	*40,5*	*36,5*	*40,5*	*48*
Tout SNdef (Nb.)	10 143		3 944	3 995	2 204

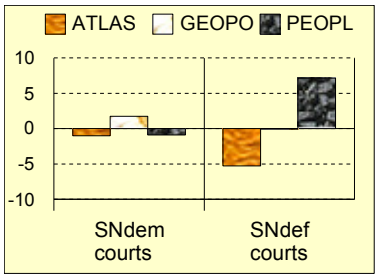


Tableau VIII.8 : Distribution des SNdef et SNdem courts (moins de 3 blancs)

La proportion des SNdef et SNdem à constituer des descriptions réduites est également un indice de coréférence (voir V.4.3.c). Le tableau VIII.8 présente la proportion pour chacun à constituer une description courte. « Description courte » car notre identification des descriptions réduites se résume simplement en un repérage des SN présentant au maximum 2 espaces typographiques. Nous estimons cependant qu'une description 'courte' est fortement susceptible de coréférer (voir VII.2.2.d).

Comme on le voit, les SN démonstratifs sont majoritairement des descriptions courtes (plus de deux tiers des SNdem présentent moins de trois espaces typographiques). Cette proportion est égale pour tous les sous-corpus (aucun écart significatif n'est mesuré). Nous verrons au niveau des variations par positions textuelles si cette égalité reste entre les trois sous-corpus. Au niveau des SNdef courts, les données sont bien différentes. Moins de la moitié des SNdef sont courts. Cette proportion est dispersée selon les sous-corpus. La proportion des 40% est en fait celle observée dans GEOPO (qui ne montre pas d'écart réduit significatif), alors que ATLAS présente 36,5% de SNdef courts contre 48% chez PEOPL. Ce dernier résultat est informatif au regard de la différence de comportement entre SNdef avec reprise et SNdef courts. Alors que ATLAS montrait plus de SNdef_R, il montre moins de SNdef courts. PEOPL montre les proportions inverses : moins de SNdef_R et plus de Sndef courts. L'extrait suivant montre un exemple de SNdef courts et sans reprise dans PEOPL.

(VIII.4) *À l'évidence cartésienne, jugée trop subjective, on préférera la certitude expérimentale. Il en résulte que le philosophe doit s'inspirer de la science ou, mieux, des sciences. Or* ***les sciences*** *ne s'éclairent que par des théories qui dépassent les sens.* ***Ces théories*** *ne sont pas celles qu'imaginent les savants pour faire progresser leurs disciplines.* ***Elles*** *sont pour le philosophe une recherche de principes. C'est revenir à la métaphysique. [PEOPL_20]*

On peut comparer cet exemple avec l'exemple VIII.5 extrait de ATLAS où l'on observe plusieurs SNdef_R longs.

(VIII.5) ***Leur dynamique démographique*** *a été très différente sur une génération. Globalement,* ***les grandes villes françaises*** *ont en effet connu une croissance vive de 1962 à 1990, avec un gain de plus de deux millions et demi d'habitants, soit près d'un quart de population supplémentaire. A elle seule,* ***l'agglomération parisienne*** *a gagné 1 735 000 habitants, soit 68 % de la hausse, ce qui est conforme à son poids actuel dans l'ensemble de la population résidant dans les grandes villes. Par contre,* ***les grandes agglomérations anglaises*** *ont vu leur population diminuer de 9 % en trente ans. Alors qu'en France, sur trente ans, ce sont les grandes villes qui ont connu la croissance la plus vive (+ 24 %), en Angleterre, ce sont au contraire celles-là qui ont enregistré la plus faible augmentation relative.* ***Les grandes villes françaises*** *n'ont ainsi cessé d'accroître, de façon très régulière d'un recensement à l'autre, leur poids dans la population régionale, tandis qu'en Angleterre, cette part a constamment diminué, y compris dans les années 1980 bien que plus modérément. Cela dit, en dehors de la capitale,* ***les grandes villes britanniques*** *ont connu une même progression que leurs équivalentes françaises. Aussi* ***le poids de l'agglomération capitale*** *dans la population des grandes villes a-t-il régressé de plus de neuf points en Angleterre alors qu'il n'a diminué que d'un point en France.[ATLAS_1]*

La gestion de l'expression co-référentielle par les SNdef semble s'opposer entre ATLAS et PEOPL, et GEOPO se retrouve entre ces deux pôles offrant un mélange de SNdef longs et courts. ATLAS, comme nous l'avions déjà remarqué relativement à son grand usage des SN_R, utilise généralement des progressions thématiques dérivées qui donnent lieu à des descriptions longues et dont la tête lexicale reprend celle de l'hyperthème (exemple VIII.1 p. 206). Nous pouvons également avoir des progressions thématiques constantes en alternance principalement dans l'Atlas Transmanche qui compare du début à la fin les données britanniques et les données françaises. Là encore, les SN sont généralement longs (exemple VIII.5).

VIII.3. Nature des éléments détachés en initiale – INIT

Les éléments détachés en initiale sont caractérisés par le fait qu'ils peuvent orienter le reste du message contenu dans la phrase, le paragraphe, la section, le texte, sans être impliqués dans la construction syntaxique obligatoire de la proposition principale. Ils constituent de ce fait une catégorie d'éléments caractérisés par un certaine indépendance qui leur permet de jouer à un niveau différent du niveau phrastique, et en ce qui nous concerne, au niveau discursif.

30% des phrases présentent au moins un INIT, ce qui représente 7 050 INIT1 et 1 085 INIT2[163]. Nous rappelons que INIT1 et INIT2 ne sont pas deux types différents d'éléments détachés en initiale mais seulement la distinction entre le premier élément détaché et les éléments suivants (voir VII.2.1). 29 titres de sections présentent un INIT1. Aucun titre ne présente deux INIT. Les INIT1 en titre sont essentiellement des adverbiaux circonstanciels comme le montrent les titres suivants :

En Russie, journalisme et littérature [PEOPL]
Après le rapport Rumsfeld, les réorganisations en cours [GEOPO]
Pour lutter contre l'échec scolaire, les zones d'éducation prioritaires (ZEP) [ATLAS]

Nous trouvons également quelques marqueurs d'organisation textuelle :

Introduction : l'espace de Bush est militaire [GEOPO]
Troisième facteur : les statuts juridiques [ATLAS]

Avec INIT1	corpus	ATLAS	GEOPO	PEOPL
Nb phrases	*7 021*	2 047	2 787	2 187
% de phrases	*30,2*	*26,9*	*35,3*	*28,3*
Écart réduit	-	**-6,2**	**+9,7**	**-3,7**
Avec INIT2	**corpus**	**ATLAS**	**GEOPO**	**PEOPL**
Nb phrases	1 085	318	343	424
% de phrases	*4,7*	*4,2*	*4,3*	*5,5*
Écart réduit	-	-2	-1,4	**+3,4**

Tableau VIII.9 : Avec ou sans INIT

Les analyses réalisées sont effectuées sur les INIT hors titre, soit 7 021 INIT1 (et 1 085 INIT2). Pour chaque sous-corpus, entre 25% et 35% des phrases présentent un INIT1. Cette proportion varie significativement pour chaque sous-corpus : GEOPO présente plus d'INIT1 (*z*=+9,7) que ATLAS ou PEOPL qui affichent tous deux un écart négatif. Au niveau des INIT2, seul PEOPL montre un écart significatif positif.

VIII.3.1. Catégorie morpho-syntaxique des INIT

Les syntagmes prépositionnels – SP – constituent la forme principale des INIT : près de 70% des INIT1 et plus de 40% des INIT2 adoptent la forme prépositionnelle. Au niveau des INIT2, la chute de proportion de SP est à lier aux 25% d'INIT2 annotés « Iautre » c'est-à-dire dont la forme n'a pas été identifiée par le programme.

	INIT(1+2)		%	
	Nb	%	INIT1	INIT2
SP	5 279	*65,1*	*68,9*	*40,6*
FIN (subordonnée)	696	*8,6*	*8,6*	*8,7*
SN	475	*5,9*	*5,4*	*8,6*
Iautre	394	*4,9*	*1,8*	*24,6*
PPA (p.passé)	336	*4,1*	*3,9*	*5,6*
ADV (adverbe)	284	*3,5*	*3,5*	*3,7*
PPR (p.présent)	216	*2,7*	*2,7*	*2,7*
INF (infinitive)	210	*2,6*	*2,7*	*1,6*
ADJ (S. adjectival)	176	*2,2*	*1,9*	*3,7*
REL (relative)	40	*0,5*	*0,5*	*0,4*
		8 106	**7 021**	**1 085**

Tableau VIII.10 : Répartition générale des formes des INIT

163 Nous rappelons que INIT1 et INIT2 ne sont pas deux types différents d'INIT mais seulement la distinction entre le premier élément détaché et les éléments suivants (voir VII.2.1).

Les INIT2 présentent des formes plus complexes à caractériser et à définir automatiquement. D'une part, un INIT2 peut comporter des ellipses (dont les référents implicites sont exprimés en INIT1) qui entraîne l'absence des éléments permettant précisément de définir sa forme (les prépositions par exemple). D'autre part, certains INIT2 sont en réalité une continuité de l'INIT1 ou une incise dans l'INIT1 et nous avons alors une erreur de programme.

(VIII.6) En fonction de l'âge des enseignants [INIT1], de leur grade, de leur statut et des taux d'encadrement [INIT2], trois grands types d'espaces peuvent être distingués. [ATLAS_2]

(VIII.7) Dans des lettres adressées à Donald Rumsfeld [INIT1], et aux responsables de la NIMA [INIT2], elles ont comparé le buy-to-deny à de la censure. [GEOPO_2]

(VIII.8) Grâce à ces auteurs[INIT1], aux troupes italiennes, au théâtre forain[INIT2], le mythe était déjà populaire en France quand Molière le reprit [...]. [PEOPL_1]

(VIII.9) Quant aux hypothèses non conventionnelles [INIT1], en particulier les scénarios terroristes [INIT2], on les tenait dans des limites imaginées d'après les expériences précédentes des années 1980 ; ou on les renvoyait aux technologies émergentes, donc à un avenir plus ou moins lointain. [GEOPO_28]

Dans ces configurations, l'INIT2 présente fréquemment une forme non reconnue par notre programme, *i.e.* une forme étiquetée « Iautre ». Mais malgré ce taux élevé d'Iautre en INIT2, la forme SP reste majoritaire recouvrant 65% de tous les INIT (1 et 2). La seconde place est occupée par les subordonnées qui représentent 8,5% des INIT.

VIII.3.2. Fonction discursive des INIT

Les adverbiaux circonstanciels constituent la fonction dominante des INIT (1 ou 2). 65% des INIT sont des adverbiaux circonstanciels – CIRC. Nous trouvons ensuite les appositions – APPO et les adverbiaux modalisateurs – MODA, sauf en INIT2 où le taux d'INIT à fonction non définie dépasse les modalisateurs d'énonciation[164]. Le tableau VIII.11 détaille également la proportion des différents rôles sémantiques des adverbiaux circonstanciels sur la totalité des INIT, ce qui fait apparaître la forte proportion des différents rôles sémantiques de CIRC : les adverbiaux temporels – CIRCtps – et les circonstants au rôle indéfini – CIRCautre – arrivent au deuxième et troisième rang des INIT les plus fréquents avec 30% et 20%. Les adverbiaux notionnels – CIRCnot – sont juste derrière les appositions. Les adverbiaux spatiaux – CIRCspa – sont les plus faibles étant, nous le verrons, quasi-exclusivement associés à ATLAS.

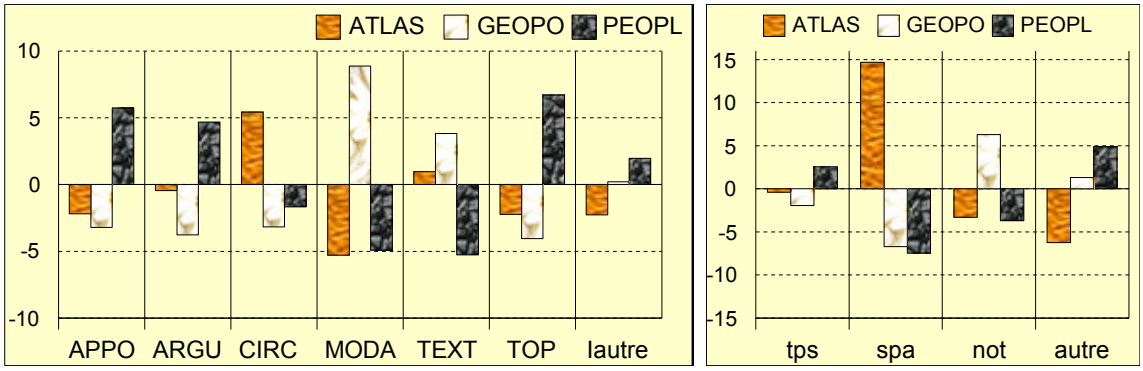
	INIT(1+2)		%	
	Nb	%	INIT1	INIT2
CIRC	5 406	66,7	68,9	52,4
CIRCautre	2 404	29,7	30,4	24,6
CIRCtps	1 641	20,2	20,8	17,0
CIRCnot	797	9,8	10,6	4,9
CIRCspa	564	7,0	7,1	5,9
APPO	1 055	13,0	12,0	19,4
MODA	664	8,2	8,7	4,8
TEXT	434	5,4	5,9	1,8
Iautre	295	3,6	1,4	18,0
ARGU	175	2,2	2,2	1,9
TOP	77	0,9	0,8	1,7
	8 106	100	**7 021**	**1 085**

Tableau VIII.11 : Répartition des différentes fonctions et rôles sémantiques des CIRC en INIT1 et INIT2

164 Cette plus grande 'indéfinitude' des INIT2 a déjà été expliquée dans la partie précédente.

Les CIRC représentent la fonction majoritaire des INIT. Les circonstances font partie prenante de l'expression de l'expérience, ce qui n'est pas le cas des adverbiaux modalisateurs et textuels. Les éléments disloqués à droite – TOP – et les arguments inversés – ARGU – ne peuvent pas non plus représenter un grand nombre d'INIT, étant lié exclusivement à des constructions spéciales ne représentant que 2,2% des phrases du corpus (voir tableau VIII.11). Seules les appositions, qui sont des prédications secondes, peuvent prétendre à une fréquence élevée, puisqu'elles participent également à la composante idéationnelle, ce qui explique certainement leurs 13%.

ATLAS est le sous-corpus qui présente le plus d'adverbiaux circonstanciels : près de 75% des INIT1 d'ATLAS sont des CIRC. GEOPO et PEOPL présentent également une forte proportion de CIRC en INIT1 (plus de 65% chacun). ATLAS est tout de même très associé aux adverbiaux circonstanciels avec un *z*=+5,4 et des écarts négatifs chez GEOPO et PEOPL, comme le montre les graphiques VIII.15.



Graphiques VIII.5 : Écarts significatifs selon les sous-corpus de la répartition des fonctions d'INIT1

Les rôles sémantiques des CIRC se répartissent chacun dans un sous-corpus particulier : ATLAS et les CIRCspa, GEOPO et les CIRCnot, PEOPL et les CIRCautre. Ces caractérisations positives chez l'un se retrouvent en négatives chez les autres. Là encore, la forte corrélation entre les CIRCspa et ATLAS se fait ressentir. ATLAS rassemble les 2/3 des adverbiaux spatiaux répertoriés (*z*=+16). Également en INIT2, ATLAS présente un écart positif de +5,5. Les adverbiaux temporels semblent insensibles au type de texte considéré. Que ce soit dans ATLAS, PEOPL ou GEOPO, il représentent autour de 12% des INIT1 et restent le rôle sémantique le plus fréquent dans tous les sous-corpus. Au vu de ces résultats, le type de circonstance associée aux INIT1 semble pouvoir assez bien caractériser nos sous-corpus.

Deuxième fonction d'INIT, l'apposition est fortement associée au sous-corpus PEOPL : plus de 40% des APPO repérées en INIT1 se situent dans ce sous-corpus, et 45% des APPO en INIT2. Alors que l'apposition représente autour de 10% des INIT1 d'ATLAS et GEOPO, chez PEOPL, elle représente 16% de INIT1. Nos données semblent signifier que l'apposition s'installe davantage dans des contextes tels que ceux rencontrés dans PEOPL.

Le cas des TOP (élément détaché d'une dislocation à droite) reste dans notre corpus très marginal. En effet, nous ne relevons que 59 INIT1 de ce type dont 47 dans PEOPL et 18 INIT2 dont 11 dans PEOPL. La grande association entre PEOPL et les dislocations (et les ThSpe en général) est également observée dans la partie VIII.5.

Les adverbiaux modalisateurs apparaissent en forte corrélation avec GEOPO. Ce type d'INIT montre un même comportement que les adverbiaux spatiaux : un écart positif élevé dans un sous-corpus et deux écarts négatifs significatifs dans les deux autres sous-corpus. Ce schéma nous assure la validité de la corrélation entre MODA et GEOPO. Une corrélation moins forte peut également être observée entre les adverbiaux textuels – TEXT et GEOPO. Ces deux fonctions d'INIT (les adverbiaux modalisateurs et textuels) boudent la position INIT2, ce qui va dans le sens de l'ordre des thèmes multiples suggéré en Systémique Fonctionnelle (voir IV.4).

Différents schémas d'écart apparaissent au travers de nos données. Premièrement, nous avons les schémas neutres pour les adverbiaux temporels et les INIT de nature indéfinie. Nous verrons que les adverbiaux temporels, fortement présents dans tous les sous-corpus montrent une forte sensibilité aux positions textuelles. Cette stabilité à travers les différents types de textes leur confère une forte capacité à organiser le discours en dépit du type de texte pris en compte. Deuxièmement, nous observons une association exclusive pour les appositions et les TOP avec PEOPL, les adverbiaux spatiaux avec ATLAS, les adverbiaux modalisateurs et notionnels avec GEOPO. L'association entre les adverbiaux spatiaux et ATLAS est particulièrement élevée. Nous verrons si ces associations se retrouvent dans toutes les positions textuelles ou seulement dans certaines positions stratégiques. Si elles sont associées à une position textuelle particulière, il y a fort à parier qu'elles participent, pour ce type de texte, au marquage de la séquentialité. Troisièmement, nous avons une dispersion des écarts pour les argument inversés et les adverbiaux textuels. Il semble donc que chacun des sous-corpus peut être caractérisé par son taux d'argument inversés et d'adverbiaux textuels. ATLAS, texte plus descriptif, présente un taux moyen de ces deux types d'INIT tandis que GEOPO et PEOPL s'opposent : GEOPO, plutôt argumentatif, favorise les adverbiaux textuels au dépit des arguments inversés et PEOPL, plus narratif, use fréquemment des arguments inversés au dépit des adverbiaux textuels. À noter que ces deux types d'INIT affichent un nombre d'apparitions relativement faible (voir tableau VIII.11). Le schéma des écarts pour les adverbiaux circonstanciels (sans distinction des rôles sémantiques) ressemble soit à une association exclusive avec ATLAS soit à une dispersion avec des adverbiaux circonstanciels en masse chez ATLAS, en nombre plus faible chez GEOPO et dans la moyenne pour PEOPL.

VIII.3.3. Des corrélations entre catégorie morpho-syntaxique et fonction discursive

Nous n'avons posé aucune association théorique entre une forme et une fonction, cependant, certaines corrélations paraissent difficiles comme celles entre une proposition relative et le rôle d'organisateur textuel. Dans un même ordre d'idée, une subordonnée à fonction d'apposition est de fait une relative. Le tableau VIII.12 présente les corrélations repérées dans notre corpus.

	MODA		APPO		ARGU		CIRC		TEXT	
	Nb	%	Nb	%	Nb	%	Nb	%	Nb	%
SP	379	**57,1**	34	3,2	162	**92,6**	4 306	**79,7**	398	**91,7**
FIN	1	0,2	-	-	4	2,3	691	12,8	-	-
lautre	5	0,8	-	-	4	2,3	90	1,7	-	-
PPA	13	2	323	**30,6**	-	-	-	-	-	-
ADV	248	**37,3**	-	-	-	-	-	-	36	8,3
SN	-	-	294	**27,9**	1	0,6	103	1,9	-	-
INF	1	0,2	-	-	3	1,7	206	3,8	-	-
PPR	-	-	205	19,4	1	0,6	10	0,2	-	-
ADJ	11	1,7	165	15,6	-	-	-	-	-	-
REL	6	0,9	34	3,2	-	-	-	-	-	-
	664		1 055		175		5 406		434	

Tableau VIII.12 : Forme de certains INIT selon leur fonction

La prédominance des SP en position initiale détachée se retrouve plus ou moins prononcée pour toutes les fonction d'INIT1 à l'exception des appositions – APPO. Les appositions présentent quatre réalisations principales : des propositions avec participe passé – PPA (30,6%), des SN (27,9%), des propositions avec participe présent – PPR

(19,4%) et des syntagmes adjectivaux – ADJ (15,6%). Nous retrouvons ici les catégories identifiées par Neveu (1998) et illustrées en V.4.2. Les adverbiaux modalisateurs – MODA – se distinguent également par leur forte propension à se réaliser sous forme d'adverbe (*Heureusement, Assurément,* etc.), même si la forme SP reste majoritaire *(En réalité, En particulier, Selon toute vraisemblance*, etc.) Les arguments inversés – ARGU – et les adverbiaux textuels – TEXT – ne sont quasiment que des SP.

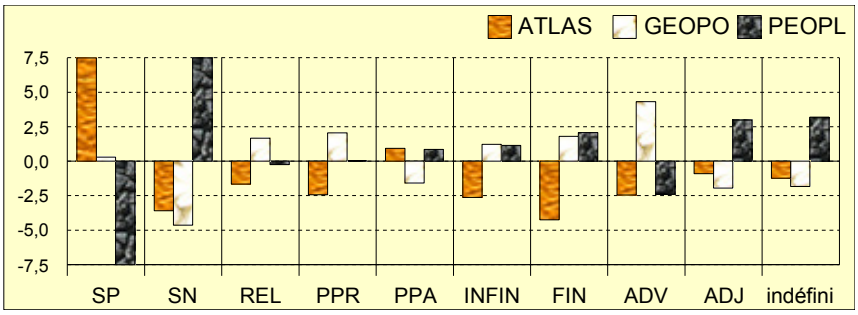
Les associations évoquées dans le tableau VIII.12 ne sont donc pas surprenantes, sauf peut-être quelques cas rares illustrés par les exemples suivants. Ces cas rares sont généralement issus d'une mauvaise catégorisation morpho-syntaxique. Pour les MODA, nous avons généralement affaire à des locutions adverbiales mal étiquetées, comme les cas des 6 relatives qui sont en réalité la locution « *quoi qu'il en soit* », ou encore les modalisateurs « *bref* » ou « *bien sûr* » qui sont étiquetés adjectifs et non locutions adverbiales. Le tableau suivant reprend les corrélations rares et y associe la liste des expressions correspondantes repérées ainsi que leur nombre d'apparitions et une phrase d'exemple mettant en texte une des expressions repérées.

corrélation rare	expressions	Nb	exemple
MODA/REL	*quoi qu'il en soit*	6	*Quoi qu'il en soit, la principale limite du projet Hariri reste dans sa propre surévaluation.[GEOPO_6]*
MODA/FIN	*plus que tout*	1	*Plus que tout, ils se méfient de l'utilisation de la force militaire, [...]. [GEOPO_19]*
MODA/INFIN	*à dire vrai*	1	*À dire vrai, tout est proportionné, puisque rien ne saurait échapper à la loi de la série : "[...] [PEOPL_22]*
MODA/ADJ	*bref*	6	*Bref, si l'accueil de collégiens et lycéens domiciliés dans d'autres départements ne gonflait pas artificiellement les effectifs des collèges et plus encore des lycées, Paris pourrait constituer le modèle de référence en matière de performance des systèmes éducatifs. [ATLAS_2]*
	bien sûr	2	
	petit à petit	2	
	sans nul doute possible	1	
MODA/PPA	*bien entendu*	5	*Autrement dit, l'attention des observateurs américains est déjà largement concentrée sur les modalités de la mise en oeuvre de la lutte anti terroriste.[GEOPO_22]*
	autrement dit [165]	8	
CIRC_not/PPR	*concernant* X	10	*Concernant les questions de surveillance et de renseignement, par exemple, l'Administration Bush a été critiquée par les parlementaires, [...]. [GEOPO_20]*

La répartition des différentes formes d'INIT se retrouve à peu près également dans les différents sous-corpus. Les SP sont régulièrement majoritaires (73% pour ATLAS, 65% pour GEOPO et 58% pour PEOPL). Cela s'explique en grande partie par la forte proportion d'INIT adverbiaux circonstanciels. Certaines particularités entre sous-corpus apparaissent, parfois en lien avec la fonction des INIT propres aux sous-corpus concerné. Ainsi, PEOPL présente une association quasi-exclusive avec les SN (fortement liés aux APPO) ; GEOPO s'associe aux adverbes (liés aux MODA) ; ATLAS s'oppose à PEOPL au niveau des SP.

L'opposition observée au niveau des SP répond à deux phénomènes : moins de circonstants dans PEOPL que dans ATLAS et une corrélation entre la fonction circonstant et la forme proposition subordonnée – FIN – plus forte chez PEOPL, ce que montre le graphique du tableau VIII.13 ci-dessous. Les deux tableaux suivants indiquent les corrélations forme/fonction pour les INIT de fonction circonstant et apposition. Ces deux types d'INIT sont les seuls qui présentent des écarts significatifs selon les sous-corpus par rapport au modèle général.

165 Ces 8 occurrences d'*Autrement dit* apparaissent dans trois textes uniquement. Même si ces trois textes n'émanent pas du même auteur, il semble que l'utilisation de cette locution relève d'une préférence auteuriale.



Graphique VIII.6 : Variations des formes des INIT selon les sous-corpus

Au niveau des circonstants, ATLAS utilise quasi-exclusivement des SP : 86,5% des INIT_CIRC sont des SP chez ATLAS. Nous observons alors un écart réduit de *z*=+6,9 pour la corrélation CIRC/SP et des écarts négatifs pour les autres corrélations CIRC-{SN, PPR, FIN}. PEOPL montre une plus grande corrélation entre les adverbiaux circonstanciels et les propositions subordonnées finies : 15,2% des INIT_CIRC sont des subordonnées – FIN – chez PEOPL. GEOPO ne montre pas de variations significatives concernant les formes de circonstants.

CIRC	*Nb phr*
SP	4 306
SN	103
PPR	10
INF	206
FIN	691
autre	90

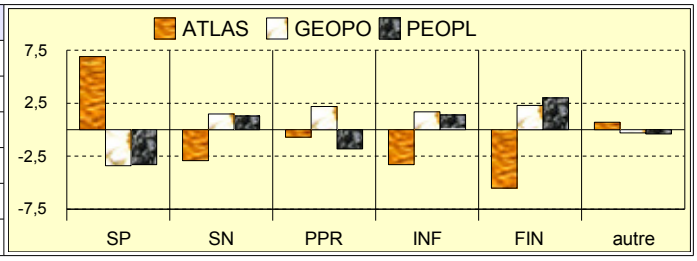


Tableau VIII.13 : écarts réduits selon les sous-corpus pour les corrélations forme/CIRC

Au niveau des appositions, nous observons une alternance entre SN chez PEOPL, participe passé chez ATLAS et participe présent ou relative chez GEOPO. Dans le tableau VIII.12, les appositions semblaient être généralement construites sous forme de SN ou de participe passé. Or, il semble que la catégorie des appositions varie selon le type de texte, ce que montre le tableau VIII.14. Ainsi, la forte proportion de SN est liée à la forte fréquence d'appositions dans PEOPL, sous-corpus qui favorise la corrélation APPO/SN.

APPO	*Nb phr*
SP	34
SN	294
REL	34
PPR	205
PPA	323
ADJ	165

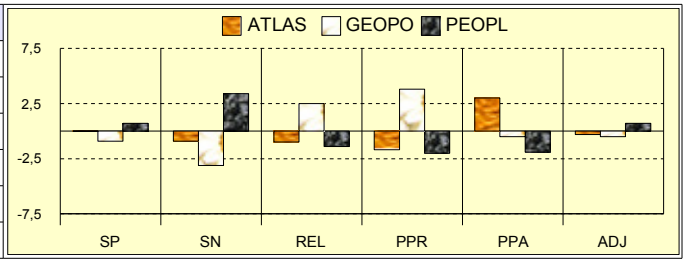


Tableau VIII.14 : écarts significatifs selon les sous-corpus pour les corrélations forme/fonction en INIT

VIII.3.4. Séquence en INIT

La majorité des phrases qui présentent un INIT n'en présentent qu'un. Cependant, 1 085 cas de séquences INIT1+INIT2 ont été repérées. À peu près tous les cas de figures apparaissent, mais dans des proportions très

disparates. Ainsi, cinq séquences n'apparaissent qu'une seule fois dans tous le corpus tandis que trois patrons couvrent la moitié des occurrences : CIRC-CIRC, CIRC-APPO et CIRC-indéfini.

En Systémique Fonctionnelle, un ordonnancement des différents éléments en position initiale est généralement admis : Thème textuel, Thème interpersonnel et Thème idéationnel (voir IV.4). Il serait donc difficile d'avoir un adverbial circonstanciel suivi d'un adverbial modalisateur ou textuel. Et impossible d'avoir une apposition suivie d'un adverbial modalisateur ou textuel. Nous ne trouvons en effet aucune apposition suivie d'un modalisateur ou d'un organisateur textuel.

Pour ce qui est des séquences CIRC-MODA ou CIRC-TEXT, quelques cas sont repérés, dont la plupart appartiennent à GEOPO. 55 séquences CIRC-MODA ou CIRC-TEXT sont repérées au total dont 34 dans GEOPO.

(VIII.10) Durant les années 1990, toutefois, ce n'est pas tant l'éventualité d'une telle résurgence qui rend la réserve si précieuse, mais plutôt [...]. *[GEOPO_30]*

(VIII.11) En toile de fond, enfin, le débat autour de la RMA s'amplifie tout au long de la période [...] *[GEOPO_30]*

La séquence la plus fréquente est constituée d'un CIRC suivi d'un second CIRC. Dans les exemples retenus, il semble que la permutation entre le premier CIRC et le second semble soit possible, surtout lorsque les deux CIRC ont un rôle sémantique différent :

(VIII.12) *Le 20 avril 1990, dans l'Ouest communautaire, les élus de 32 régions littorales, de l'Écosse à l'Andalousie, officialisent la création de la Commission Arc Atlantique.* *[ATLAS_1]*

(VIII.13) *En physique, pour obtenir sa mutation dans le Lot, un certifié devait avoir 257 points, c'est-à-dire, en l'absence de points de rapprochement de conjoint, 15 à 20 ans d'ancienneté.* *[ATLAS_2]*

(VIII.14) *Le 19 septembre 2002, après avoir présenté à l'Assemblée générale des Nations unies sa position sur la situation en Irak, Gearges W. Bush s'est adressé au Congrès pour lui demander un vote l'autorisant à faire usage de la force, afin de [...]* *[GEOPO_20]*

(VIII.15) *À Rouen, en 1640, Pascal envisagea de construire une machine effectuant les quatre opérations arithmétiques élémentaires.* *[PEOPL_4]*

Cependant, si l'on remet les exemples en contexte, il apparaît vite que l'ordre des circonstants n'est pas anodin. Pour l'exemple VIII.13 , remis en contexte, l'ordre des circonstant permet de réaliser une mise en contraste plus nette avec le circonstant mentionné juste avant en position postverbale *en histoire-géographie pour obtenir sa mutation dans la Loire-Atlantique*. Une inversion de l'ordre des circonstants semble alors moins cohérente.

(VIII.13') *Nous avons retenu le barème de mutation des professeurs de type lycée (56% des enseignants du secondaire), le principe utilisé étant voisin dans le cas des professeurs de lycée professionnel. [...] Il est possible, sur ce principe, d'établir la hiérarchisation spatiale opérée par les professeurs. Le niveau départemental peut être facilement étudié: afin de donner des points de repère aux enseignants dans la formulation de leurs voeux, les syndicats produisent chaque année, par discipline, les barèmes du "dernier entré" dans chaque département (on prend le plus petit barème qui a permis "d'obtenir" le département). À l'occasion du mouvement de 1988, il fallait par exemple 69 points en histoire-géographie pour obtenir sa mutation dans la Loire-Atlantique: ce département était accessible, en premier poste, à un certifié sortant de stage, bénéficiant des points de rapprochement de conjoint.* ***En physique, pour obtenir sa mutation dans le Lot,*** *un certifié devait avoir 257 points, c'est-à-dire, en l'absence de points de rapprochement de conjoint, 15 à 20 ans d'ancienneté. [?****Pour obtenir sa mutation dans le Lot, en physique,*** *un certifié devait avoir 257 points, c'est-à-dire, en l'absence de points de rapprochement de conjoint, 15 à 20 ans d'ancienneté.] En affectant aux barèmes de chaque discipline un coefficient correspondant à la proportion d'enseignants qui en relèvent, nous avons calculé des barèmes synthétiques pour les mouvements 1987, 1988 et 1989 des professeurs de type lycée.*

D'autres raisons peuvent également être avancées comme la longueur des circonstants : il semble préférable de commencer avec l'élément le plus court, comme en (VIII.14') où l'ordre des constituant n'est pas en relation avec une organisation temporelle spécifique.

(VIII.14') ***3. Le vote su congrès*** [titre niveau 2]
Le 19 septembre 2002, après avoir présenté à l'Assemblée générale des Nations unies sa position sur la situation en Irak, *George W. Bush s'est adressé au Congrès pour lui demander un vote l'autorisant à faire usage de la*

force, afin de " faire appliquer les résolutions susmentionnées du Conseil de sécurité des Nations unies, de défendre les intérêts de sécurité nationale des Etats-Unis contre les menaces émanant de l'Irak, et de restaurer la paix internationale et la sécurité dans la région". Secondé par Donald Rumsfeld et Colin Powell, le président a justifié la nécessité de ce vote, dans un souci d'unité, pour faire face de façon plus crédible à la " menace que fait planer le régime de Saddam Hussein". Cette initiative a été saluée par les parlementaires, qui y ont vu une volonté d'ouverture en direction du pouvoir législatif, et une reconnaissance de leurs prérogatives.

Enfin, la permutation peut être tout à fait réalisable, comme en (VIII.12') et (VIII.15').

(VIII.12') *La construction européenne s'accélère. Le Marché Unique, l'ouverture des frontières à l'Est, les réformes de la politique régionale mobilisent les régions périphériques en retard de développement.* ***Le 20 avril 1990, dans l'Ouest communautaire,*** *les élus de 32 régions littorales, de l'Écosse à l'Andalousie, officialisent la création de la Commission Arc Atlantique. Dans le cadre de la Conférence des Régions Périphériques Maritimes de l'Union européenne, les objectifs annoncés sous la présidence d'Olivier Guichard sont de créer un lobby atlantique face au recentrage de l'Europe vers l'est, d'engager des programmes de coopération interrégionale le long de la façade atlantique, d'obtenir de la Communauté européenne des financements spécifiques.*

(VIII.15') ***2.6) La machine arithmétique*** [titre niveau 2]
Seule contribution de Pascal au progrès des sciences appliquées, la machine arithmétique est une réalisation profondément novatrice. [...]
À Rouen, en 1640, *Pascal envisagea de construire une machine effectuant les quatre opérations arithmétiques élémentaires. Son objectif était de faciliter les pénibles opérations comptables dont son père avait la charge.*
Surtout depuis le début du XVIIe siècle, l'accroissement et la complexité des opérations numériques qui résultaient du développement aussi bien de l'algèbre et de l'astronomie que des opérations commerciales et bancaires avaient conduit à la création, en plus des systèmes de jetons et de bouliers d'usage ancien, de réglettes utilisant les logarithmes récemment inventés par John Napier. Chronologiquement, la première machine [...]

31,6% des séquences sont constituées de deux circonstants. Cette proportion varie entre 25,4% chez GEOPO (*z*=-2,5) à 41,5% chez ATLAS (*z*=+8), ce qui concorde avec le schéma de dispersion des circonstants (graphique VIII.5). Malgré ces variations, la séquence CIRC-CIRC reste la plus importante pour tous les sous-corpus. D'ailleurs, la majorité des séquences – 685 précisément, soit 63% des séquences – ont pour INIT1 un CIRC. Viennent ensuite celles commençant par une apposition (17%), un adverbial modalisateur MODA (9%) et textuel (8%).

INIT1	*INIT2*	*Nb Phrases*	*%*
CIRC	CIRC	343	*31,5*
CIRC	APPO	133	*12,3*
CIRC	Iautre	129	*11,9*
MODA	CIRC	76	*7,0*
APPO	CIRC	73	*6,7*
TEXT	CIRC	64	*5,9*

Tableau VIII.15 : Répartition des séquences d'INIT les plus fréquentes

Nous retrouvons ici les mêmes résultats qu'en cumulant les fréquences isolées des fonctions en INIT1 et en INIT2. Aucune donnée nouvelle n'apparaît. Nous conservons l'ordre de fréquence d'apparition : CIRC, APPO, MODA chez INIT1 et CIRC, Iautre, APPO chez INIT2.

Cet ordre reste relativement le même dans les différents sous-corpus. Seuls trois écarts significatifs sont mesurés : ATLAS présente plus de CIRC-CIRC (*z*=+3,8) et GEOPO plus de CIRC-MODA (*z*=+3,3) et de MODA-CIRC (*z*=+2,5). De façon générale, GEOPO montre plus de MODA, que ce soit en INIT1 ou en INIT2. PEOPL n'affiche aucun écart significatif.

Concernant les 343 séquences CIRC-CIRC, 39% d'entre elles ont en INIT1 soit un CIRCtps, soit un CIRCautre. La majorité des séquences circonstancielles sont constituées de deux CIRCautre (22%). 18% d'entre elles sont constituées de deux CIRCtps qui se suivent. Là encore, nous observons un cumul des comportements observés pour chaque rôle sémantique pris isolément. Le tableau VIII.16 présente tous les types de séquences repérées et leur

nombre d'apparitions, ce qui montre bien que sur 23 217 phrases, la présence de deux circonstants successifs est bien faible.

sem1	autre				tps				not			spa				total
sem2	tps	spa	not	autre	tps	spa	not	autre	tps	not	autre	tps	spa	not	autre	
ATLAS	6	5	5	15	23	6	3	22	3	2	6	7	13	4	12	132
GEOPO	10	3	5	19	19	3	4	12	1	0	7	1	0	1	2	87
PEOPL	16	1	8	41	20	4	3	15	2	1	4	5	3	0	1	124
Corpus	32	9	18	75	62	13	10	49	6	3	17	13	16	5	15	343
	9	*2,5*	*5*	*22*	*18*	*4*	*3*	*14*	*2*	*1*	*5*	*4*	*5*	*1,5*	*4,4*	*100%*

Tableau VIII.16 : Rôles sémantiques dans les séquences CIRC-CIRC

Si l'on compare les sous-corpus et leur utilisation des séquences CIRC-CIRC, quatre variations significatives apparaissent qui font écho aux variations observées précédemment. ATLAS présente significativement plus de séquences impliquant des adverbiaux spatiaux : *z(CIRCspa-CIRCspa)*=+2,8 et *z(CIRCspa-CIRCautre)*=+2,7. ATLAS présente également un écart négatif pour les séquences CIRCautre-CIRCautre (*z*=-2,9) qui apparaissent significativement plus dans PEOPL (*z*=+3). GEOPO ne présente aucune variation pertinente.

VIII.4. Des connecteurs aux formes variables

Les connecteurs 'purs' que nous avons retenus dans notre annotation ont pour principale caractéristique leur absence de contenu idéationnel, c'est d'ailleurs pour cela que nous les avons exclus de la catégorie INIT. Leur fonction est d'expliciter les relations de discours entre deux propositions. Nous n'avons effectué aucune classification des connecteurs 'purs' selon le type de relation de discours qu'ils pouvaient exprimer. De ce fait, nous nous retrouvons avec une catégorie très hétérogène. L'analyse d'une telle catégorie ne paraît pas toujours pertinente. Cependant les variations selon les différents sous-corpus nous semblent tout à fait intéressantes.

Moins de 10% des phrases présentent un connecteur en initiale. La liste des connecteurs apparaissant plus d'une fois en position initiale est donnée en annexe I. Tout comme le remarque Romera (2004), la majorité des occurrences de connecteurs ne correspondent, au final, qu'à très peu de formes différentes. La distribution des connecteurs dans notre corpus confirme cet état de fait. Le tableau VIII.17 montre que seulement 9 formes de connecteurs couvrent 77%

Forme	Total		
	occ.	%	
1. Mais	704	*31,7*	
2. Et	239	*10,8*	
3. Ainsi	220	*9,9*	*52,4*
4. Or	166	*7,5*	
5. Enfin	124	*5,6*	
6. Cependant	95	*4,3*	
7. Car	61	*2,7*	
8. Certes	53	*2,4*	
9. Pourtant	51	*2,3*	*24,8*
Total pour ces 9 formes	**1 713**		*77,2*
Total final pour 87 formes	**2 220**		*100*

Tableau VIII.17 : Connecteurs les plus fréquents repérées dans notre corpus

des occurrences de connecteurs repérés. Si l'on ne considère que les trois premières lignes, nous voyons qu'à peine 3 formes de connecteurs couvrent plus de 50% des occurrences.

D'un autre point de vue, 23% des occurrences de connecteurs relèvent de 78 formes différentes. En d'autres termes, nous avons d'un côté 77 formes qui apparaissent en moyenne 7 fois dans tout le corpus et 9 formes qui présentent en moyenne 190 occurrences chacune.

Le travail de Romera présente des rapports similaires bien qu'il se base sur un corpus oral composé de 68 conversations (95 683 mots) issues du CREA[166]. La similarité des résultats est très troublante.

Forme (traduction)	*Fq*	*%*
1. Pero (Mais)	439	*19,2*
2. Y (Et)	298	*13*
3. Porque (Parce que)	256	*11,2*
4. Pues (Donc)	224	*9,8*
5. Es Que (Est-ce-que)	174	*7,6*
6. O Sea (C'est-à-dire)	122	*5,3*
7. Entonces (Alors)	121	*5,3*
8. Bueno (Bon)	73	*3,2*
9. Si (Si)	51	*2,2*
total intermédiaire pour ces 9 formes	**1758**	*76,8*
total final pour 108 formes	**2292**	*100*

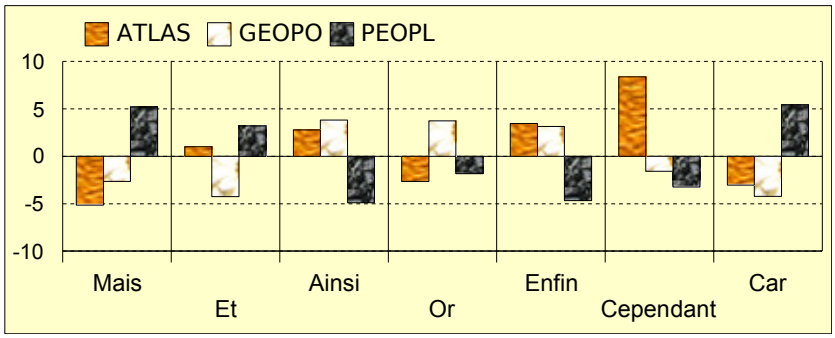
Tableau VIII.18 : Connecteurs les plus fréquents en espagnol oral selon Romera (2004 : 81)

Dans son corpus oral, 4 formes couvrent plus de 50% des occurrences, 9 formes plus des trois-quarts et les 99 formes restantes couvrent les derniers 23%. Le repérage de ces connecteurs a été réalisé manuellement selon la définition suivante : un connecteur est une catégorie fonctionnelle qui apparaît à la frontière d'unités de discours et qui établit une relation de cohérence entre ces unités (cf. Romera 2004:72). Il s'agit donc d'un repérage fonctionnel[167] et non morpho-syntaxique comme le nôtre (voir partie VII.2.1). Mais malgré la différence de méthodologie et surtout de registre (oral *vs.* écrit, on est frappé par la ressemblance entre les deux listes de connecteurs les plus 'couvrants'. Ce rapprochement se renforce si l'on met de côté les connecteurs typiques de l'oral tels que *Es Que*, *O Sea*, *Bueno* ou encore *Si*.

Si l'on observe à présent les distributions de ces formes selon les sous-corpus, la variation d'usage selon le type de texte devient évidente. À un niveau général tout d'abord, l'usage des connecteurs semble plus fort dans des textes descriptifs ou narratifs présentant un cadre mono-référentiel tels que ceux composants PEOPL. En effet, près de la moitié des connecteurs repérés (46,4% soit 1 029 occurrences) se trouvent dans PEOPL, contre moins de 40% dans GEOPO (37,6% soit 835 occurrences) et 16% dans ATLAS (356 occurrences). Le test de l'écart réduit montre qu'aucun sous-corpus n'est dans la 'moyenne acceptable'. Ainsi, les connecteurs sont significativement moins fréquents dans ATLAS (z=-14,4) et plus fréquents dans PEOPL (z=+11,2). GEOPO montre également plus de connecteurs que la moyenne, mais dans des proportions moindres (z=+3).

166 Le « *Corpus de Referencia des Español Actual* » est composé d'une série de textes oraux et écrits contemporains provenant de différents pays hispanophones. Il a été construit par l'Universidad Autónoma de Madrid.

167 Romera (2004) ne s'intéresse pas qu'aux connecteurs, mais à ce qu'elle nomme des *Discourse Functional Units – DFU*, qui correspondent à la définition citée précédemment. Selon cette définition, des éléments autres que connecteurs 'purs' sont annotés. Par exemple, *En segundo lugar* est pour Romera une DFU. Dans notre étude, un élément de ce type (*En second lieu*) est annoté INIT à fonction textuelle (TEXT). La partie V.3.4) soulève cette difficulté de différenciation entre adverbiaux textuels et connecteurs.



Graphique VIII.7: Variations selon les sous-corpus des connecteurs les plus fréquents

Les formes de connecteurs varient également significativement d'un sous-corpus à l'autre. Si l'on mesure les écarts réduits des distributions par sous-corpus, les sept connecteurs les plus fréquents montrent tous des écarts réduits significatifs, comme le montre le graphique VIII.7

D'après les résultats représentés par le graphique VIII.7, les variations significatives ne vont pas spécialement dans le sens de la répartition générale des sous-corpus[168]. Ces écarts ne sont donc pas une simple conséquence du mouvement général mais une particularité des sous-corpus. Cela nous permet d'associer les connecteurs {*Mais*, *Et, Car*} à PEOPL, {*Ainsi*, *Or, Enfin*} à GEOPO et {*Cependant*, *Enfin*} à ATLAS.

VIII.5. Composition des Thèmes spécifiques – ThSpe

Un sixième des phrases sont des ThSpe (3 857 phrases). Dans PEOPL, ils représentent près de 22% des phrases (PEOPL est le sous-corpus qui présente le plus de ThSpe *z(ThSpe)*=+12,3, cf. tableau VIII.3 p.202). Dans 7,5% des cas, un connecteur (sans INIT) précède le Thème spécifique ; dans 26% c'est un INIT (non précédé d'un connecteur) et dans 3% ce sont un connecteur et un ou plusieurs INIT. Tout comme les Thèmes topicaux, les Thèmes spécifiques apparaissent majoritairement seuls (dans 63,5% des cas).

	corpus entier				*%*		
	Nb	***%***		***%'***	***ATLAS***	***GEOPO***	***PEOPL***
clivée	933	*24,2*	*63,6*	*4,0*	*3,6*	*3,0*	*5,5*
On...	836	*21,7*		*3,6*	*2,8*	*3,1*	*4,9*
ILimp	682	*17,7*		*2,9*	*1,8*	*3,9*	*3,1*
ThSpe autre	539	*14,0*		*2,3*	*1,5*	*1,9*	*3,5*
SujInv	440	*11,4*		*1,9*	*2,0*	*1,1*	*2,6*
Present	351	*9,1*		*1,5*	*1,2*	*1,9*	*1,5*
Disloc	76	*2,0*		*0,3*	*0,2*	*0,1*	*0,8*
		100			*100*	*100*	*100*
ThSpe	3 857	*16,6*			*12,9*	*15,1*	*21,8*
Nb Phr.	23 217				7 592	7 901	7 724

* la colonne % affiche la proportion pour les ThSpe et la colonne %' affiche la proportion pour le corpus entier.

Tableau VIII.19 : Répartition des différents types de ThSpe dans notre corpus

La catégorie des Thèmes spécifiques rassemble des constructions textuelles aux fonctions variées : certaines permettent de mettre une entité ou une circonstance en avant (les clivées ou les dislocations – Disloc) quand d'autres

168 Beaucoup de connecteurs dans PEOPL, assez dans GEOPO et peu dans ATLAS (voir partie VIII.1).

ont pour effet de mettre le Thème en arrière-plan (les impersonnelles, les constructions en On..., les sujets inversés). Nous nous retrouvons face à une catégorie très hétérogène et dont les variations selon les différents sous-corpus ne sont pas toujours pertinentes à comparer, vu la différence fondamentale entre certaines de ces constructions. Sans hypothèses fortes quant à l'implication des Thèmes spécifiques dans le marquage de la séquentialité, nous avons effectué les même analyses qu'au niveau des ThTop pour les ThSpe.

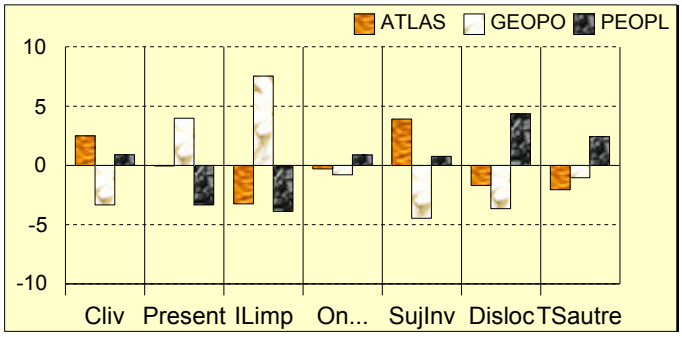
Si l'on compare la fréquence des différents types de ThTop à celle des différents types de ThSpe, nous remarquons que les clivées, type de ThSpe le plus fréquent, représentent à peine 4% des phrases du corpus entier, alors que presque tous les ThTop dépassent les 6% de phrases, en exceptant les SN possessifs et les pronoms démonstratifs (voir tableau VIII.5 p.204).

Chaque type de construction distinguée représente entre 10 et 25% des ThSpe, si l'on exclut les dislocations. Vu leur très faible fréquence, on peut dire que les dislocations relèvent d'un phénomène rare (à peine 76 phrases). Cela s'explique sûrement par leur association au discours oral plus qu'à l'écrit (cf. V.4.4). Nos données montrent que les dislocations sont spécifiques à PEOPL qui représente 78% des dislocations repérées. Cela peut être mis en relation avec le style plus littéraire de ce sous-corpus, comme l'illustrent les deux exemples suivants (il est difficilement imaginable de rencontrer de telles phrases dans ATLAS ou PEOPL).

(VIII.16) ***L'école, c'est*** *la solitude parmi les autres et le début de la conscience de solitude, c'est le premier monde où l'on est jeté sans clés ni indices, où les autres sont déjà un mystère que l'on pressent noir sans comprendre. Solitude encerclée déjà, car les autres sont ensemble, solitude humiliée déjà avec ce père douteux, et l'insécurité commence avec l'humiliation pressentie. Vient la première injustice, les coups de règle sur les doigts, qu'on n'a pas mérités, et le visage sinistre et cruel du bourreau, penché contre soi. Il ne sera plus seul dans la solitude, mais avec les héros-victimes, depuis le Christ jusqu'à Parnell. Il s'oriente vers une pose de solitude. [PEOPL_2]*

(VIII.17) ***L'Agneau sauveur, c'est*** *l'équilibre harmonieux de la double nature et, par extension, le rythme masculin équilibré et attendri par le rythme ternaire féminin. [PEOPL_18]*

Si l'on compare les répartitions des différentes constructions spéciales selon les sous-corpus, les écarts observés sont beaucoup plus faibles que ceux observés pour les phrases à Thème topical[169], ce que montre le graphique VIII.8. La plupart des schémas sont des schémas de dispersion. Seuls deux schémas d'écart sont neutres, relatifs aux constructions en On... et aux constructions non définies – TSautre.



Les écarts réduits sont mesurés par rapport à la totalité des ThSpe (et non la totalité des phrases)

Graphique VIII.8 : écarts réduits des différents types de ThSpe selon les sous-corpus

Malgré le fait que PEOPL soit le sous-corpus présentant la plus grande proportion de ThSpe pour quasiment tous les types de ThSpe distingués, seules les dislocations s'associent à PEOPL, nous en avons déjà parlé. Pour les autres types de construction, dans le schéma de dispersion, PEOPL représente la moyenne. En d'autres termes, sa très forte

169 Le graphique VIII.8 affiche un axe des abcisses allant de 10 à -10, alors que le graphique VIII.2 représentant les variations des différents types de ThTop affiche un axe des abcisses allant de 25 à -25.

proportion de Thèmes spécifiques est répartie équitablement selon les différents types caractérisés et ne se concentre pas sur un type particulier.

ATLAS se caractérise par un écart positif au niveau des clivées (avec un écart à la limite du significatif : *z*=+2,5) et des constructions inversées. Ces deux constructions spéciales s'accordent effectivement bien à l'introduction de circonstances spatiales ou temporelles (voir la description des clivées et des inversions dans la partie V.4.4) comme le montrent les exemples suivants.

(VIII.18) ***Vers l'Ouest, sur une plage de 25 kilomètres, de Weymouth à Bridport,*** *se trouve un élevage de cygnes - Abbotsbury Swannery - connu dans le monde entier. [ATLAS_1]*

(VIII.19) ***C'est dans le Midi méditerranéen et aquitain et en Île-de-France*** *que les enfants d'employés sont en plus grand nombre - jusqu'à près de 20% des enfants en Corse, à Nice, en Seine-Saint-Denis ; par contre, les départements ruraux du Centre-Ouest et du Centre-Est en comptent souvent moins de 10%. [ATLAS_2]*

(VIII.20) ***C'est au cours des années 70****, la décennie où les grands traits de la matrice historique et spatiale dans laquelle nous vivons se sont affirmés, que le développement des lignes maritimes traversant la Manche a été le plus fort. [ATLAS_1]*

Les clivées sont généralement utilisées pour mettre le focus sur une entité ou une circonstance pour laquelle le phénomène étudié est caractéristique. Nous trouvons ainsi des clivées de type : *c'est* X *qui est le plus/moins {important, fort, concerné, etc.} ou,* comme dans les exemples (VIII.19) et (VIII.20), *c'est SP que X est les plus/moins {important, fort, concerné}* ou encore *c'est (particulièrement) le cas {de, pour, en, dans, etc.} X*. Dans les cas de clivées mettant le focus sur un SP, il semble bien que ce soit la particularité de la circonstance exprimée par le SP qui est soulignée (et non la particularité de l'entité *X*)

À cet usage emphatique s'ajoute un usage conclusif (exemple VIII.5) où le focus est mis sur le connecteur *ainsi*.

(VIII.21) *Cette intensification des contacts, de la connaissance, des relations, cette multiplication des médias culturels au sens large, crée de l'interconnaissance et de la sympathie. Cette interconnaissance concerne, nous venons de le voir, des fractions de population très significatives en nombre, mais également diversifiées. Ce processus est adossé au contexte d'ouverture européenne. Celui-ci plus institutionnel, plus général, est facilitant, incitant, pour nombre des relations qui se créent. Ces relations en retour trouvent des occasions plus fréquentes, plus faciles, dans ce contexte.* ***C'est ainsi que les facteurs de rapprochement jouent entre les sociétés littorales de la Manche.*** *[ATLAS_1]*

Les clivées conclusives restent peu fréquentes. 5,6% des clivées (52/933) ont cette forme. Une étude précise sur ce sujet pourrait préciser le fonctionnement de ces types de clivées qui trouveraient certainement leur place dans la liste des marqueurs méta-discursifs.

Les impersonnelles – ILimp – se retrouvent en masse et de façon exclusive chez GEOPO (*z*=+*7,5*). Cette association avec ILimp peut être mise en relation avec l'aspect argumentatif de GEOPO, les impersonnelles permettant de porter indirectement un jugement sur un fait (V.5.2). Nous retrouvons cet aspect argumentatif de GEOPO dans sa proportion d'adverbiaux modalisateurs (voir partie VIII.3.2). GEOPO utilise également beaucoup de constructions présentationnelles (*z*=+4). L'association avec les présentationnelles recoupe l'idée d'une hétérogénéité dans les progressions thématiques chez GEOPO (voir exemple VIII.2 p.207). En effet, ces constructions spéciales ne « jouent pas de rôle actif dans la progression thématique » (Carter-Thomas 2000, V.4.4).

Ces deux écarts positifs sont les seuls observés chez GEOPO. Tous les autres ThSpe (hors schéma neutre) y sont significativement moins présents. GEOPO s'oppose ainsi autant à ATLAS (au niveau des clivées et des inversées) qu'à PEOPL (au niveau des présentationnelles et des dislocations).

Au vu de ce panorama sur les répartitions des constructions à thème spécifique, nous voyons que chaque type de construction peut être significativement plus important dans un certain type de texte. La plupart des Thèmes spécifiques affichent un schéma d'écart de dispersion, ce qui indique que ces éléments ne sont absolument pas stables

à travers les différents types de textes. Ils peuvent de ce fait constituer des indicateurs dans l'élaboration d'une catégorisation automatique des textes.

VIII.6. Degré d'accessibilité – DegAccess

Notre annotation nous permet également de mesurer le degré d'accessibilité des phrases selon l'échelle établie par Ariel (1990), échelle présentée en V.4.3. Cette mesure ne distingue aucun type de construction spéciale puisque les ThSpe se voient attribués un DegAccess_0, égal à celui des descriptions indéfinies. Le calcul du degré d'accessibilité nous apporte un autre regard sur la répartition des ThTop en effectuant certains regroupements. Cependant, ainsi que plusieurs travaux l'ont remarqué (notamment Schnedecker 2005), ce regroupement n'est peut-être pas aussi générique que l'affirme Ariel (1990) puisqu'il semble qu'à l'intérieur d'une même langue, le marquage de l'accessibilité peut fortement varier d'un type de texte à un autre.

DegAccess (description - rappel)		Corpus Nb	Corpus %
0	SNind, ThTop autres et ThSpe	6 759	29,0
1	Nom Propre nouveau	801	3,5
2	SNdef long	4 229	18,0
3	SNdef court et/ou avec reprise	5 914	25,5
4	Nom Propre repris	792	3,5
5	SNdem long	309	1,0
6	SNdem court et/ou avec reprise	1 229	5,5
7	Pronoms et SNpossessifs	3 185	14,0

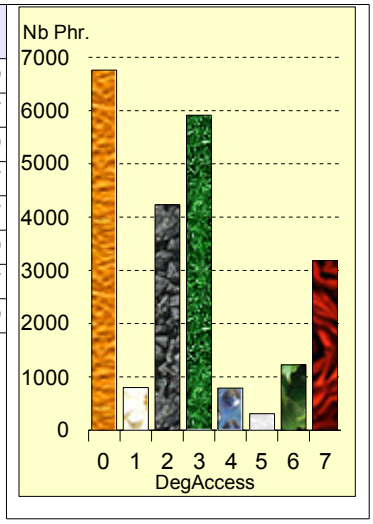


Tableau VIII.20 : Répartition des degrés d'accessibilité – DegAccess – dans le corpus

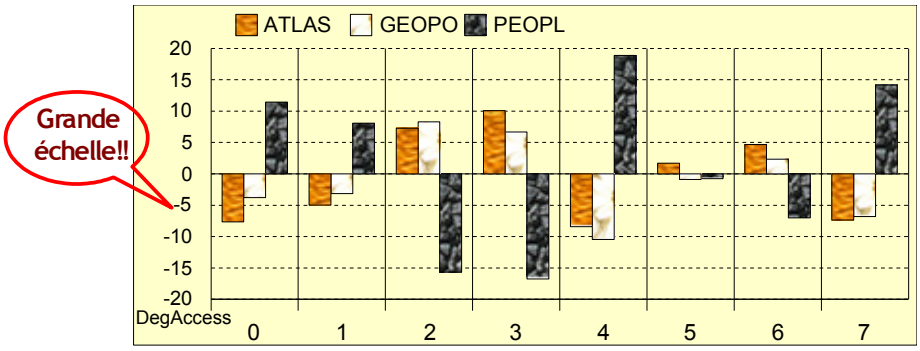
Le DegAccess_0 constitue le degré d'accessibilité le plus fréquent. Ce degré regroupe une grande variété de formes différentes dont la classe hétérogène des « formes autres »[170]. Les DegAccess_2 et 3 sont les plus fréquents après le DegAccess_0, ce qui ne nous surprend pas puisqu'ils correspondent aux descriptions définies. Le fait de regrouper descriptions courtes et descriptions avec reprise modifie la donne : il y a plus de SNdef 'co-référentiels'[171] – réduits ou présentant une reprise – que de SNdef complets et sans reprise. Mais ce regroupement ne modifie pas les schémas d'écart observés pour les Thèmes topicaux. Nous voyions déjà dans le tableau VIII.7 qu'ATLAS usait davantage de SNdef_R (et des reprises en général) que GEOPO. Cela se retrouve au DegAccess_3 où ATLAS présente un *z*=+10 et GEOPO un *z*=+7 (voir graphique VIII.9).

Les spécificités de chaque sous-corpus telles que décrites précédemment sont conservées. PEOPL s'oppose de façon générale à GEOPO et ATLAS. PEOPL affiche significativement plus de formes au degré d'accessibilité prétendu faible (DegAccess_0, *i.e.* les ThSpe et SN indéfinis et DegAccess_1, *i.e.* les noms propres sans reprise lexiclae ou

170 Les « formes autres » ne correspondent pas totalement aux « ThTop autre » puisque les pronoms indéfinis (rangés dans les ThTop autres) apparaissent ici avec un DegAccess_7.

171 Les guillemets rappellent que le terme 'co-référentiel' n'est pas à prendre au pied de la lettre. En effet, les caractéristiques utilisées pour identifier les SN 'co-référentiels' automatiquement sont beaucoup moins complexes que les définitions linguistiques de la co-référentialité.

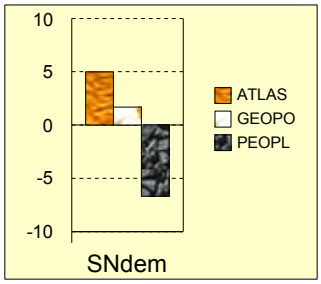
posiitionnés en première phrase de section) ou très élevé (DegAccess_7 associé aux ProPoss). À ces deux extrêmes s'ajoute l'utilisation spécifique des formes au DegAccess_4, *i.e.* les redénominations par nom propre. ATLAS et GEOPO affichent des degrés plus intermédiaires, en opposition à PEOPL. Les écarts au niveau des DegAccess_2 et 3 reflètent bien l'emploi fort des SNdef, emploi amoindri dans PEOPL par celui des noms propres.



Graphique VIII.9 : Variations des degrés d'accessibilité - DegAccess - selon les sous-corpus

Suite aux variations selon les sous-corpus observées précédemment et en fonction de notre conclusion quant à une permutation entre SN définis chez ATLAS/GEOPO et noms propres chez PEOPL, il semble évident que cette attribution de degrés d'accessibilité ne fonctionne pas de la même façon pour PEOPL que pour ATLAS et GEOPO. Nous observons le même phénomène que celui décrit par Schnedecker (2005) : dans les portraits, sites mono-référentiels, le nom propre est une alternance au pronom qui peut, en plus, marquer un déplacement. Il serait donc judicieux d'adapter l'échelle d'accessibilité au type de texte considéré comme par exemple pour PEOPL, augmenter le degré d'accessibilité des noms propres répétés.

Les variations selon les sous-corpus sont toutes significatives, à l'exception du DegAccess_5 correspondant aux descriptions démonstratives complètes. En repensant au graphique VIII.2 (p.205, rappelé par le graphique VIII.10) qui ne distinguait pas de sous-catégories aux SN démonstratifs, nous remarquons que les SNdem 'sensibles au type de texte' sont en fait ceux qui constituent une description réduite et/ou avec reprise. Le DegAccess_6 représente 80% des SNdem en général. En d'autres termes, il est rare d'avoir une description démonstrative complète. Ce constat peut mettre en doute la nécessité de distinguer les degAccess_5 et 6 dans une échelle d'accessibilité.



Graphique VIII.10 : variation des SNdem selon les sous-corpus

Certaines inadéquations entre la variation textuelle et une échelle d'accessibilité fixe telle qu'établie par Ariel seront à nouveau observées au niveau des variations selon les positions textuelles, ce qui précisera la relation entre les descriptions définies et les noms propres dans PEOPL. La mesure du DegAccess est moins instructive pour comparer les sous-corpus que pour comparer les positions textuelles. En effet, le recours au degré d'accessibilité est davantage lié aux variations selon la position textuelle (avec l'hypothèse que plus on entre à l'intérieur des paragraphes, plus le degré d'accessibilité augmente).

VIII.7. Collocations entre INIT1 et ThTop/ThSpe

Observons maintenant la composition de la position initiale en prenant en compte la fonction d'INIT ou l'absence d'INIT **et** le type de ThTop/ThSpe, ce qui nous permet de représenter toutes les compositions possibles en position

initiale. Dans le tableau VIII.21 apparaissent toutes les collocations dont la fréquence d'apparition dépasse les 2% de phrases (*i.e.* 465 phrases). Ces collocations représentent 11,7% des 120 types de collocations différentes recensées.

Position initiale			Nb phrases	%
INIT1	ThTop	ThSpe		
.	SNdef	.	4 559	*19,6*
.	SNdef_R	.	2 315	*10,0*
.	PRO3	.	1 441	*6,2*
CIRC	SNdef	.	1 382	*5,9*
CIRC	SNdef_R	.	921	*3,9*
.	ThTop autre	.	764	*3,5*
.	.	Cliv	769	*3,3*
.	SNdem	.	755	*3,25*
.	SNindef	.	736	*3,0*
.	.	On....	584	*2,5*
CIRC	PRO3	.	541	*2,3*
.	NP	.	535	*2,3*
.	SNdem_R	.	513	*2,2*
.	NP_R	.	485	*2,1*

Tableau VIII.21 : Collocations générales en position initiale

Le tableau VIII.21 nous indique que la composition la plus fréquente est celle présentant un Thème topical de forme SNdef(_R) non précédé d'un INIT. Ce résultat est sans surprise puisque les SN définis avec ou sans reprise constituent les sujets grammaticaux les plus fréquents, tout comme les phrases sans INIT. Cette composition rassemble presque 30% des phrases (19,6% + 10%).

Les 5 premières lignes font apparaître des collocations présentant les éléments les plus fréquents pris isolément : l'absence d'INIT – abrégé ssINIT, les adverbiaux circonstanciels, les SNdef(_R) et les PRO3. En fait, la fréquence des collocations semble être directement liée aux éléments qui les composent :

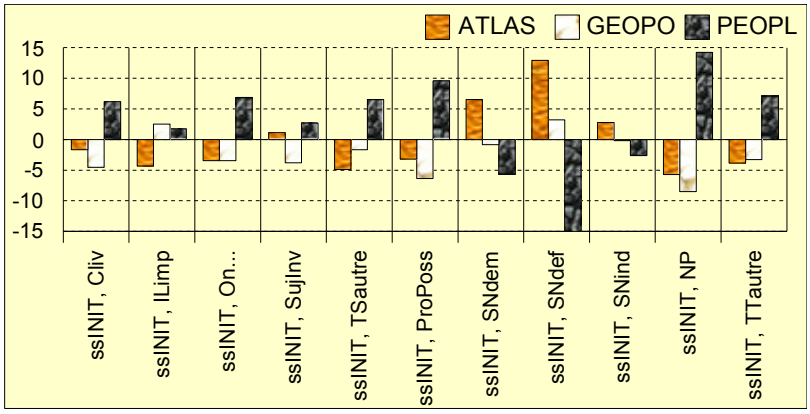
- en position INIT1, soit il n'y a pas d'INIT, soit il y a un adverbial circonstanciel ;
- en position sujet, nous retrouvons à peu près l'ordre affiché dans le tableau VIII.5 (p.204) : des ThTop de forme SNdef sans reprise, avec reprise, des PRO3, etc.

Aucune collocation surprenante n'apparaît. Cela signifie que l'ordre dessiné par les éléments isolés reste stable lorsque l'on considère les éléments dans leur apparition conjointe. Cependant, cela ne signifie pas que la présence/absence d'un certain type d'élément en INIT n'est aucunement corrélée à la présence/absence d'un certain type d'élément en ThTop/ThSpe. D'ailleurs, si nous appliquons le test du Khi² à nos données, nous observons une probabilité infinitésimale d'attribuer la collocation des INIT et des ThTop/ThSpe au hasard. Nous présentons justement ce type de corrélations dans le chapitre X.

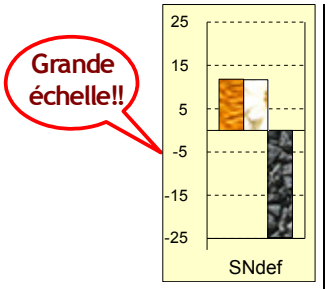
Parmi les 120 collocations recensées, 9 n'apparaissent qu'une seule fois sur les 23 217 phrases annotées. L'annexe F présente la liste des collocations n'apparaissant que très faiblement (moins de 23 fois, ce qui représente moins de 1% du corpus). Ces collocations rares se caractérisent soit par la présence d'un élément rare (argument inversé, TOP et dislocation), soit par la présence d'un élément non défini. Dans le dernier cas, l'indéfinitude de la catégorie *autre* va de pair avec une grande diversité d'emploi. Cela explique que, malgré le nombre de phrases présentant un ThTop/ThSpe « autre » (9,5%), plus de la moitié des collocations impliquant un ThTop/ThSpe « autre »

(précisément 10 sur 18 types recensés) apparaissent moins de 20 fois. Le constat est encore plus évident pour les INIT « autres » qui s'affichent tous dans des collocations rares (18/18).

Si l'on considère maintenant les différents sous-corpus, beaucoup de variations significatives apparaissent. PEOPL affiche 34 variations significatives dont 21 positives, GEOPO 27 dont 14 positives et ATLAS 27 dont 9 positives. Ces variations sont très souvent liées aux caractéristiques isolées de chaque élément dans chaque sous-corpus. Nous retrouvons les associations vues précédemment : apposition, pronom de 3e personne, nom propre et construction spéciale dans PEOPL ; adverbial modalisateur, construction impersonnelle dans GEOPO ; adverbial spatial dans ATLAS. Les graphiques suivants font apparaître tous les écarts significatifs observés. Pour plus de lisibilité nous avons séparé les écarts concernant les 'collocations' avec INIT (graphiques VIII.13 et VIII.14) de celles sans INIT – ssINIT (graphique VIII.11).



Graphique VIII.11: Écarts significatifs selon les sous-corpus pour des différentes collocations sans INIT



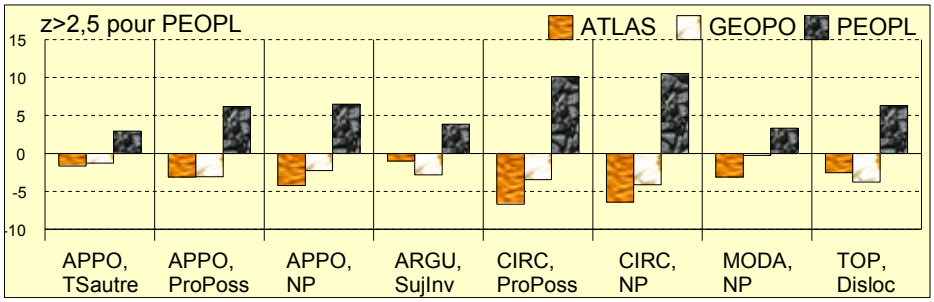
Graphique VIII.12: Variations des SNdef en ThTop selon les sous-corpus

Les écarts observés au niveau des phrases sans INIT sont tout à fait similaires aux écarts observés de façon générale pour les différents ThTop et ThSpe, sauf pour les ThTop de forme SNdef, où l'on passe d'un schéma de dissociation exclusive avec PEOPL (extrait du graphique VIII.2 p.205 repris ci-contre dans le graphique) à un schéma de dispersion, GEOPO passant d'un écart réduit de *z*=+11,7 (égal à celui observé dans ATLAS) à *z*=+3,2. Cette forte diminution de l'écart ne se retrouve pas dans ATLAS ou *z(ssINIT-SNdef)*=+13. Comme l'indique le graphique VIII.14, GEOPO fait précéder les SNdef d'un adverbial (circonstanciel ou modalisateur) là où ATLAS utilise plus fréquemment un SNdef seul.

En dehors de cette exception, nous retrouvons les écarts observés en général, et notamment l'association entre PEOPL et la plupart des constructions spéciales. Cela se traduit par des écarts positifs associés à PEOPL pour la majorité des configurations sans INIT avec Thème spécifique (entre +2,72 pour les sujets inversés et +6,9 pour les constructions en On...). Ces écarts positifs ne s'observent plus lorsque les Thèmes spécifiques sont précédés d'un INIT. Aucune collocation avec construction spéciale n'est présente dans les graphiques VIII.13 et VIII.14 sauf celles – logiques – entre un argument mis en initiale et une construction à sujet inversé ou entre un TOP et une dislocation. Les écarts observés pour les collocations de type ssINIT-x restent très importants par rapport à ceux observés pour les autres collocations (les graphiques VIII.11, VIII.13, VIII.14 sont à la même échelle).

Les graphiques VIII.13 et VIII.14 représentent les écarts significatifs positifs observés dans GEOPO et dans PEOPL. Il n'y a pas de graphique concernant ATLAS car les seuls écarts significatifs positifs observés dans ATLAS se trouvent soit au niveau de phrases sans INIT (graphique VIII.11), soit au niveau de la collocation CIRC-SNdef, indiquée dans le graphique concernant GEOPO (graphique VIII.13). Les écarts observés dans ATLAS sont en fait des écarts

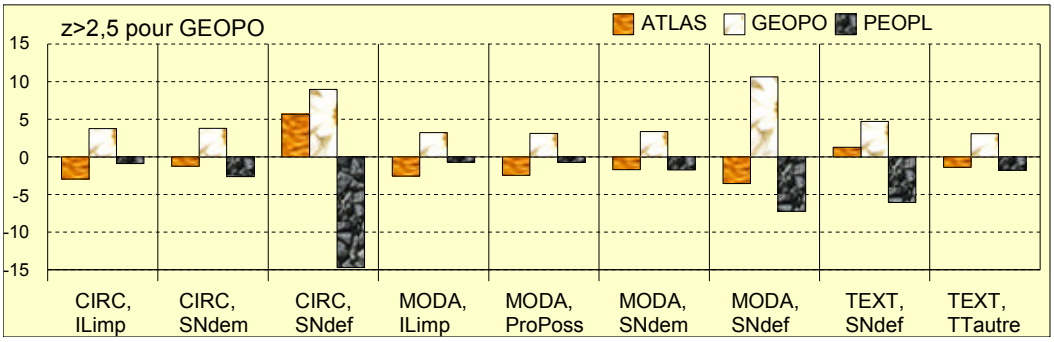
négatifs correspondant aux écarts positifs observés dans GEOPO ou PEOPL. Les variations sont classées selon le type d'INIT présent dans la collocation concernée. Les Thèmes topicaux avec reprise ne sont pas distingués de ceux sans reprise – la dénomination SNdef représente à la fois les SNdef et les SNdef_R – et la catégorie ProPoss rassemble les PRO3, PROdemo et SNposs.



Graphique VIII.13: Écarts selon les sous-corpus des différentes collocations pour lesquels z(PEOPL)>2,5

PEOPL constitue le sous-corpus présentant les écarts significatifs les plus importants. Les schémas d'écart dans lesquels il est significativement positif sont généralement des schémas d'association exclusive, excepté les schémas d'éclatement observés pour les collocations APPO-NP, ARGU-SujInv et MODA-NP. Les éléments associés positivement à PEOPL (APPO, NP et ProPoss) se retrouvent mêlés les uns aux autres ou, pour les éléments ThTop, en collocation avec le type d'INIT le plus fréquent : les adverbiaux circonstanciels.

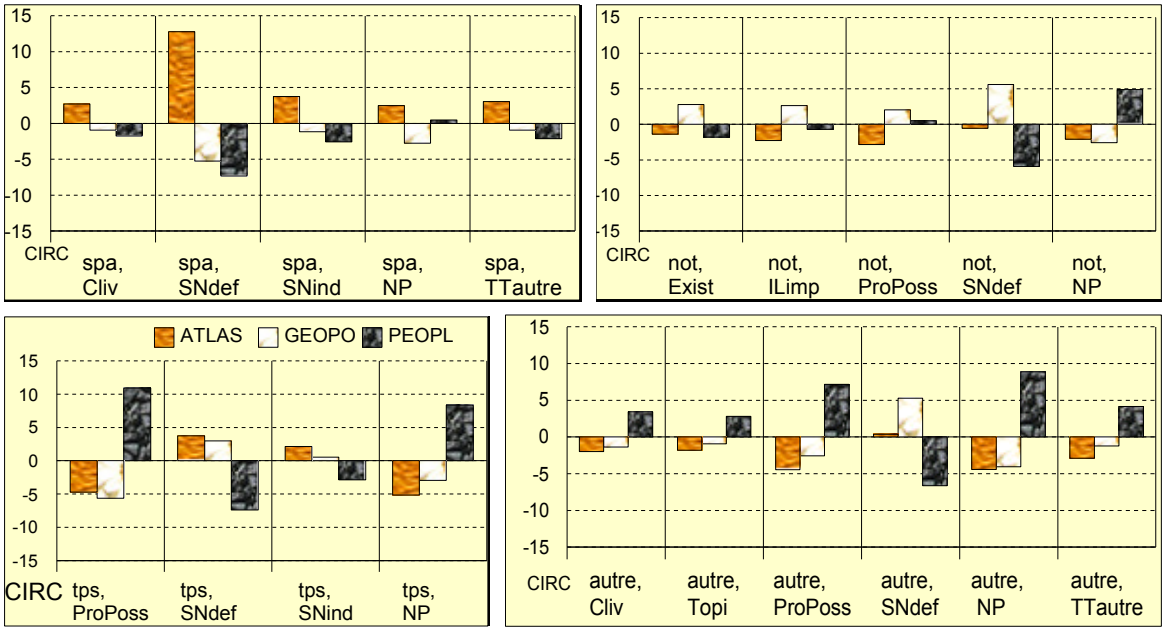
La collocation entre un adverbial circonstanciel et un Thème topical de type pronominal ou possessif (ProPoss) montre un *z*=+10,1 et celle entre un circonstanciel et un nom propre affiche un *z*=+10,5. La 'collocation' la plus significative dans PEOPL, si l'on peut parler de collocation, correspond à la présence d'un nom propre seul non précédé d'un INIT1 (*z*=+14,2). Comme attendu, les collocations significativement moins présentes dans PEOPL sont celles impliquant un SNdef en ThTop (*z(ssINIT-SNdef_R)*=-16 ; *z(CIRC-SNdef)*=-14,7).



Graphique VIII.14: Écarts selon les sous-corpus des différentes collocations pour lesquels z(GEOPO)>2,5

Pour GEOPO, ce sont les collocations comprenant un adverbial modalisateur qui affichent un écart positif significatif (*z(MODA-SNdef)*=+10,6). Celles constituées d'un adverbial circonstanciel sont également significativement plus présentes dans ce sous-corpus (*z(CIRC-SNdef)*=+9).

Si l'on considère les rôles sémantiques des circonstants, nous remarquons que seuls les rôles sémantiques spatial et notionnel modifient les variations observées au niveau des collocations générales CIRC-x. ATLAS est exclusivement associé aux CIRCspa. Il est le seul corpus présentant des écarts positifs pour ce rôle sémantique. GEOPO affiche presque la même exclusivité avec les CIRCnot, sauf lorsque ceux-ci sont suivis d'un NP. Dans ce dernier cas (CIRCnot+NP), c'est PEOPL qui affiche un écart significatif.



Graphiques VIII.15: Écarts significatifs selon les sous-corpus pour les différents rôles de CIRC en collocation

Les associations entre PEOPL et les collocations présentant un ProPoss ou un nom propre se retrouvent au niveau des circonstants temporels, notionnels et de catégorie indéfini. À l'opposé, la dissociation entre PEOPL et les SNdef se retrouve pour tous les rôles circonstanciels. Cet écart négatif correspond :

(i) à une dissociation exclusive pour la collocation CIRCtps-SNdef ;
(ii) au reflet de l'association exclusive avec ATLAS pour la collocation CIRCspa-SNdef ;
(iii) à une opposition vis-à-vis de GEOPO dans un schéma de dispersion, pour les collocations CIRCnot-SNdef et CIRCautre-SNdef.

Les collocations contenant un CIRCtps ou un CIRCautre suivent les variations générales observées entre sous-corpus selon le type de ThTop ou de ThSpe. On retrouve surtout les différents écarts observés au niveau de PEOPL (sous-corpus qui présente le plus de variation), *i.e.* plus de ProPoss et de nom propre et moins de SNdef.

La multitude de collocations possibles et des écarts significatifs observés rend certainement la lecture de cette partie douloureuse et nous nous en excusons. Cependant, cette partie est nécessaire pour comprendre et souligner les associations dominantes entre un sous-corpus et un élément en position initiale. Ainsi, il semble que les CIRCtps et les CIRCautre n'ont pas d'influence sur le type de Thème topical qui le suit. En effet, les Thèmes topicaux qui cohabitent avec ces circonstants restent associés significativement au même sous-corpus que celui auxquels ils sont associés sans tenir compte du rôle sémantique ou de la fonction d'INIT. À l'opposé, les adverbiaux spatiaux, notionnels, modalisateurs et l'absence d'INIT modifient les variations observées d'un point de vue général.

Nous verrons en détail certains cas particuliers de collocations dans le chapitre X. Ce sera alors l'occasion de revenir de façon plus précise et sans doute plus claire sur les données présentées ici.

VIII.8. Récapitulatif de la distribution générale et par sous-corpus des éléments en position initiale

Cette description générale de la composition de la position initiale fait apparaître différents patrons plus ou moins fréquents. Le patron de position initiale le plus fréquemment rencontré est la présence seule d'un ThTop. En d'autres termes, la plupart des phrases commencent directement par le sujet grammatical et ne présentent pas de constructions spéciales. Nous retrouvons donc le schéma typique : Sujet-Thème + Prédicat-Rhème. En précisant la nature du Thème topical, nous arrivons au patron : SNdef + Prédicat.

Si l'on considère chaque élément séparément, tous ont leur nature qui varie significativement selon le sous-corpus. Au niveau des INIT, aucune fonction n'est insensible au sous-corpus (sauf la fonction indéfinie INITautre). Au niveau des Thèmes topicaux, seuls les pronoms démonstratifs et les SN indéfinis ne varient pas significativement selon le sous-corpus. De fait, chaque sous-corpus affiche des spécificités dans sa composition de la position initiale. Nous verrons dans le chapitre suivant que la plupart de ces spécificités se complexifient selon les positions textuelles, sauf pour ce qui est des spécificité liées aux constructions à Thème spécifique. Le tableau VIII.22 résume l'ensemble des spécificités observées et pertinentes pour la suite des analyses.

Deux types de schémas d'écart mettent en évidence les spécificités d'un sous-corpus : les schémas d'association/dissociation exclusive et les schémas de dispersion. Nous rappelons qu'un schéma de dispersion s'observe lorsque la répartition de l'indice concerné n'est absolument pas stable à travers les types de texte représentés par nos trois sous-corpus. Le tableau VIII.22 représente cette information par l'utilisation des + et des -. Ainsi, un élément spécifique à un sous-corpus est précédé de deux signes ++ et un élément significativement absent d'un sous-corpus est précédé de deux signes --. Cet élément n'est pas mentionné dans les deux sous-corpus non concernés par la spécificité ou l'absence significative. Lorsqu'il y a dispersion, les éléments sont précédés d'un seul signe. Les éléments ne montrant pas de variation significative dans un sous-corpus ne sont pas mentionnés.

	spécificité	ATLAS	GEOPO	PEOPL
Patron		++ThTop seul --Connect+ThTop	+INIT+ThTop	-INIT+ThTop ++ThSpe
Connecteur		-Connect		+Connect
INIT1	**Catégori morpho-syntaxique / fonction**	CIRC/SP APPO/PPA	APPO/PPR	CIRC/SUB APPO/SN
	Fonction	+CIRC	-CIRC -ARGU ++MODA +TEXT	 ++APPO +ARGU -TEXT ++TOP
	Rôle sémantique	++spa -autre	++not	 +autre
ThTop	**Forme**			++NP ++PRO3 --SNdef
	coréférence	+_R -SNdef courts		ProPoss -_R +SNdef courts
ThSpe		+Cliv +SujInv	-Cliv +Present ++ILimp -SujInv -Disloc	 -Present +Disloc

Tableau VIII.22 : Récapitulatif des spécificités des sous-corpus en position initiale

Chapitre IX
Des positions textuelles influentes : initiales de sections, de paragraphes et de phrases

Sommaire

Une des hypothèses fortes de ce travail consiste à poser que ce qui est mis en position initiale varie selon le niveau de segmentation. Nous avons donc dégagé trois positions textuelles : S1 (position initiale de sections), P1 (position initiale de paragraphes) et P2 (position initiale de phrases intraparagraphiques), voir la partie IV.5. Nous supposons qu'*a priori*, les changements de sections et de paragraphes sont à même de signaler une discontinuité dans le discours. En S1, cette discontinuité serait forte, s'associant à des phénomènes de rupture. En P1, elle serait mélangée à de la continuité s'associant ainsi à des phénomènes de déplacement. À l'opposé, la position P2, qui ne comporte pas d'autre indication que le simple changement de phrase, serait associée à de la continuité.

De fait, des corrélations d'indices de rupture se rencontreraient spécifiquement en S1, des corrélations d'indices de déplacement en P1 et des indices de continuité en P2 :

PosTxt	Phénomène	Connect	INIT	ThTop/ThSpe
S1	rupture	pas de connecteur	INIT participant à une TSC globale	introduction d'un nouveau topique (ThSpe ou ThTop sans reprise)
P1	déplacement	?	INIT participant à une TSC (globale ou locale)	redénomination du topique, reclassification, certaines ThSpe
P2	continuité	connecteur	Pas d'INIT	continuation

Tableau IX.1 : Hypothèses quant aux traits propres aux différentes positions textuelles

Bien entendu, nous conservons notre regard comparatif entre sous-corpus. Pour faciliter la comparaison des écarts entre positions textuelles selon les différents sous-corpus, nous utilisons exclusivement les représentations graphiques qui permettent, en un regard, de se faire une idée sur les différences et similarités de comportement entre nos trois sous-corpus. Lorsqu'il y a similarité entre les trois sous-corpus, nous avons généralement opté pour ne pas représenter les schémas d'écart afin d'alléger le graphique et la page. Nous avons ainsi choisi de n'afficher que les schémas d'écart pertinents, *i.e.* indiquant une spécificité d'un ou deux sous-corpus. Ce choix est totalement influencé par un souci de lisibilité. Aussi, si la mise en forme matérielle le permet, tous les schémas d'écart sont représentés.

D'un point de vue général, nous remarquons que les variations entre positions textuelles sont moins importantes que celles observées précédemment entre sous-corpus. Ainsi, aucun observable ne présente d'écart réduit supérieur à 10, qu'il soit négatif ou positif, alors qu'entre sous-corpus, nous avons observé des écarts réduits allant jusqu'à -25. Cette remarque générale implique que la différence entre genres est très importante et que l'on peut difficilement décrire une organisation du discours sans prendre en compte, avec finesse, la nature des textes analysés.

IX.1. Patrons de position initiale selon les positions textuelles – PosTxt

IX.1.1. Des éléments associés aux différentes positions textuelles

Au niveau de la composition générale, nous observons une répartition significativement différente des éléments détachés en initiale, Thèmes topicaux (et par reflet, des Thèmes spécifiques[172]) et connecteurs selon les différentes positions textuelles :

	Corpus		%		
	%	Nb	S1	P1	P2
ThTop	*83,4*	19 360	*87,0*	*85,2*	*82,7*
ThSpe	*16,6*	3 857	*13,0*	*14,8*	*17,3*
Connect	*9,6*	2 220	*1,4*	*6,7*	*10,8*
INIT	*30,2*	7 022	*33,6*	*33,3*	*29,3*
Nb. Phr.		23 217	1 165	4 118	17 954

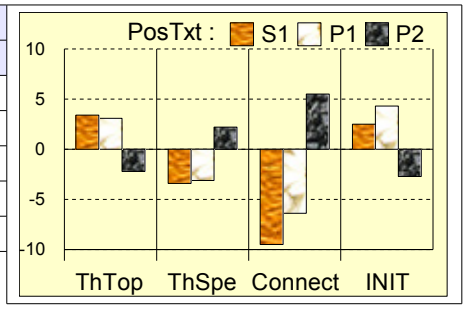


Tableau IX.2 : Variations de la composition générale de la position initiale selon les différentes PosTxt : S1, P1 et P2

172 Les mesures effectuées au niveau des Thèmes topicaux s'inversent au niveau des Thèmes spécifiques, puisque ces deux éléments s'excluent. Ainsi, lorsque l'on dénombre 87% de ThTop en S1, cela signifie qu'il y a 13% de ThSpe en S1. Pour les écarts réduits, il suffit d'inverser la polarité. Ainsi, si l'on observe un z=+3,4 pour les ThTop en S1, on a z=-3,4 pour les ThSpe en S1. Le graphique du tableau IX.2 illustre cet effet de reflet.

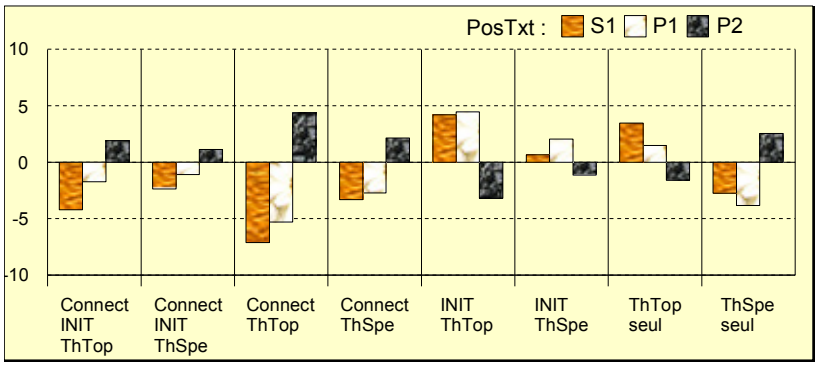
Il semble effectivement que les connecteurs soient associés très fortement à la position intraparagraphique P2. Les cas de connecteurs commençant une section ou un paragraphe sont très rares : 1,4% des phrases commençant une section (16 des phrases S1) et 6,7% des phrases commençant un paragraphe (274 des phrases P1). Nous retrouvons ici la fonction généralement attribuée aux connecteurs : articuler deux propositions.

Les Thèmes spécifiques affichent un comportement similaire à celui des connecteurs, mais avec des écarts plus faibles. Seuls les écarts négatifs en S1 et P1 sont significatifs (en S1, *z*=-3,4 et en P1, *z*=-3,1). Ces résultats tendent tout de même vers l'idée d'un positionnement intraparagraphique des constructions spéciales et non en début de paragraphe ou de section, comme on aurait pu le présager. Nous verrons en VIII.5 que ce comportement dépend grandement du type de construction considérée.

La présence d'un élément détaché en initiale varie différemment : un écart positif net en P1 et un écart négatif en P2. Cette présence semble bien accompagner un déplacement tel que l'indique le changement de paragraphe et non une continuité plus appropriée aux relations entre phrases intraparagraphiques. Au niveau du changement de section, l'écart est certes positif mais à la limite du significatif (+2,5). Il y a donc plus d'INIT en S1 qu'en P2, mais dans des proportions moindres qu'en P1.

Les trois patrons les plus fréquents ([ThTop seul] > [INIT+ThTop] > [ThSpe seul]) restent les mêmes quelle que soit la position textuelle envisagée. En observant uniquement les fréquences d'apparition, seul le patron [Connect+ThTop] apparaît moins fréquemment en S1 et P1 qu'en P2. Il semble effectivement difficile de commencer une section par un connecteur ! Un paragraphe peut toutefois commencer par un connecteur, mais dans des proportions très inférieures aux phrases intraparagraphiques. Cela tend à montrer que les connecteurs ont un rôle cohésif très local[173]. Nous observons le même phénomène pour le patron [ThSpe seul] qui présente un écart négatif pour S1 et P1 et un écart positif pour P2 (*z(S1)*=-3,4, *z(P1)*=-3,1 et *z(P2)*=+2,3).

En comparant les différents patrons selon leur écart réduit par position textuelle, presque tous affichent des variations significatives (graphique IX.1). La plus forte variation correspond bien au **patron [Connect+ThTop]**, ce qui confirme l'association de ce patron aux phrases intraparagraphiques, comme vu précédemment. Mais d'autres variations nous permettent d'associer certains patrons aux différentes positions textuelles. Ainsi, les premières phrases de section favorisent les patrons **[ThTop seul]** et **[INIT+ThTop]**. Ces deux patrons représentent 86% des phrases en S1. Tous les autres patrons évitent significativement les phrases en S1, exception faite des **[INIT+ThSpe]**. Ce dernier est d'ailleurs le seul à n'afficher aucune variation significative dans aucun sous-corpus.



Graphique IX.1 : Écarts des patrons de position initiale selon les différentes PosTxt

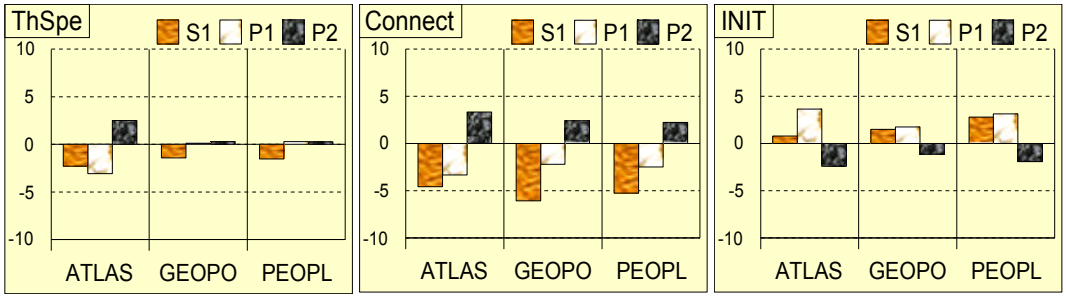
173 Un exemple de connecteur à rôle 'local'*vs.* 'global' est donné en partie I.1.2.

Du côté des phrases en **P1**, seul le patron **[INIT+ThTop]** est significativement plus présent, nous retrouvons les écarts observés dans le tableau IX.2. Les patrons impliquant une construction à Thème spécifique sont soit insensibles à la position textuelle, soit significativement moins présentes en initiale de paragraphes.

Les phrases en **P2** présentent l'opposée de celles en S1 et P1 : plus de connecteurs (notamment le patron [Connect+ThTop] avec *z*=+4,4), plus de Thèmes spécifiques (le patron [ThSep seul] affiche un *z*=+2,5) et moins d'éléments détachés suivis d'un Thème topical, le patron [INIT+ThTop] étant caractéristique des positions S1 et P1.

IX.1.2. Des variations entre positions textuelles dans chaque sous-corpus

Si l'on observe maintenant les différentes positions textuelles selon les sous-corpus, quelques nouvelles particularités apparaissent, qui restent dans des proportions relativement faibles. Les écarts observés dans le corpus entier ne se retrouvent pas systématiquement au niveau de chaque sous-corpus. Le graphique IX.2 nous montre des sous-corpus aux différents comportements selon le type d'élément considéré.



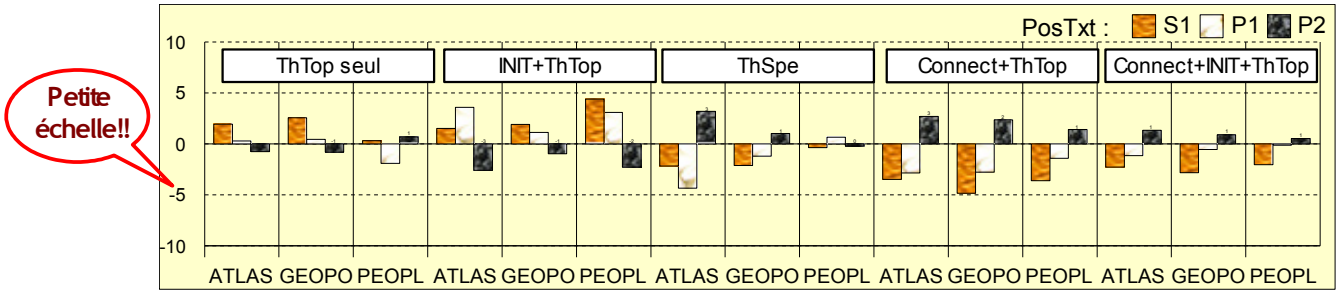
Graphique IX.2 : Écarts réduits des PosTxt par sous-corpus pour les éléments ThSpe, Connect et INIT

Le diagramme central indique que la distribution selon la position textuelle des phrases présentant un **connecteur 'pur'** ne varie pas selon les sous-corpus. Nous retrouvons effectivement les mêmes schémas d'écart (et qui plus est, dans des proportions égales) au niveau d'ATLAS, de GEOPO, de PEOPL et du corpus entier : moins de connecteurs en S1 et en P1 et plus en P2. La proportion de connecteurs en S1 et P1 est si faible que nous sommes incapables d'observer des variations entre positions textuelles au niveau des différentes formes de connecteur. Les spécificités de chaque sous-corpus observées dans le chapitre précédent sont à associer principalement aux spécificités en position intraparagraphique.

Au niveau des **éléments détachés en initiale** et des **constructions à Thème spécifique**, c'est ATLAS qui montre le plus de particularités, GEOPO et PEOPL présentant à peu près les mêmes schémas d'écart. ATLAS semble réserver ses éléments détachés pour les premières phrases de paragraphe, alors que PEOPL et GEOPO suivent le modèle général : plus d'INIT en S1 et en P1. Du côté des Thèmes spécifiques, GEOPO et PEOPL montrent exactement le même comportement, c'est-à-dire aucun écart réduit significatif. En revanche, ATLAS affiche une répartition particulière. En effet, dans ce sous-corpus, les ThSpe se situent précisément au niveau intraparagraphique, comme l'illustre l'exemple IX.1.

(IX.1) *Dans une perspective de développement régional, les zones frontalières terrestres constituent un objectif particulier de rééquilibrage des disparités régionales, visant à combler le déficit de liens ou de développement liés à la position périphérique.* ***C'est particulièrement le cas pour les zones frontalières du sud de l'Europe.*** *Dans une perspective d'intégration accrue et de dynamisation de l'Union Européenne, les zones frontalières constituent des points d'appui de démultiplication d'activités, d'échanges et d'initiatives dans de nouvelles aires fonctionnelles auxquelles l'histoire des États Nations n'avaient pas laissé l'occasion de se développer à l'époque contemporaine.* ***C'est particulièrement le cas pour les zones frontalières du nord de l'Union Européenne****. Ici* ***c'est moins le manque de développement ou de densités (...) que le déficit encore existant de dynamiques communes de part et d'autre des frontières.*** *[ATLAS_1]*

Si l'on observe maintenant les huit patrons répertoriés, nous voyons que presque tous présentent des variations significatives selon les sous-corpus, même si ces variations restent dans des écarts faibles (les écarts réduits ne dépassent pas +/- 5). Les patrons **[INIT+ThSpe], [Connect+ThSpe], [Connect+INIT+ThSpe]** qui montraient les écarts les plus faibles dans le graphique IX.2 ne montrent aucun écart réduit significatif si l'on considère les sous-corpus séparément (c'est pourquoi nous ne présentons pas leurs schémas d'écart dans le graphique IX.3). Les patrons **[Connect+ThTop], [Connect+INIT+ThTop]** ne varient pas selon les sous-corpus : les répartitions selon les différentes positions textuelles restent les mêmes quel que soit le type de texte considéré. Cette invariabilité est à rapprocher de celle observée pour les phrases présentant, de façon générale, un connecteur (tableau IX.2).



Graphique IX.3: Écarts réduits des PosTxt par sous-corpus pour les différents patrons

En dehors des patrons avec connecteur, il reste trois patrons qui montrent des variations significatives selon la PosTxt et dont les schémas d'écart varient eux aussi selon les trois sous-corpus : les patrons [ThTop seul], [INIT+ThTop] et [ThSpe seul].

Les patrons **[INIT+ThTop]** et **[ThSpe seul]** sont ceux qui montrent le plus d'écart selon la position textuelle dans les différents sous-corpus. Ainsi, concernant les **[ThSpe seul]**, seul ATLAS montre des écarts significatifs indiquant que ces patrons sont beaucoup plus situés à l'intérieur des paragraphes, en P2, qu'en initiale de paragraphes, en P1. Pour GEOPO et PEOPL, les écarts ne sont pas significatifs, *i.e.* les [ThSpe seuls] sont indifféremment situés en S1, P1 ou P2. Au niveau des **[INIT+ThTop]**, ATLAS et PEOPL montrent une association similaire au niveau des P1 et une association différente au niveau des S1 (GEOPO affiche un schéma d'écart neutre). Cela signifie que dans ATLAS comme dans PEOPL, il est très fréquent de commencer un nouveau paragraphe par un Thème topical précédé d'un élément détaché. Par contre, au niveau des initiales de sections, ce n'est que dans PEOPL que ce patron est significativement plus présent.

Au niveau du **patron [ThTop seul]**, seul GEOPO affiche une légère association entre ce patron et les initiales de sections. ATLAS et PEOPL ne montrent ici pas de variations significatives. L'exemple suivant illustre un début de section par le patron [ThTop seul], dans GEOPO (nous recensons 190/348 débuts de section similaires).

(IX.2) ***L'ÉLABORATION D'UN CADRE DE LUTTE CONTRE LE TERRORISME*** [titre niveau 1]
Le rapide panorama des mesures d'urgence que nous venons d'établir a pris place dans un cadre légal établi à l'automne 2001, puis complété par [...] *[GEOPO_1]*

Comme l'illustre cet exemple, les [ThTop seul] en début de section sont généralement des descriptions longues et complètes, ce que nous verrons plus en détail en IX.2.4.b.

IX.2. Répartition des Thèmes topicaux selon la position textuelle

	corpus entier		%*		
	Nb	%*	S1	P1	P2
SNdef	10 143	52,4	69,2	62,0	49,0
PRO3	2 221	11,5	2,6	4,2	13,7
Ttautre	1 675	8,7	6,8	7,3	9,1
NP	1 592	8,2	11,7	8,6	7,9
SNdem	1 538	7,9	4,5	8,4	8,1
SNindef	1 447	7,5	4,8	8,0	7,6
SNposs	435	2,2	0,4	1,0	2,7
PROdemo	309	1,6	0,1	0,5	2,0
		100	100	100	100
ThTop	19 360	83,4	1 007(86,5%)	3 507(85%)	14 839(83%)
Nb Phr.	23 217		1 165	4 118	17 934

* la colonne % affiche la proportion pour les ThTop et la colonne %'affiche la proportion pour le corpus entier.

Tableau IX.3 : Répartition des différents types de ThTop selon la position textuelle

Pour présenter les données (tableau IX.3), nous avons choisi l'ordre décroissant observé sur le corpus entier comme dans le chapitre précédent. Si l'on observe ce qui se passe au niveau des différentes positions textuelles, nous voyons que l'ordre décroissant n'est pas respecté, mis à part pour les SN définis qui sont toujours les types de Thème topical les plus fréquents. La position P2 est celle qui se rapproche le plus de la moyenne générale (ce qui est assez normal étant donné que les phrases en P2 représentent plus de 77% de toutes les phrases du corpus). Ainsi, l'ordre décroissant y est quasiment le même qu'en général, sauf pour les SN démonstratifs qui, au niveau de P2 seulement, sont plus fréquents que les noms propres – NP. Pour les phrases commençant une section, il y a de grosses variations qui seront illustrées dans la sous-section suivante. Cependant, nous voyons bien que les pronoms de 3e personne (PRO3) ne sont plus du tout la deuxième forme la plus fréquente de ThTop, remplacée, en initiale de sections, par les noms propres. Cette interversion entre les PRO3 et NP se retrouve également en P1. Ainsi, ce n'est qu'en position intraparagraphique, position typique de continuation, que le pronom *il*, indice typique de continuation, apparaît en force. En initiale de sections ou de paragraphes, son emploi semble plus contraint.

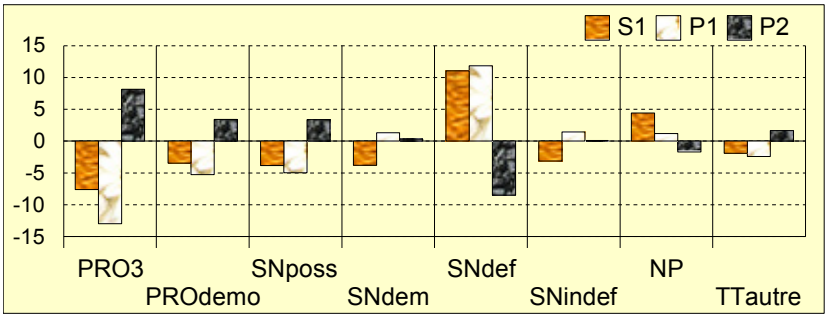
IX.2.1. Forme des Thèmes topicaux et degré d'accessibilité

Les variations observées au niveau des fréquences dans les phrases initiales de sections se retrouvent dans le fait que S1 montre le plus d'écarts significatifs et les écarts les plus forts. À l'opposé, les initiales de phrases intraparagraphiques affichent peu d'écarts significatifs. P1 se situe entre ces deux positions textuelles. Le graphique IX.4 indique que les formes les plus sensibles à la position textuelle sont les mêmes que les formes les plus sensibles au type de texte*, i.e.* les PRO3 et les SNdef. Par contre, nous ne retrouvons pas les noms propres. Nous verrons que dans PEOPL, les noms propres montrent une répartition particulière. D'un point de vue général, il se dessine une combinatoire entre P2 et PRO3 (auxquels on peut adjoindre les autres pronoms et les SN possessifs) d'une part et entre S1/P1 et SNdef d'autre part.

Au niveau des **PRO3**, l'écart négatif plus faible en S1 qu'en P1 semble paradoxal. Cette absence moins forte de PRO3 en S1 vient du fait que plusieurs sections de ATLAS et PEOPL commencent par un PRO3 reprenant, chez PEOPL, le nom de la personnalité objet du portrait et, chez ATLAS, le référent introduit en titre de section (des

exemples sont donnés en X.5). GEOPO ne comporte qu'une seule phrase en S1 dont le ThTop est un PRO3. Cette phrase est données en (IX.3). On y voit le référent antécédent du PRO3 introduit dans l'adverbial circonstanciel détaché.

(IX.3) ***"NI GUERRE, NI PAIX"*** [titre niveau 1]
Les implications internes [titre niveau 2]
Bien que ***la prolongation pour une période indéfinie de cette situation de "ni guerre, ni paix"*** *convienne à la majeure partie de l'establishment syrien,* ***elle*** *n'augure rien de bon à moyen et long termes pour les perspectives de développement du pays. [GEOPO_5]*



Graphique IX.4: écarts dans la forme des ThTop selon la PosTxt

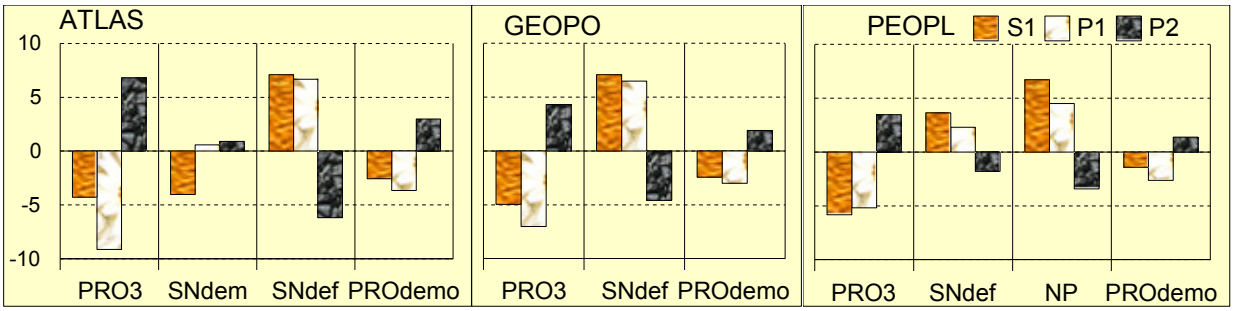
Ces données montrent bien qu'à un niveau global, GEOPO et ATLAS n'ont pas de continuité topicale comme PEOPL et que la fonction discursive des titres dans ATLAS est bien différente de celle dans GEOPO et PEOPL (voir partie X.5). Dans les chiffres, nous avons 26 PRO3 en S1 (11 dans ATLAS, 14 dans PEOPL et 1 seul dans GEOPO) et 149 en P1 (35 dans ATLAS, 24 dans GEOPO et 90 dans PEOPL).

Mis à part les SNdef, il n'y a que les NP qui se montrent significativement plus en S1. Alors que nous nous attendions à significativement plus de SNindef en S1 (l'indéfini étant associé à un phénomène de rupture car indicateur d'un degré d'accessibilité nul), il semblerait qu'il n'est pas si fréquent de commencer une section avec un SNindef (z=-3,3). Moins de 5% des phrases avec ThTop en S1 (48/1 007) commencent par un SNindef, contre 7,5% en P2 (1 121/14 839).

IX.2.2. Des associations entre formes de ThTop et PosTxt différentes selon les sous-corpus

Les graphiques suivants permettent de comparer les variations observées en faisant varier à l'intérieur de chaque sous-corpus le facteur PosTxt. Les schémas d'écart neutres ne sont pas représentés, ce qui explique pourauoi les ThTop de forme SNposs, PROdemo, SNindef et SNautre n'apparaissent pas, leur schéma d'écart étant neutre dans les trois sous-corpus. Les SNdem et NP n'apparaissent, eux, que dans les sous-corpus où ils affichent des variations significatives.

De façon générale, nous assistons aux corrélations : SNdef/S1-P1 et PRO3/P2, ce qui était relativement attendu. On retrouve le rôle des NP spécifique chez PEOPL et ce rôle se précise. Il semble en effet que les NP dans PEOPL ont le même comportement que les SNdef. Alors que nos données précédentes nous faisaient croire en une substitution – là où ATLAS et GEOPO utilisaient un SNdef, PEOPL utilisait un NP – nous nous rendons compte que les NP ajoutent de la variation plus qu'ils ne remplacent la variation des SNdef. Ainsi, les SNdef, dans PEOPL comme dans les autres sous-corpus se dissocient de la position P2.



Graphique IX.5 : Schémas d'écart pertinents des différentes ThTop par PosTxt et sous-corpus

ATLAS est le sous-corpus qui affiche les plus grands écarts au niveau ThTop. Nous retrouvons au niveau des PRO3 ce que nous avons vu précédemment, *i.e.* un certain nombre de sections dans ATLAS commencent par une reprise pronominale du titre de section, ce qui rend l'écart négatif de PRO3 en S1 moins élevé qu'en P1. La dissociation observée dans le corpus entier entre les SNdem et la position S1 se retrouve exclusivement dans ATLAS. Dans ce sous-corpus, à peine 3,5% des phrases en S1 ont un SNdem en ThTop (contre un minimum de 5% dans les autres sous-corpus).

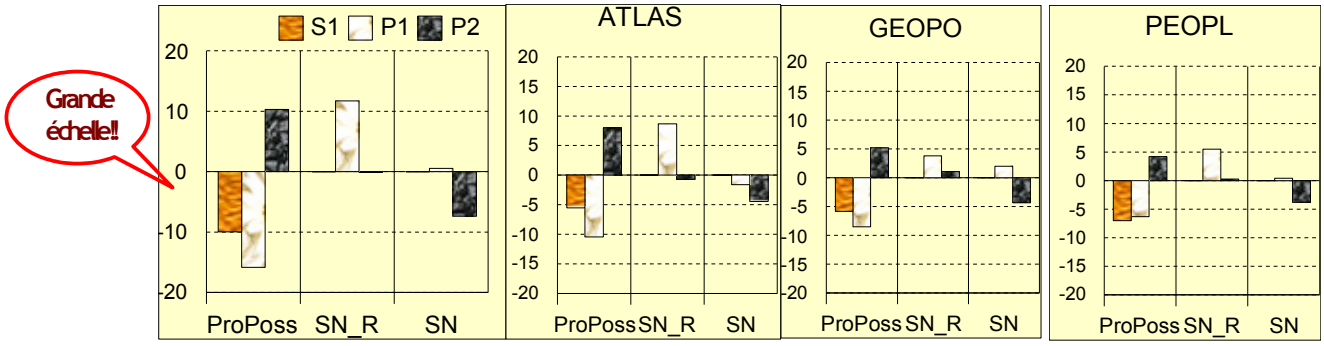
Les PRODemo sont les seules formes qui affichent le même schéma d'écart pour les trois sous-corpus. Il s'agit d'un schéma d'éclatement où S1 est dans la moyenne, P1 affiche significativement moins de PRODemo et P2 significativement plus. Les écarts sont plus importants dans ATLAS que dans PEOPL ou GEOPO.

IX.2.3. Coréférence en ThTop selon la position textuelle

IX.2.3.a) Pronominalisation : quelques cas troublants

Nous avons déjà noté que les ProPoss (pronoms et SNposs) apparaissent préférentiellement en P2 et cela, quel que soit le sous-corpus. Le graphique suivant nous montre une répartition claire des trois groupes de ThTop : les ThTop 'anaphoriques', les ThTop 'coréférentiels' et les ThTop pleins.

Les résultats de l'analyse vont totalement dans le sens de notre intuition : l'anaphore s'associe exclusivement à la continuité présumée de P2, la co-référence par reprise lexicale (redénomination ou réduction de terme) est très présente en P1 où le changement de paragraphe favorise (requiert?) une reprise lexicale du référent. Les ThTop pleins sont significativement moins présents en P2. Les graphiques IX.6 montrent bien que cette répartition est la même pour tous les sous-corpus. On voit effectivement les mêmes écarts significatifs pour tous les sous-corpus, même si ceux-ci sont moindres que ceux observés dans tout le corpus.



Graphiques IX.6 : Pronominalisation et reprise lexicale en ThTop selon les PosTxt

ATLAS est certainement le sous-corpus qui montre les plus fortes associations, notamment entre P1 et les SN_R. La seule petite différence entre sous-corpus se situe au niveau de la différence des écarts entre S1 et P1 au niveau des ProPoss. Alors que ATLAS et GEOPO montrent un écart négatif deux fois moins fort en S1 qu'en P1, PEOPL ne semble pas faire de différence, au niveau des ProPoss, entre ces deux positions textuelles. L'association entre P2 et ProPoss chez PEOPL est d'ailleurs moins forte que chez les deux autres sous-corpus. Cela peut certainement s'expliquer par deux choses. D'une part, le caractère mono-référentiel des textes de PEOPL permettent à la continuité topicale générale de se réaliser sous forme de pronom même en début de section ou de paragraphe (et même en début de texte, voir exemple en partie X.5). D'autre part, il y a la possibilité, dans PEOPL, d'employer en alternance des NP_R et des PRO3 pour maintenir la continuité topicale, comme le montre l'exemple suivant :

(IX.4) ***Léonard**, quittant l'atelier de Verrocchio, a pu être quelque temps au service de Laurent de Médicis; c'est ce qu'affirme l'Anonyme Gaddiano, [...]*
*En 1482-1483, **Léonard** est au service de Ludovic le More qui vient de s'emparer du duché de Milan (1480). **Il** devient le grand animateur de la cour. Après 1499, **il** cherche un autre protecteur princier : [...]. **Il** intéresse César Borgia pour la guerre en Romagne (1502), Charles d'Amboise pour l'architecture (projet de villa au bord du le Noviglio) et la décoration des demeures (à partir de 1506). Aux yeux des princes français comme à ceux de César Borgia, **Léonard**, si célèbre qu'il soit comme peintre, compte pour ses autres capacités. On est frappé aussi par la facilité avec laquelle l'artiste-ingénieur passe du service d'un protecteur à celui de son adversaire. **Il** revient à Milan avec les princes français qui ont chassé Ludovic ; à la fin de 1504, il est à Piombino, auprès de Jacoppo IV d'Appiano, qui, l'année précédente, avait été chassé par César Borgia, le patron de Léonard. Les grands esprits n'ont pas de camp. **Léonard** appartient à qui se l'attache et lui laisse un loisir pour l'étude. Paul Jove a été frappé de ses capacités comme organisateur de fêtes, musicien, etc., et conclut que ces aptitudes "l'ont rendu cher à tous les princes qui l'ont connu." En dehors des décors de théâtre ou de parade, **Léonard** a composé des rébus, constitué des recueils de devinettes et de fables, rédigé des devises, des [...]* [PEOPL_11]

Cet exemple montre bien comment la conjugaison des PRO3 et des NP_R permet de créer toute une zone de continuation autour du personnage de Léonard de Vinci. Cependant, comme on le voit ici, la présence d'un NP_R n'est jamais le fruit d'une redondance. En effet, à chaque fois que la redénomination « Léonard » apparaît, soit il y a eu activation d'un autre participant (l'Anonyme Gaddiano qui relate les faits retranscrits ici), soit il y a nécessité structurelle d'avoir un nom et non un pronom (avant l'incise appostive *« si célèbre qu'il soit comme peintre »*), soit il y a changement de style (« *Léonard appartient...* » expose l'avis du locuteur plus que ne narre l'histoire de Léonard). Ce n'est *a priori* que dans le dernier cas que le NP_R est indice de discontinuité. Cependant, nous pouvons également supposer que le deuxième NP_R a d'autres raisons d'être (l'adverbial circonstanciel de temps et le changement de paragraphe qui le précèdent).

IX.2.3.b) Reprise lexicale en SN selon la position textuelle

La distinction entre SN et SN_R n'est pas effectuée au niveau des S1, car la présence de reprise lexicale est impossible en S1. Non pas que linguistiquement il est impossible de commencer une section par une redénomination, mais parce que notre programme calcule la reprise (_R) à l'intérieur d'une section (voir VII.2.2.b). Nous considérons en effet que le titre de section remet les 'compteurs référentiels' à zéro. De plus, il était absolument nécessaire de délimiter la taille du contexte dans lequel chercher une reprise (élargir la taille du contexte à tout le texte aurait été non pertinent, nous avons déjà vu large en considérant toute la section). Si l'on avait indiqué les écarts en S1, un écart négatif pour les SN_R et un écart positif pour les SN auraient été observés, écarts reflétant notre stratégie d'annotation et non une analyse linguistique fiable. Les seules reprises observées en S1 sont celles concernant les éléments du titre (reprises décrites en X.5).

La proportion à présenter une reprise lexicale est plus importante lorsque le ThTop se situe en P1 (P2 reste dans la moyenne). Cette tendance ne se retrouve pas pour tous les types de ThTop.

Le tableau IX.4 indique les proportions des différents ThTop à présenter une reprise (_R)[174]. Entre P1 et P2, les différences de proportions soulignent la corrélation observée entre la position P1 et les SNdem_R, SNdef_R et NP_R. Ainsi, ces formes affichent plus de reprises en P1 qu'en P2, ce qui n'est pas le cas des SNindef et autres ThTop.

	Proportion à présenter une reprise _R (%)		
	P1	***P2***	***Corpus***
SNdem	*50*	*41*	*42*
SNdef	*42*	*37*	*35*
SNindef	*30*	*29*	*29*
NP	*60*	*52*	*50*
TTautre	*32*	*39*	*36*
ToutThTop	*43*	*38*	*37*

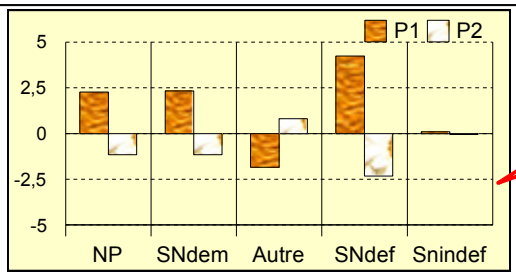
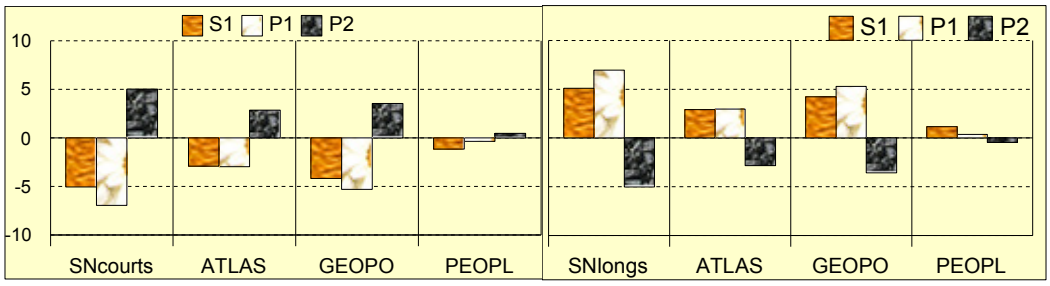


Tableau IX.4 : Proportion et écarts des différents SN ThTop à présenter une reprise lexicale en P1 ou P2

Le test de l'écart réduit montre peu de variations significatives selon la position textuelle. P2 ne montre aucun écart significatif (*i.e.* les répartitions des reprises en cette PosTxt suivent le modèle théorique). Si l'on observe ces mêmes écarts à l'intérieur de chaque sous-corpus, nous remarquons que seul ATLAS montre également un écart significatif positif des SNdef_R en P1. Ces résultats montrent que dans ATLAS l'indice de reprise n'apporte d'information sur les stratégies textuelles qu'au niveau des ThTop de forme SNdef.

IX.2.3.c) Longueur des descriptions selon la position textuelle

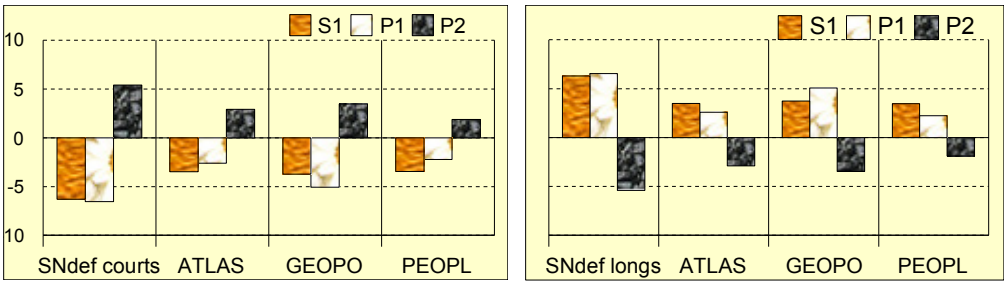
Le dernier indice de co-référentialité en ThTop concerne la longueur des descriptions et plus précisément des SNdem et SNdef. En nous basant sur notre compétence de la langue, nous supposons que les descriptions longues sont tout à fait justifiées en début de section, là où l'installation des objets du discours concernés par la section se fait et où le locuteur doit s'assurer que le lecteur activera les 'bons' référents. Les descriptions courtes sont certainement préférables en début de paragraphe lorsqu'il y a continuité référentielle. Par contre, lorsque le changement de paragraphe coïncide avec un changement thématique, la présence d'une description plus longue est sans doute nécessaire. Nous avons là un indice de segmentation thématique qui a de fortes chances de varier selon la PosTxt à l'intérieur d'un sous-corpus (et selon la présence de certains INIT, voir le chapitre suivant). La position P2 quant à elle montre dans tous les sous-corpus une forte proportion de PRO3, ce qui diminue la proportion de descriptions définies (longues ou courtes).



Graphiques IX.7: Écarts selon les PosTxt de la répartition des SN *courts* vs. *longs* dans les sous-corpus

174 Cette proportion est égale au nombre de _R (par exemple SNdem_R) sur le nombre total des ThTop correspondant (tous les SNdem, qu'ils présentent ou non une reprise).

Les graphiques IX.7 nous indiquent qu'effectivement les SN longs sont associés à S1 et P1 et dissociés de P2. En revanche, ce schéma d'écart ne se retrouve pas dans PEOPL qui ne montre aucune variation significative dans la longueur des SN selon la PosTxt. Cela peut certainement être mis en relation avec le nombre important de NP dans PEOPL qui, dans tous les cas, sont des SN courts. En effet, si l'on ne prend en compte que les SNdef[175], PEOPL affiche des variations significatives selon la PosTxt, comme le montre les graphiques suivants :



Graphiques IX.8 : Écarts selon les PosTxt de la répartition des SNdef *courts* vs. *longs* dans les sous-corpus

Comme on le voit dans ces graphiques IX.8, le schéma d'écart neutre observé précédemment au niveau de PEOPL était dû à la forte proportion de NP dans ce sous-corpus. Le fait que PEOPL, dans son organisation interne, montre également moins de SNdef courts en S1 signifie bien que les SNdef ont, dans ce sous-corpus aussi, un rôle similaire à celui dans ATLAS ou PEOPL.

Si l'on compare les graphiques IX.7 et IX.8, on voit que, pour ATLAS et GEOPO les deux groupes de graphiques sont les mêmes. Le comportement des SNdef courts *vs.* longs est le même que celui des SN en général. En fait, c'est le comportement général des SN courts *vs.* longs qui est fortement conditionné par celui des SNdef. En effet, quelle que soit la position textuelle, plus de 50% des ThTop dans ATLAS ou GEOPO sont de forme SNdef. Cette proportion atteint les 80% de ThTop en S1 dans GEOPO. Alors que dans PEOPL, cette proportion tourne autour de 40%. Le tableau IX.5 donne la mesure de ces données.

	ATLAS			***GEOPO***			***PEOPL***		
	S1	***P1***	***P2***	***S1***	***P1***	***P2***	***S1***	***P1***	***P2***
SNdef %	*77*	*68*	*55*	*80*	*69*	*56*	*47*	*40*	*35*
Nb Phr.	324	1 116	2 504	241	739	3 016	132	321	1 751
ThTop	423	1 647	4 538	303	1 066	5 339	281	794	4 962
Nb phr.	469	1 841	5 282	348	1 257	6 296	348	1 020	6 356

Tableau IX.5 : Proportion des SNdef dans les trois sous-corpus selon la position textuelle

Les sous-corpus ne montrent pas la même distribution des SNdef courts *vs.* longs selon les différentes PosTxt. Dans PEOPL, seul l'écart positif observé en S1 pour les SNdef longs est significatif, ce qui signifie que les autres positions textuelles ne montrent pas de préférence pour des Sndef longs ou courts. Dans ATLAS, il semble y avoir un continuum : plus on est en initiale d'une unité textuelle de haut niveau, plus on a des SNdef longs. Ainsi il y a significativement plus de SNdef longs en S1 qu'en P1 et significativement plus de SNdef courts en P2. Enfin, GEOPO affiche plus de SNdef longs en P1 qu'en S1 (S1 reste toutefois associée à un *z*(SNdef longs)=+3,7). Ces résultats

175 Dans le chapitre précédent, nous avons vu que la longueur des SNdem ne varie pas selon le sous-corpus, alors que les SNdef oui. Dans le début de cette section, nous avons vu que ce n'est que dans ATLAS que l'on observe une variation de la répartition des SNdem selon les PosTxt (beaucoup moins de SNdem en S1). Cette variation ne se retrouve pas au niveau de la longueur des SNdem. Dans tous les cas, il y a significativement peu de SNdem en S1. Pour simplifier les choses, nous avons décidé de ne pas représenter les écarts pour les SNdem. Jusqu'ici, les SNdem ne sont pas très spécifiques à un emploi.

semblent également indiquer que ce sont les SNdef courts qui réalisent dans GEOPO les progressions thématiques intraparagarphiques, en complément des pronoms.

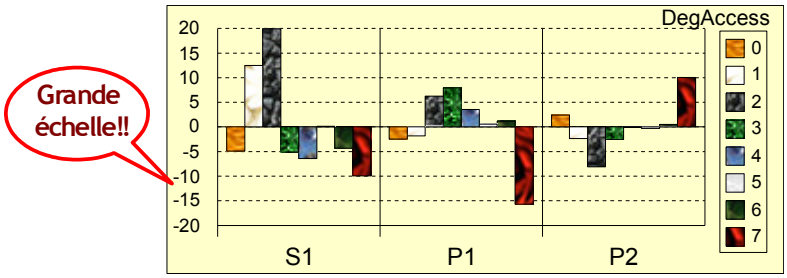
IX.2.4. Récapitulatif des variations en ThTop

Les deux points vus précédemment montrent à quel point les trois sous-corpus gèrent différemment la co-référence en ThTop. PEOPL utilise principalement des pronominalisations et des redénominations tout au long de son texte ; et bien que celles-ci varient significativement selon les PosTxt, elles semblent pouvoir être employées dans toutes les PosTxt. Seules les phrases en S1 s'alignent à peu près sur nos intuitions : descriptions longues et sans reprise. ATLAS suit l'intuition que l'on pouvait se faire : des descriptions longues et sans reprise en S1, moins de descriptions longues et plus de reprises en P1 et enfin, des ProPoss en P2. GEOPO se distingue par son association plus forte entre les changements de paragraphes et les descriptions longues. Les SN_R sont assez peu utilisés dans ce sous-corpus, ce que nous avions déjà remarqué dans le chapitre précédent.

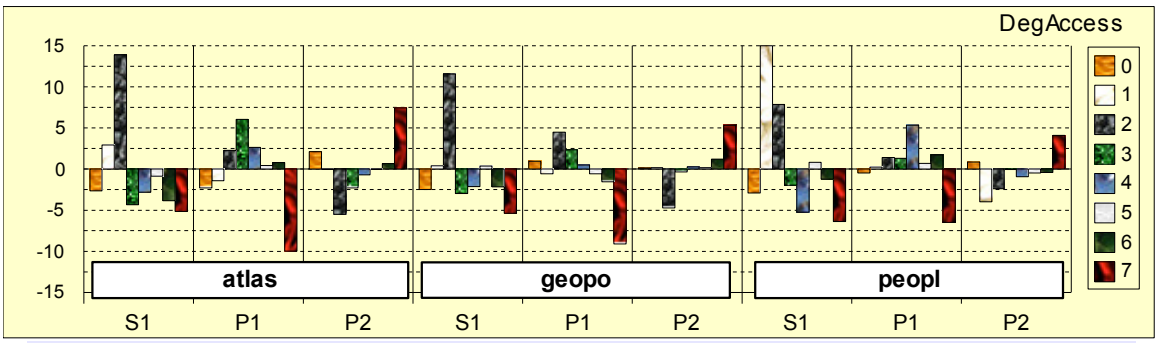
	S1 ———> P1 ———> P2	
ATLAS	SNdef ----------> SNdef_R ----------> ProPoss	
	SN longs ----------> SN courts	
GEOPO	SNdef ----------> (SNdef_R) ----------> ProPoss	
	SN longs ----------> SN courts	
PEOPL	SNdef ----------> SNdef_R ----------> ProPoss	
	NP ----------> NP_R ----------> ProPoss/NP_R	
	SNdef longs	

Tableau IX.6 : Récapitulatif des spécificités des sous-corpus au niveau ThTop selon les différentes PosTxt

Ce récapitulatif peut également se faire en observant les variations selon les PosTxt du degré d'accessibilité selon l'échelle d'Ariel (1990) dans les différents sous-corpus.



Graphique IX.9 : Variations des degrés d'accessibilité - DegAccess - selon les PosTxt



Graphique IX.10 : Schémas d'écart (comportant au moins un écart significatif) des DegAccess selon les PosTxt dans les différents sous-corpus

Nous retrouvons dans les graphiques IX.9 et IX.10 les faits suivant :

DegAccess_0 Les **ThSpe** (voir section suivante) et les **indéfinis** sont significativement moins présents en S1 et cela dans les trois sous-corpus.

DegAccess_1 Les **NP** se trouvent préférentiellement en S1 lorsqu'ils correspondent à une première mention, et cela uniquement dans PEOPL (graphique IX.16). Cette association est à prendre avec prudence, puisque la présence de reprise lexicale est impossible en S1 du fait de notre programme d'annotation (voir l'explication en IX.2.3.b).

DegAccess_2 et 3 Les **SNdef** sont fortement associés à S1 (association de deux indices de rupture) et dans de moindres mesures à P1 lorsqu'ils sont longs et sans reprise lexicale – DegAccess_2. Lorsqu'ils comportent une reprise – DegAccess_3, ils sont uniquement associés à P1 (association de deux indices de déplacement). La position P2 est généralement dissociée des SNdef, mais si l'on observe la répartition des SNdef selon leur longueur, on voit significativement plus de SNdef courts en P2. Cette répartition des SNdef courts ne se trouve pas dans ATLAS – qui montre significativement moins de SNdef courts (VIII.2.2).
Dans les trois sous-corpus, le schéma d'écart du DegAccess_2 se rapproche d'un schéma de dispersion où P1 est la position la plus neutre vis-à-vis de la répartition de ces SNdef. Une nette diminution s'observe d'ATLAS à GEOPO puis à PEOPL. Mais malgré cette diminution, les écarts restent élevés en S1 (avec, nous le rappelons encore, une certaine prudence à avoir vis-à-vis de cet écart du fait de la non prise en compte des reprises lexicales en S1).

DegAccess_4 Les **NP_R** se situent préférentiellement en P1 (association de deux indices de déplacement). Cette association est presque exclusive à PEOPL. L'écart observé dans ATLAS égal z=+2.

DegAccess_5 et 6 Les **SNdem** ne montrent pas de variation selon la PosTxt lorsqu'ils sont longs et sans reprise – DegAccess_5. Par contre, lorsqu'ils sont courts et/ou avec reprise – DegAccess_6 (co-référence plus directe), ils sont significativement moins présents en S1, ce qui ne s'observe que dans ATLAS. Il faut toutefois noter la très faible fréquence des DegAccess_5, les SNdem étant généralement courts et/ou avec reprise lexicale (voir VIII.2.2).

DegAccess_7 Les **ProPoss** sont associés exclusivement à P2 (association de deux indices de continuité), quel que soit le sous-corpus considéré.

IX.3. Répartition des INIT selon la PosTxt

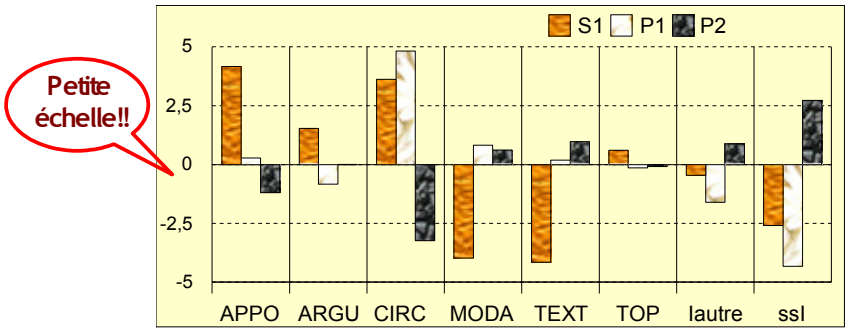
D'un point de vue général, il y a plus d'INIT1 en S1 et P1, surtout en P1 (voir tableau IX.2 en IX.1.1). Pour les INIT2, aucune variation significative n'est mesurée ($z(S1)$ =-0,1 ; $z(P1)$ =+0,4 ; $z(P2)$ =-0,2). Concernant la corrélation forme/fonction des INIT, aucun écart significatif n'apparaît à la comparaison des trois PosTxt (alors que de nombreux écarts significatifs sont mesurés à la comparaison des trois sous-corpus, cf. chapitre précédent). Concernant la répartition des différentes fonctions, aucun écart significatif n'est mesuré au niveau des différents types d'INIT2. En revanche, au niveau des INIT1, des associations entre les différentes PosTxt et certaines fonctions et rôles sémantiques d'INIT apparaissent. Ces associations sont moins fortes que celles observées au niveau des ThTop mais plus importantes que celles notées au niveau des ThSpe.

INIT1	corpus (Nb|%)		S1 (Nb)	P1 (Nb)	P2 (Nb)
CIRC	4839	70	293	984	3562
CIRCautre	2138	30	125	404	1609
CIRCtps	1457	21	120	312	1025
CIRCnot	744	11	32	135	577
CIRCspa	500	7	16	133	351
APPO	844	12	69	153	622
MODA	612	9	9	117	486
TEXT	414	6	2	75	337
ARGU	154	2	12	23	119
Iautre	100	1,5	4	11	85
TOP	59	1	4	10	45
Nb d'INIT1	7022		393	1373	5256

Tableau IX.7 : Répartition des fonction d'INIT selon les PosTxt

IX.3.1. Fonctions discursive des INIT : les appositions et les circonstants en position de discontinuité

Quelle que soit la position textuelle, les CIRC restent les fonctions d'INIT les plus importantes (autour de 70% : 75% en S1, 72% en P1 et 68% en P2). Que ce soit au niveau des différents rôles sémantiques ou au niveau des autres fonctions d'INIT, l'ordre décroissant reste identique selon les différentes PosTxt. La seule exception se situe au niveau des ARGU qui sont remarquablement plus importants en S1, mais le nombre des ARGU est trop faible pour que cet écart soit significatif (voir graphique IX.11)

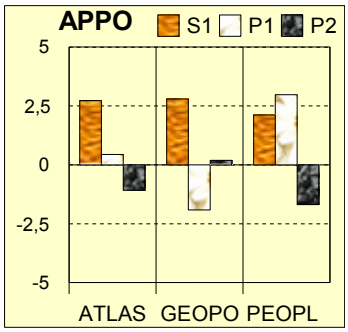


Graphique IX.11 : Variations des fonctions d'INIt selon la PosTxt

À partir de la mesure des écarts réduits, quatre fonctions d'INIT apparaissent comme étant significativement sensibles à la position textuelle : les appositions, les adverbiaux circonstanciels, les adverbiaux modalisateurs et les organisateurs textuels. Ces deux derniers types d'INIT (les MODA et les TEXT) sont dissociés de la position S1 sans montrer d'association pour une autre position textuelle. Alors que nous nous attendions à une association entre TEXT et un indice matériel de discontinuité, aucune donnée ne va dans ce sens. Si l'on observe les écarts dans les trois sous-corpus, les mêmes écarts apparaissent mais moins élevés et du coup, non significatifs. L'absence prononcée de MODA et de TEXT en S1 est donc générale à tous les sous-corpus.

Le schéma obtenu pour les APPO est assez troublant. En effet, étant considérées comme des indices de continuité, il est étonnant que celles-ci se situent souvent en début de section, à moins qu'elles ne permettent de créer

de la continuité là où le changement de section marque une rupture. Si l'on observe les schémas d'écart au sein de chaque sous-corpus, nous voyons trois schémas d'écart différents (graphique IX.12). En ne s'intéressant qu'aux écarts significatifs, nous avons d'un côté ATLAS et GEOPO qui montrent plus d'APPO en S1 et de l'autre PEOPL, sous-corpus présentant le plus d'APPO, qui montre plus d'APPO en P1. Les trois exemples suivants montrent des sections 'introduites' par une APPO dans les différents sous-corpus[176].



Graphique IX.12 : Variations des APPO selon la PosTxt dans les trois sous-corpus

(IX.5) ***Le modus vivendi avec Guernesey*** *[titre niveau 2]*
Proposé aux professionnels par les Autorités françaises et britanniques, *le modus vivendi, par définition ne visait pas à régler les problèmes de fond, mais constituait un accord préalable, en l'occurrence valable un an et renouvelable. L'accord était basé essentiellement sur le principe des concessions mutuelles. Il prévoyait, [...]* *[ATLAS_1]*

(IX.6) ***LES SCÉNARIOS*** [titre niveau 1]
Les deux scénarios identifiés et analysés ici sont ceux qui déterminent une configuration spécifique du triangle syro-libano-israélien, avec ses prolongements sur les situations internes et ses implications régionales et internationales propres : le scénario du statu quo et l'option d'un retrait israélien du Liban-sud . Ces deux cas de figure [...]
Maintien du statu quo actuel *[titre niveau 2]*
Le scénario de " ni guerre , ni paix " est sans aucun doute le plus plausible aujourd'hui [...]
Variations autour du scénario du retrait unilatéral du Liban-sud... ou comment décomposer le triangle ? [titre niveau 2]
Impensable à la veille de l'opération "Raisins de la colère"*, ce scénario avec toutes ses variantes fait désormais partie du domaine du "politiquement" envisageable. Il a été retenu ici en raison de l'évolution du débat en Israël sur le Liban qui, [...]* *[GEOPO_4]*

Dans les exemples extraits d'ATLAS et de GEOPO, on voit que l'apposition permet d'effectuer une continuité entre le titre et le contenu de la section. Il est frappant de constater que de chaque apposition précède une reprise du titre ou d'un élément du titre. Par contre, dans les exemples issus de PEOPL (comme l'exemple IX.7 suivant), les appositions ont pour fonction de faciliter la continuité topicale globale du texte (dans l'exemple, Fiodor Dostoïevski est le topique global du texte) en 'sautant' par dessus les titres qui ne servent, dans ce corpus, qu'à poser un repère dans la chronologie relatée[177].

(IX.7) ***1. CHRONOLOGIE SOMMAIRE*** *[titre niveau 1]*
Né à Moscou, *Dostoïevski vécut surtout à Saint-Pétersbourg, où il mourut. Il ne reçut pas de formation universitaire, mais [...]*
La vie active de Dostoïevski peut se diviser en trois périodes. Dans la première, jusqu'à vingt-sept ans, il se forme et s'essaie . Dans la deuxième, il réfléchit et écrit [...]
2. AVANT LE BAGNE *[titre niveau 1]*
L'enthousiaste des lettres *[titre niveau 2]*
Second fils d'un médecin-major, *Fiodor ne fut pas un enfant martyr. Son père n'était pas un [...]* *[PEOPL_3]*

La fréquence des CIRC diminue progressivement selon le niveau de segmentation envisagé. En S1, 75% des INIT sont des CIRC. 25% des premières phrases de section commencent par un CIRC. En P1, ces proportions passent à 72% et 24%. En P2, elles descendent à 68% et 20%. Comme le montrent ces résultats, les CIRC sont souvent employés en début de nouvelle section soit pour reprendre un élément du titre (exemple IX.8), soit pour poser les localisations circonstancielles nécessaires à la construction du *text-world* relaté dans la section (exemple IX.9).

(IX.8) ***LE PROBLÈME DES 15 000 PIEDS*** *[titre niveau 2]*
En dessous de 15 000 pieds, *les avions de combat tactiques sophistiqués et coûteux restent vulnérables à l'action de moyens pléthoriques et peu coûteux comme l'artillerie anti aérienne automatique (AAA) légère de tout calibre, les SAM, et surtout les systèmes portables à guidage infrarouge comme les missiles américains Stinger. [...]* *[GEOPO_10]*

176 D'autres exemples d'apposition en S1 dans ATLAS sont donnés dans la partie X.5, où l'on voit des reprises pronominales d'éléments du titre précédées d'une apposition.

177 Nous reparlerons de ce fonctionnement de l'apposition dans la partie X.5 consacrée aux reprises des éléments du titre.

(IX.9) ***LE COMBAT LITTORAL*** *[titre niveau 2]*
Depuis la fin de la guerre froide,[178] *l'U.S. Navy a voulu montrer qu'elle offrait des réponses adaptées aux réalités contemporaines. Au début des années 1990, n'ayant plus d'adversaire en mer, elle a commencé à [...]* *[GEOPO_10]*

Les cas de reprises du titre en CIRC comme dans l'exemple IX.8 dépendent de l'implication du titre dans la construction du text-world. Il faut en effet que le titre de section exprime autre chose que le référent principal de la section. Ces cas sont relativement rares. Les cas de CIRC en S1 sont principalement similaires à l'exemple IX.9. Nous trouvons en S1 de beaux exemples de world-builders, ces expressions qui permettent de poser les fondations du text-world à construire. De nombreux textes et de nombreuses sections de haut niveau commencent par un CIRC. En voici quelques exemples :

(IX.10) *LA LUTTE CONTRE LE TERRORISME : ESSAI DE BILAN INSTITUTIONNEL [titre de l'article]*
Depuis une quarantaine d'années, *les républicains se sont fait fort de réduire le poids de l'État fédéral. Or l'actuelle lutte contre le terrorisme, [...], remettrait en cause [...]*
FACE À L'URGENCE : LES PREMIÈRES DÉCISIONS DE L'ADMINISTRATION BUSH [titre niveau 1]
Dans le mois qui a suivi l'attentat du 11 septembre, *l'administration a procédé [...]* *[GEOPO_1]*

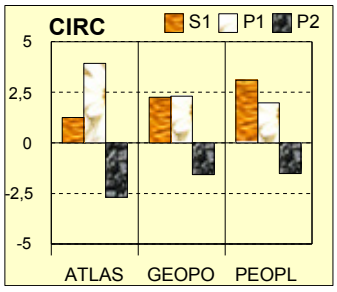
(IX.11) ***LA MAÎTRISE DES ESPACES, FONDEMENT DE L'HÉGÉMONIE MILITAIRE DES ÉTATS-UNIS [titre de l'article]***
Depuis la fin de la guerre froide, *les spécialistes de politique étrangère se sont demandé quel nouvel ordre mondial succéderait à la bipolarité Est-Ouest, et quelle nouvelle doctrine remplacerait pour les Etats-Unis celle du containment. Ceux qui pensent que nous sommes arrivés à un "moment unipolaire" de l'Histoire et prônent pour les Etats-Unis une politique de suprématie, [...]* *[GEOPO_10]*

Ces exemples sont tous tirés du sous-corpus GEOPO. Il y a assez peu de sections commençant par un CIRC dans ATLAS. Dans PEOPL, les CIRC semblent avoir un rôle relativement différent. Les CIRC de PEOPL sont souvent liés au topique (le personnage du portrait), ce qui se retrouve dans les CIRC par une forte présence d'anaphores ou de cataphores, réalisées soit dans subordonnées par un pronom référant au personnage (exemple IX.6) soit par des SN possesifs comme en (IX.5).

(IX.12) ***Les sources de l'imaginaire*** *[titre niveau 2]*
Comme il connaît les ars moriendi – ces gravures du savoir-mourir, les clefs des songes et le Tarot, les traités d'alchimie et ceux d'astrologie, *Bosch a lu les ouvrages des mystiques, La Nef des fous de Brant, La Légende dorée où sont décrites les tentations de saint Antoine, et Les Visions de Tungdal, poème traduit de l'irlandais et qui montre une sorte de don Juan du XIIIe siècle gratifié, pour son salut, du spectacle même de l'enfer, quintessence de l'horrible. Matière de rêverie, répertoire de figures. [...]* *[PEOPL_12]*

(IX.13) ***Un art de la persuasion*** *[titre niveau 2]*
Dans ses activités, *Pascal est servi par un grand sens de la communication : plus que tout autre en son temps, il a saisi les exigences du contact entre les esprits et de l'art de persuader [...]* *[PEOPL_4]*

Les trois sous-corpus montrent des répartitions de CIRC différentes selon les PosTxt, comme le montre le graphique IX.13. ATLAS associe les CIRC (et surtout les CIRCspa, voir infra) à la position P1 en opposition à la position P2 (nous verrons des exemples dans la sous-section suivante). GEOPO ne montre pas d'écart significatif, même si les CIRC sont proches d'être significativement associés aux initiales de sections et de paragraphes. À l'inverse d'ATLAS, PEOPL associe les CIRC à S1, ce qui confère aux CIRC dans PEOPL un fort pouvoir de segmentation. Nous verrons dans la sous-section suivante quel type de CIRC suit ce schéma d'écart.



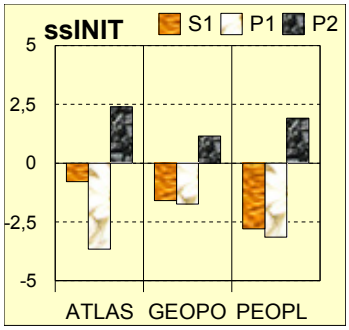
Graphique IX.13 : Variation des CIRC selon la PosTxt dans les trois sous-corpus

Les APPO et les CIRC montrent un fonctionnement assez proche dans PEOPL, même si les schémas d'écart que ces INIT affichent dans ce sous-corpus diffèrent. Il semble qu'en cumulant les APPO et les CIRC, nous obtenons

178 Il est intéressant de remarquer que cet article commence lui-même par le circonstant : *Depuis la fin de la guerre froide,* voir l'exemple (IX.10) infra.

approximativement le schéma d'écart pour tous les INIT en général. D'ailleurs, si l'on observe les écarts des autres fonctions d'INIT, aucune variation significative n'est mesurée, mise à part celle concernant l'absence d'INIT – ssINIT. Nous retrouvons ce rapprochement entre APPO et CIRC dans le chapitre suivant.

Nous avons pris la décision de représenter les phrases sans INIT dans tous les graphiques précédents car l'absence d'INIT est apparue comme un indice de séquentialité que l'on ne soupçonnait pas. En effet, dans le graphique IX.14, les ssINIT montrent un schéma d'écart intermédiaire entre un schéma d'association exclusive avec P2 et un schéma de dispersion où P1 tient le rôle de l'écart négatif. En mesurant ces schémas d'écart dans les trois sous-corpus, nous observons que seul ATLAS montre effectivement un schéma de dispersion où l'on observe significativement moins de ssINIT en S1, les ssINIT en P2 sont à la la limite de l'écart positif significatif. PEOPL montre une dissociation de S1 et P1 trop faible pour entraîner un écart significatif en P2. GEOPO montre un schéma neutre. ATLAS et PEOPL sont ainsi les seuls à montrer une variation significative pour les phrases sans INIT, variation qui semble associer les ssINIT à un indice de continuité, ce que nous vérifierons dans le chapitre suivant.

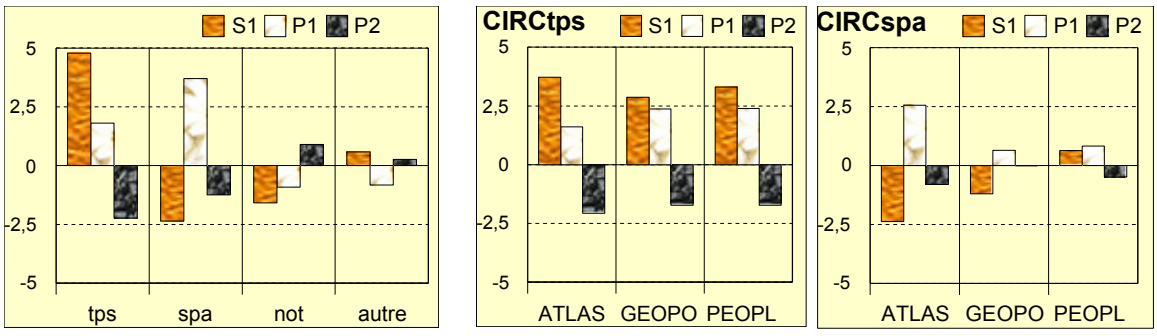


Graphique IX.14 : Variation des ssINIT selon la PosTxt dans les trois sous-corpus

IX.3.2. Rôles sémantiques des adverbiaux circonstanciels

Le chapitre précédent a montré que chaque rôle sémantique de CIRC, à l'exception des CIRCtps, est associé spécifiquement à un sous-corpus (CIRCspa et ATLAS, CIRCnot et GEOPO et CIRCautre et PEOPL). Au niveau des positions textuelles, seuls deux rôles textuels montrent des écarts significatifs : les CIRCtps et les CIRCspa (graphiques IX.15).

Les CIRCtps s'affichent préférentiellement au niveau des changements d'unités typodispositionnelles. Cette association se retrouve quel que soit le sous-corpus considéré. Ainsi, chaque sous-corpus montre un écart significatif positif pour les CIRCtps en S1. Les CIRCtps seraient ainsi de bons indices de discontinuité au niveau de l'organisation globale et cela indépendamment du type de texte. Nous vérifierons dans le chapitre suivant si cette capacité à indiquer une discontinuité est propre aux CIRCtps ou simplement aux changements de sections.



Graphiques IX.15 : Variations des fonctions d'INIt selon la PosTxt

Nous avions vu dans le chapitre précédent que 66% des CIRCspa proviennent d'ATLAS. Dans GEOPO et PEOPL les CIRCspa sont parsemés dans le texte sans prise en compte de la position textuelle. Dans ATLAS en revanche, ils semblent associés aux indices de déplacement en P1. Ainsi, 5,5% des paragraphes dans ATLAS commencent par un CIRCspa (contre 1,5% dans GEOPO et 1% dans PEOPL). À l'inverse des CIRCtps, les CIRCspa

ne commencent que très rarement les sections. Dans le corpus entier, moins d'1,5% des premières phrases de sections commencent par un CIRCspa, contre plus de 10% par un CIRCtps ou un CIRCautre. Les CIRCnot introduisent 2,7% des phrases en S1. Dans ATLAS, les CIRCspa introduisent à peine plus de phrases en S1 que dans le corpus entier (2%). Même en INIT2, nous ne trouvons qu'un seul CIRCspa en S1 (il y a 54 séquences d'INIT en S1, ce qui représente environ 5% des premières phrases de sections). Les deux exemples suivants montrent comment ATLAS emploie les CIRCspa en introduction de paragraphe.

Dans l'exemple IX.14, les adverbiaux circonstanciels spatiaux permettent d'opérer une énumération indexée selon des zones géographiques différentes, ce qui correspond à une TSC spatiale qui organise les informations selon leur répartition sur un territoire bien délimité (en l'occurrence, la France). Nous remarquons que les zones ne sont pas toujours localisées par un CIRCspa. Il est assez fréquent de voir une TSC spatiale construite par une alternance de CIRC et de ThTop ou de compléments postverbaux référant à une zone géographique.

(IX.14) ***Un élève étranger sur trois en Île-de-France*** [titre niveau 3]

Plus du tiers (38%) des élèves étrangers vit ***en Île-de-France,*** *région qui rassemble 20% à peine de la population scolaire française. Un élève sur six y est de nationalité étrangère - et même plus d'un sur quatre en Seine-Saint-Denis, département qui scolarise à lui seul plus de 80 000 enfants étrangers. Les Algériens, les Portugais [...]*

Les académies de Lyon et Grenoble *accueillent près de 150 000 enfants étrangers, dont plus de 50 000 dans le seul département du Rhône, qui est avec la Seine-Saint-Denis, Paris et les Hauts-de-Seine l'un des quatre départements français où la proportion d'élèves étrangers dépasse 15%. Les enfants d'origine maghrébine sont de loin les plus nombreux, avec une forte composante algérienne ; et 20% des élèves turcs vivent* ***dans la région Rhône-Alpes****.*

L'Est, de Montbéliard à Strasbourg et Nancy, *forme le troisième ensemble à forte proportion d'élèves étrangers : environ 100 000 enfants scolarisés - de 8 à 12% des élèves selon les départements - se répartissant selon trois nationalités principales : Marocains, puis Turcs, puis Algériens.*

Dans les académies méditerranéennes *qui comptent également une centaine de milliers d'élèves étrangers, les taux d'élèves étrangers avoisinent 10% dans le Var, les Bouches-du-Rhône et en Corse ; ailleurs ils sont inférieurs à la moyenne nationale.*

Dans le Nord, *les élèves de nationalité étrangère sont surtout nombreux dans l'agglomération lilloise.* ***Le département du Nord*** *compte 50 000 élèves étrangers, 8% de sa population scolaire, dont les deux tiers d'origine maghrébine, les Marocains étant plus nombreux que les Algériens. En revanche,* ***le Pas-de-Calais****, qui fut un département de forte immigration du temps du charbon, fait maintenant partie des départements où les enfants étrangers sont peu nombreux.*

À l'ouest d'une ligne Le Havre-Montpellier, *la population scolaire comprend généralement moins de 5% d'enfants étrangers - et même moins de 1% dans les Côtes-du-Nord, le Morbihan et la Vendée.* ***Les académies de l'Ouest,*** *Caen, Rennes, Nantes et Poitiers, accueillent quatre fois moins d'élèves étrangers que celles de Lyon, ou de Lille. Les enfants portugais sont habituellement les plus nombreux, précédant les Marocains ou les Turcs.* *[ATLAS_2]*

On voit bien dans cet exemple une organisation de l'information selon une TSC spatiale globale (*i.e.* chaque paragraphe correspond à une unité spatiale). Cependant, ce ne sont pas que les adverbiaux circonstanciels antéposés qui indiquent la TSC. La localisation géographique des informations contenues dans le premier 'item' de la TSC est même exprimée dans un complément postverbal totalement intégré à la proposition. Le second 'item' est, lui, introduit par un ThTop *(Les académies de Lyon et Grenoble)* qui ne localise pas vraiment toute la zone géographique concernée, précisée au final en position finale : « *dans la région Rhône-Alpes* ». Au final, seule la moitié des segments de la TSC sont introduits par un CIRC. Un autre exemple de la sorte est donné en (IX.15), où l'on voit une TSC locale (*i.e.* intraparagraphique) construite autour d'une utilisation équivalente de ThTop et de CIRCspa.

(IX.15) ***Dans l'Avon (35,3%),*** *les Français sont surtout présents à Bristol, ville portuaire, et dans une moindre mesure à Woodspring et Bath lorsque l'on prend en compte les résultats par districts.* ***Dans le Surrey (35,4%),*** *hormis le district d'Elmbridge, la répartition de la population de nationalité française est uniforme. Enfin* ***le Kent (89,3%),*** *porte de l'Angleterre lorsque l'on se rend dans l'Ile par le Pas de Calais, semble avoir profité de l'ouverture du Tunnel sous la Manche en 1994. Le nombre de Français y est passé de 1 000 à plus de 2 000 en trois ans.* *[ATLAS_1]*

Ce type d'alternance CIRC/ThTop, très fréquente pour les TSC spatiales, est assez rare pour les TSC temporelles. Les TSC temporelles sont plus souvent construites par des CIRCtps, comme le montre l'exemple IX.16.

(IX.16) ***LES FRANÇAIS EN GRANDE-BRETAGNE*** *[titre niveau 1]*

La France envoie ses concitoyens en grand nombre vers l'Angleterre. La Grande Bretagne représente la troisième destination d'expatriation pour les Français après la Belgique et l'Allemagne. Selon les chiffres consulaires, la population française présente en Angleterre serait d'environ 160 000 personnes.

En 1993, *les statistiques européennes estimaient la communauté française au Royaume Uni à environ 42 000 personnes. Ce chiffre est sujet à erreur étant donné qu'il n'existe aucun recensement global de la population française. Les Français, en effet, ne sont pas tenus de déclarer officiellement leur présence au Royaume Uni. Le chiffre de 42 000 personnes est basé sur le nombre de personnes immatriculées au Consulat Français.*

En 1994, *11 000 personnes ont transité de France au Royaume Uni afin d'y élire résidence. La même année, en 1994, le consulat général de France à Londres , enregistrait 5 286 nouvelles immatriculations. Cela signifie que l'immatriculation représente environ la moitié du nombre réel de Français se rendant en Angleterre afin d'y élire résidence.* ***En 1995,*** *le registre d'immatriculation des Français à Londres indique le chiffre de 7 469 nouvelles entrées. L'hypothèse avancée est que le flux réel est supérieur à 15 000 personnes pour l'année 1995. Cette hypothèse est étayée par l'amorce dès 1993, de conditions favorables à l'établissement des Français en Angleterre. L'Angleterre a, dès le milieu des années 1990, recouvré une croissance économique favorable à l'emploi alors que la Livre, relativement faible par rapport au Franc Français jusqu'en 1996, jouait en faveur des investissements français. D'autre part, le nombre de personnes de nationalité française présentes en Angleterre a subi une très forte augmentation depuis le début des années 1990.*

En 1996, *la communauté des Français immatriculés atteint 48 767 personnes, soit plus de 6 000 personnes depuis 1993. Il est possible que le chiffre réel soit le double.*

En 1998, *la communauté française immatriculée atteint 60 000 personnes. Cela signifie que près de 20 000 français ont choisi de migrer au Royaume Uni entre 1993 et 1998. La communauté française en Angleterre a quasiment augmenté de moitié depuis l'ouverture des frontières européennes si l'on s'en tient aux chiffres officiels. Officieusement, le Consulat de France à Londres estime qu'il y aurait près de 100 000 Français présents dans la capitale britannique !* *[ATLAS_1]*

Il peut également y avoir des 'cadres' isolés, *i.e.* des cadres qui n'appartiennent pas à une TSC. C'est principalement le cas dans GEOPO et PEOPL, où les CIRCspa introduisent généralement une localisation spatiale isolée au milieu d'une zone où soit les propos ne sont pas forcément localisés spatialemenet soit la localisation spatiale n'est pas exprimée mais inférée car elle correspond à la localisation générale du texte (ici les États-unis). Il semble que ces CIRC n'ont pas une très grande portée, comme dans l'exemple suivant où la portée du CIRCspa s'arrête à la fin de sa phrase d'accueil. Cette hypothèse sera approfondie dans le chapitre suivant.

(IX.17) *Au niveau des Gouverneurs, les changements sont moins massifs qu'au sein de le Législatif. Néanmoins, ce n'est pas une fonction à négliger : tous les derniers Présidents d'envergure - Clinton, Bush Sr, Reagan et Carter - ont été des Gouverneurs avant d'atteindre la Présidence. Que ce soit l'Arkansas, la Californie, ou le Texas, l'accession à le poste de Gouverneur semble maintenant être un marche-pied efficace pour atteindre le poste le plus élevé du pays. A ce niveau, une autre figure montante du Parti démocrate a acquis une certaine visibilité. Il s'agit de Bill Richardson, qui vient d'être élu Gouverneur du Nouveau Mexique en battant le républicain John Sanchez, 57% à 38%.* ***En Europe,*** *sa réputation vient essentiellement de son action diplomatique, notamment à l'ONU, entre 1997 et 1998. Il était devenu membre de l'équipe présidentielle de Clinton en 1998, comme Secrétaire à l'Energie, avant de se lancer dans une carrière politique nationale. Sa récente élection constitue ainsi son premier succès sur la voie de l'enracinement électoral, un élément qui, jusqu'à présent, avait toujours manqué à ce haut fonctionnaire. Ses prises de position traduisent une modération certaine, même si ses engagements en faveur de la lutte contre la pollution ou l'extension de la couverture-santé (health care) sont solides . A part ce nouveau venu sur la scène étatique, les autres résultats étaient attendus [..]* *[GEOPO_8]*

IX.3.3. Récapitulatif des variations des différents INIT

	S1 ———> P1 ———> P2
ATLAS	APPO ----------> CIRC ----------> ssINIT
	CIRCtps ---------> CIRCspa -------> ssINIT
GEOPO	APPO CIRC(tps) --------> CIRC(tps) -------> ssINIT
PEOPL	CIRC(tps) ----------> APPO ----------> ssINIT

Tableau IX.8 : Spécificité des sous-corpus au niveau INIT selon les différentes PosTxt

Les variations observées au niveau des INIT sont assez faibles comparées à nos attentes. Les différents types d'INIT spécifiques à certains sous-corpus (les MODA par exemple) ne se retrouvent pas tous dans l'analyse des

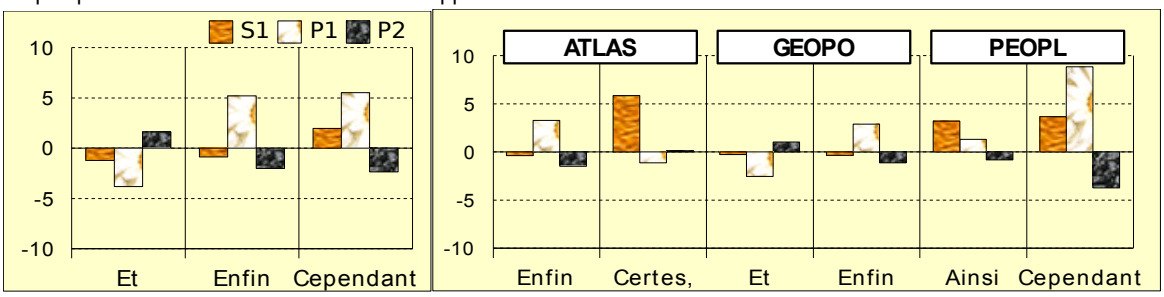
variations selon les positions textuelles. Au vu de nos résultats, les MODA ne se conjuguent pas avec les indices physiques du découpage matériel du texte. Seuls les INIT de type APPO et CIRC coïncident avec des positions textuelles particulières, ce qui est résumé dans le tableau IX.8.

Cependant, il est tout à fait possible que les autres types d'INIT participent au marquage de la séquentialité par combinaison avec certains types spécifiques de ThTop ou de ThSpe. En effet, selon notre définition du marquage discursif, ce sont des configurations d'indices et non des indices seuls qui indiquent une étape dans la séquentialité du discours. L'étude de ces configurations constitue l'objectif principal du chapitre suivant.

IX.4. Des connecteurs spécifiques à P1 ou P2

Nous avions vu dans le chapitre précédent que les différents sous-corpus montraient des préférences en matière de connecteur. Ainsi, ATLAS est associé aux connecteurs *Cependant* et *Enfin,* GEOPO à *Ainsi*, *Or* et *Enfin* et PEOPL à *Mais*, *Et* et *Car*.

Si l'on observe le comportement des connecteurs les plus fréquents en position initiale, seuls *Et*, *Enfin* et *Cependant* affichent des variations significatives, comme le montre le premier graphique en IX.16. *Enfin* et *Cependant*, s'associent préférentiellement avec la PosTxt P1, à l'inverse *de Et.* En observant le deuxième graphique qui indique les variations significatives observées dans chaque sous-corpus, on remarque que seul *Enfin* reste associé à ses sous-corpus préférés. De nouvelles associations apparaissent.



Graphiques IX.16: Des connecteurs différents selon les position textuelles

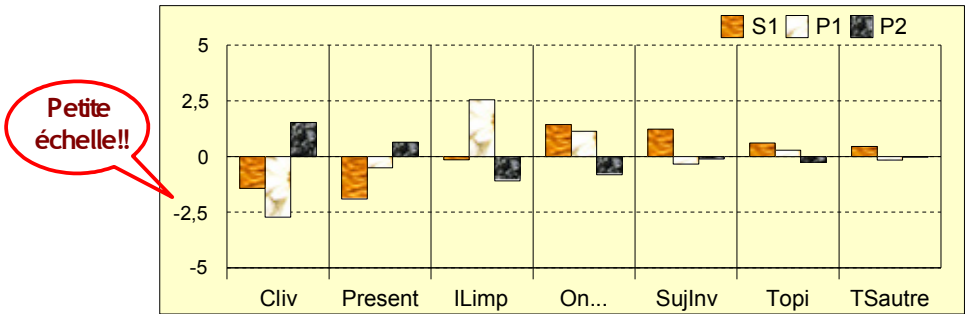
Les connecteurs qui s'associent, dans un sous-corpus particulier, aux positions S1[179] ou P1 peuvent constituer de bons candidats à l'indication d'un déplacement *(Enfin* et *Certes* dans ATLAS, *Enfin* dans GEOPO, *Ainsi* et *Cependant* dans PEOPL). Ces associations peuvent constituer le point de départ d'une étude plus approfondie.

IX.5. Répartition des ThSpe selon la PosTxt

De façon générale, les ThSpe montrent des variations assez faibles selon les PosTxt (voir tableau IX.1 en IX.1.1). Cette faiblesse de variation se retrouve au niveau de tous les types de ThSpe. Ainsi, comme le montre le graphique IX.17, seules les constructions clivées et impersonnelles affichent un écart réduit à peine supérieur à 2,5.

Les clivées sont significativement moins présentes en P1, à l'inverse des ILimp significativement plus présentes en P1. En dehors de ces deux écarts, aucun autre ThSpe ne montre de sensibilité à l'égard de la PosTxt. En observant les écarts à l'intérieur des différents sous-corpus, aucun comportement particulier n'apparaît.

179 Les fréquences en S1 des différents connecteurs sont si faibles que les écarts observés sont plus à prendre comme un exemple qu'une généralité (en confondant les 9 formes les plus fréquentes, nous arrivons à peine à 10 phrases en S1, dont 8 dans PEOPL).

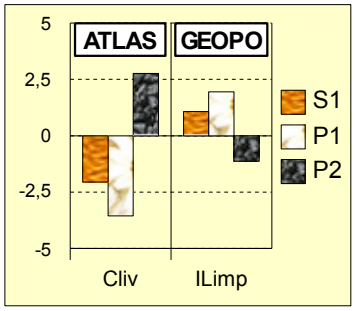


Graphique IX.17: écarts des types de ThSpe selon la PosTxt

Les clivées sont généralement associées à un effet de déplacement, notamment, un déplacement en douceur d'un topique à un autre (voir V.4.4), ce qui aurait pu correspondre à notre conception du changement de paragraphes. Cependant, face à la réalité de nos données, les clivées semblent être davantage utilisées pour mettre le focus sur une entité ou une circonstance particulière plutôt que d'effectuer un déplacement topical. Ainsi, nous avons des clivées de type *c'est X qui est le plus/moins {important, fort, concerné, etc..}* ou *c'est SP que X est le plus/moins {important, fort, concerné}* ou encore *c'est (particulièrement) le cas {de, pour, en, dans, etc..} X* (voir les exemples donnés dans le chapitre précédent, en VIII.5). Ce fonctionnement a d'ailleurs été souligné par Péry-Woodley (1992) qui remarque, à l'issue d'une analyse comparative du fonctionnement des clivées en anglais et en français, que, dans des productions de textes français par des locuteurs natifs, les clivées ont un rôle principalement local de mise en focus alors que dans des productions de textes anglais par des locuteurs natifs elles ont un rôle davantage organisationnel (cf. Péry-Woodley 1992:155)[180].

Ce rôle davantage local des clivées (en français) se retrouve dans notre analyse des variations selon les différentes PosTxt. Comme le montre le graphique IX.17, les constructions clivées sont significativement moins présentes en P1. Si l'on observe ce qui se passe dans ATLAS, le seul sous-corpus à montrer significativement plus de constructions clivées, nous obtenons le graphique IX.18 qui montre que le schéma d'écart obtenu au niveau du corpus entier est uniquement le fruit des variations observées dans ATLAS. Ce dernier graphique nous montre également un écart significatif positif des clivées en P2 (nous avons là un schéma d'éclatement). Nous avions donné un exemple de l'emploi des ThSpe (et notamment des clivées) dans ATLAS en IX.1.2.

Nous n'avions pas d'intuitions quant à la fonction discursive des ILimp, à part celle de leur implication au niveau de la composante interpersonnelle. Selon les résultats affichés dans le graphique IX.18, les ILimp montrent un écart significatif positif pour la position P1. Dans le chapitre précédent, une forte association est apparue une forte association entre les constructions impersonnelles et le sous-corpus GEOPO. Mais si l'on regarde le schéma d'écart des ILimp dans GEOPO selon les différentes PosTxt, aucun écart significatif n'apparaît (graphique IX.18).



Graphique IX.18: Variations des clivées et des ILimp selon les PosTxt

À la différence de l'association Cliv-ATLAS, l'association ILimp-GEOPO n'est pas responsable du schéma général. L'écart positif en P1 observé au niveau du corpus entier est le

180 Dans notre partie V.4.4 sur les constructions focalisantes, il est intéressant de noter que la plupart des travaux sur les clivées sont d'une part basés sur l'anglais et d'autre part, inscrit dans le modèle de la structure informationnelle à un niveau essentiellement phrastique (Lambrecht 1994).

résultat de l'amalgame des schémas dans les trois sous-corpus, *i.e.* de façon générale, les ILimp se situent davantage en P1, indépendamment du type de texte.

Nous voyons bien que les variations des ThSpe sont extrêmement faibles et n'apportent pas d'information réellement interprétable en l'état relativement à l'analyse de l'organisation du discours. Il faudrait certainement réaliser une étude plus fine des différentes constructions spéciales rassemblées dans la catégorie ThSpe. Mais nous gardons toutefois certaines réserves quant à un fonctionnement discursif de telles constructions. Il semble que la justification de ces arrangements syntaxiques se situe plutôt au niveau de la structure informationnelle de la phrase que de la structure du discours.

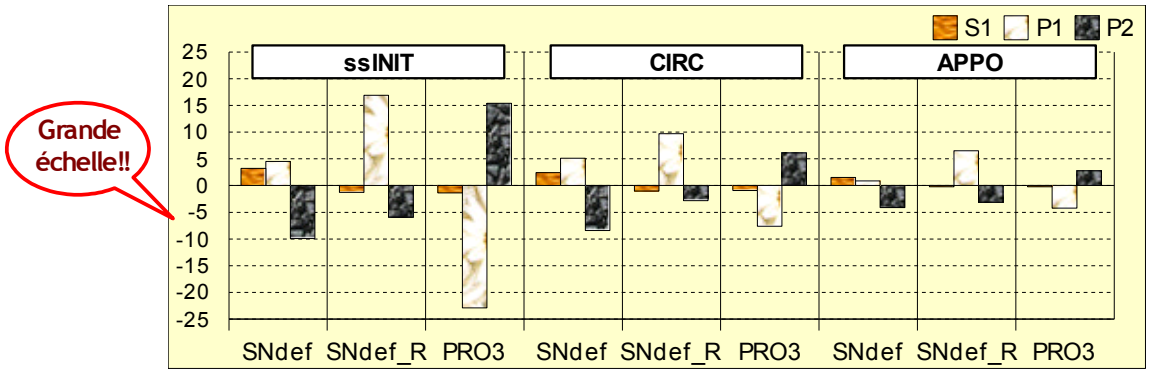
IX.6. Collocations selon la PosTxt

Si l'on observe comment se répartissent les collocations entre un INIT et un ThTop/ThSpe selon les positions textuelles, nous voyons que les collocations préférées repérées au niveau des sous-corpus restent les mêmes quelle que soit la position textuelle. Ainsi, dans toutes les positions, ce sont plutôt des 'non collocations' (*i.e. d*es ThTop seuls non précédés d'INIT) qui dominent.

INIT1	*ThTop*	*ThSpe*	*S1*	*P1*	*P2*
.	SNdef	.	*36,05*	*21,15*	*18,28*
.	SNdef_R	.	*4,89*	*14,16*	*9,34*
CIRC	SNdef	.	*13,39*	*6,95*	*5,24*
.	PRO3	.	*1,89*	*1,63*	*7,54*
CIRC	SNdef_R	.	*1,37*	*5,54*	*3,77*
.	TTautre	.	*4,03*	*3,16*	*3,27*
.	SNdem	.	*3,09*	*3,18*	*3,28*
.	NP	.	*4,89*	*1,99*	*2,21*
.	SNind	.	*2,15*	*3,06*	*3,26*
.	.	Clivée	*1,72*	*2,16*	*3,68*

Tableau IX.9 : Répartitions des collocations préférées selon les différentes positions textuelles

D'un point de vue général, les écarts significatifs selon les positions textuelles suivent ceux observés au niveau des ThTop/ThSpe et non au niveau des INIT. Ainsi, les collocations avec SNdef affichent une fréquence significativement supérieure en S1 et P1 et inférieure en P2. Celles avec SNdef_R sont spécialement supérieures en P1 et celles avec PRO3 ont une fréquence très supérieure en P2 et, par contrecoup, très inférieure en P1.



Graphique IX.19 : Variations des collocations les ThTop les plus fréquents et les INIT les plus fréquents (dont ssINIT)

Le graphique IX.19 relate les écarts observés dans le corpus entier pour les collocations entre un SNdef, SNdef_R ou un PRO3 et un INIT de type CIRC APPO. Comme on le voit, les écarts sont atténués lorsqu'il y a un INIT, quel qu'il soit.

Si l'on observe ce qui passe à l'intérieur de chaque sous-corpus, les spécificités de chaque sous-corpus associées aux traits spécifiques de chaque positions textuelles se combinent. Si l'on prend les écarts les plus significatifs, nous retrouvons dans PEOPL des collocations avec NP (avec NP en S1 et NP_R en P1) ou avec des ThSpe (uniquement en P2), dans GEOPO des collocations avec MODA (en P1 et P2) et des TEXT (en P2 uniquement). ATLAS montre particulièrement plus de phrases sans INIT en P2.

Nous n'insistons pas davantage dans cette façon d'analyser les collocations. En effet, il est assez vain de vouloir voir émerger des configurations d'indices particulières en prenant toutes les données d'un bloc. Certains types de ThTop et d'INIT semblent particulièrement pertinents à analyser en profondeur. Le chapitre suivant reprend chaque indice qui est ressorti comme étant une spécificité d'un sous-corpus et/ou comme étant associé à une certaine position textuelle et analyse les variations que sa présence entraîne sur la nature de son environnement.

IX.7. Récapitulatif général des variations selon les positions textuelles

Ces analyses des écarts selon les trois positions textuelles distinguées montrent bien l'importance de ce facteur de variation. Dans l'ensemble, nous retrouvons approximativement nos hypothèses d'associations entre les différents éléments en position initiale et les trois PosTxt. Nous avons repris ici les deux tableaux récapitulatifs présentés à la fin des parties IX.2 et IX.3. Dans ces tableaux, les ThSpe ne figurent pas. Les écarts réduits mesurés pour les différents ThSpe ne nous ont pas appris grand chose, ou si : les constructions spéciales n'ont vraisemblablement pas de rôle dans le signalement de la séquentialité, alors que nous espérions en une certaine corrélation entre les constructions thématisantes et les phénomènes de déplacement. Même les clivées, généralement associées à un phénomène de discontinuité, montrent un fonctionnement principalement limité à la structure informationnelle de la phrase. Dans cette situation, les différentes positions textuelles ne peuvent pas vraiment influer sur le fonctionnement des ThSpe.

Au niveau des ThTop et des INIT, certaines formes apparaissent comme de bons indices potentiels du marquage des différentes étape de la séquentialité. Ainsi, les formes significativements plus présentes en S1 peuvent postuler au poste d'indice de rupture, celles en P1 au poste d'indice de déplacement et celles en P2 au poste d'indice de continuité. Nous testons ces corrélations dans le chapitre suivant.

ThTop	***S1 ————> P1 ————> P2***
ATLAS	SNdef ----------> SNdef_R ----------> ProPoss
	SN longs ----------> SN courts
GEOPO	SNdef ----------> (SNdef_R) ----------> ProPoss
	SN longs ----------> SN courts
PEOPL	SNdef ----------> SNdef_R ----------> ProPoss
	NP ----------> NP_R ----------> ProPoss/NP_R
	SNdef longs
INIT	***S1 ————> P1 ————> P2***
ATLAS	APPO ----------> CIRC ----------> ssINIT
	CIRCtps ---------> CIRCspa -------> ssINIT
GEOPO	APPO CIRC(tps) --------> CIRC(tps) -------> ssINIT
PEOPL	CIRC(tps) ----------> APPO ----------> ssINIT
Connect	***S1 ————> P1 ————> P2***
ATLAS	Dissociation ----------> Association
GEOPO	
PEOPL	

Tableau IX.10 : Récapitulatif des spécificité des sous-corpus selon les différentes PosTxt

Chapitre X
Configurations d'indices de séquentialité

Sommaire

Ce chapitre s'attache à expliquer les variations occasionnées par les indices émergeant des chapitres précédents sur leur environnement, c'est-à-dire sur les autres éléments présents en position initiale. Ces variations nous permettent de déterminer les 'modalités de cohabitation' entre les différents éléments ThTop et INIT. Nous parlons de **« cohabitations »** plutôt que de « collocations » car il ne s'agit plus d'observer la présence conjointe de deux éléments en position initiale (ce que nous avons fait en VIII.7 et en IX.6), mais d'examiner les tendances 'environnementales' d'un élément.

L'environnement pris en compte pour les Thèmes topicaux correspond à l'élément détaché en initiale. Pour représenter toutes les phrases, nous considérons l'absence d'INIT (ssINIT) comme une catégorie d'INIT. L'environnement des éléments détachés est représenté par le degré d'accessibilité, qui permet de référer simplement aux différentes caractéristiques analysées dans les chapitres précédents : les différents regroupements des catégories morpho-syntaxiques de ThTop, les cas de reprise lexicale, la longueur des SN définis et démonstratifs. De plus, le degré d'accessibilité permet de regrouper les Thèmes topicaux et spécifiques (les ThSpe sont considérés comme des formes au degré d'accessibilité 0) et donc, de prendre en compte toutes les phrases du corpus. Pour rappel, nous redonnons ci-dessous le tableau définitoire des différents degrés d'accessibilité.

DegAccess (description - rappel)	
0	Descriptions indéfinies, formes autres et ThSpe
1	Nom propre nouveau (NP)
2	Description définie complète (SNdef)
3	Description définie courte (SNdef court) et/ou avec reprise (SNdef_R)
4	Nom propre répété (NP_R)
5	Description démonstrative avec modifieur (SNdem)
6	Description démonstrative courte (SNdem court) et/ou avec reprise (SNdem_R)
7	Pronoms et SN possessifs (ProPoss)

Tableau X.1 : Échelle des degrés d'accessibilité selon Ariel (1990) adaptée à l'étude - Rappel

Il faut rester vigilant vis-à-vis de l'interprétation des DegAccess_1, 2, 3 et 4. Concernant les DegAccess_1 et 2, il faut se souvenir que les reprises lexicales ne sont pas calculées en position S1 (nous avons considéré qu'à chaque changement de section s'opérait une sorte de réinitialisation des référents activés). Ainsi, en position S1 uniquement, les DegAccess_1 et 2 peuvent correspondre à des SN présentant une reprise lexicale (notamment dans PEOPL), le DegAccess_3 ne correspond qu'aux descriptions courtes et le DegAccess_4 est impossible. Enfin, concernant le DegAccess_4, les formes correspondantes à ce degré ne sont pas assurément co-référentielles puisque notre calcul des reprises lexicales est très lâche (tout élément dont l'occurrence a déjà été rencontrée dans la section en cours est considéré comme une reprise, voir VII.2.2.b). Notre repérage des descriptions réduites est également très brut puisqu'il ne prend en compte que le nombre de mots (VII.2.2.d). Ces réserves ne rendent pas pour autant les distinctions entre ces différents DegAccess inutilisables, les analyses précédentes ont déjà conforté cette caractérisation automatique.

Toujours concernant l'analyse des variations occasionnées par les différents INIT, notamment les adverbiaux circonstanciels et les appositions, l'environnement est étendu à la phrase suivante (Phr+1), relativement à l'hypothèse de l'encadrement du discours. Nous ne prenons pas toutes les phrases suivantes : (1) la Phr+1 doivt obligatoirement appartenir au même paragraphe que la phrase d'accueil – Phr0 (nous ne tentons pas l'exploration des continuités topicales au delà d'un changement de paragraphe); (2) elle ne doit pas comporter un INIT de même fonction (et de même rôle sémantique dans le cas des circonstants), car celui-ci entraînerait la fin de portée de l'élément détaché de Phr0; (3) afin de ne pas être biaisé par les variations selon les positions textuelles, la Phr0 doit correspondre à une phrase intraparagraphique (et non en initiale de sections ou de paragraphes). Cette étape de l'analyse nous permet de mesurer sur la portée de certains éléments détachés. Ainsi, les adverbiaux circonstanciels qui cohabitent, en Phr0, avec des formes non pronominales mais, en Phr+1, avec des Thèmes topicaux pronominaux sont très probablement des introducteurs de cadre, leur apparition entraînant un déplacement dans la progression thématique.

Les deux chapitres précédents nous ont permis de faire émerger des indices plus pertinents que d'autres pour l'étude de l'organisation discursive, car ils montrent des variations significatives :

(A) selon les **positions textuelles dans le corpus entier**

Élément	***Indice***	***Rappel de l'abréviation***
Élément détaché en initiale – INIT	Absence d'élément détaché en initiale	ssINIT
	Apposition	APPO
	Circonstant	CIRC
	Circonstant temporel	CIRCtps
Thème topical – ThTop	forme PRO3, PROdemo ou SNposs	ProPoss ou DegAcess 7
	Opposition SNdef longs / SNdef courts	DegAccess_3 / 4
Connecteur pur		Connect

(B) selon les **positions textuelles dans un sous-corpus particulier**

Élément	*Indice*	*Rappel de l'abréviation*	*Sous-corpus*
Élément détaché en initiale – INIT	Circonstant spatial	CIRCspa	ATLAS
En position Sujet Grammatical – ThTop	Reprise lexicale dans une description définie	SNdef_R DegAccess_4	
	Reprise lexicale ou réduction dans une description démonstrative	DegAccess_6	
	Redénomination par Nom propre	NP / NP_R DegAccess_1 / 2	PEOPL

(C) selon un **sous-corpus particulier, sans variation significative au niveau des positions textuelles**

Élément	*Indice*	*Rappel de l'abréviation*	*Sous-corpus*
Élément détaché en initiale – INIT	Circonstant notionnel	CIRCnot	GEOPO
	Adverbial modalisateur	MODA	

Pour les indices de type (A) et (B), il s'agit de voir si, en dehors de la position textuelle à laquelle ils sont associés, leur environnement varie ou non (en prenant compte bien entendu des variations générales observées selon les positions textuelles dans le chapitre précédent). Cela nous permet de peser le poids de l'indice de position textuelle dans les associations décelées précédemment : ces différentes formes peuvent-elles constituer des indices de séquentialité par elles-mêmes ou doivent-elles nécessairement être dans une position textuelle particulière ? Nous ne consacrons pas de partie à l'influence des connecteurs 'purs' sur leur environnement car nos analyses ne montrent pas de variations significatives très informatives. Les seules influences observées concernent leur incidence sur l'absence ou la présence d'un INIT, ce que nous avions déjà remarqué dans les chapitres précédents (il y a significativement plus de phrases sans INIT après un connecteur). Même si le chapitre précédent nous a montré des spécificités relatives aux différentes formes de connecteur 'pur' tout à fait intéressantes, nous ne disposons pas d'un assez grand nombre de données pour étudier ces formes particulières de connecteur. Ajouter une partie pour, finalement, ne pas dire grand chose n'aurait pas vraiment apporter de plus-value à ce chapitre. Dans le même souci d'économie, nous ne consacrons pas de parties aux indices de type (C). Notre observation des positions textuelles n'a rien donné et la mesure des écarts occasionnés par la présence de ces indices sur leur environnement n'a donné aucun résultat significatif, ce qui justifie tout à fait leur absence dans ce chapitre.

X.1. Les appositions en rupture?

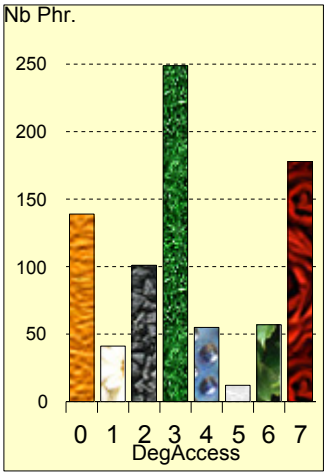
Dans le chapitre précédent, nous avons été surprise d'observer significativement plus d'apposition en initiale de sections et cela indépendamment du sous-corpus considéré (même si PEOPL montre également plus d'apposition en initiale de paragraphes, voir IX.2).

Le graphique X.1 montre la fréquence (en nombre de phrases) des cohabitations entre les appositions et les différents degrés d'accessibilité. À ses côtés, le graphique X.2 indique les écarts réduits occasionnés par la présence d'une apposition sur les différents degrés d'accessibilité. Pour rappel, le calcul de ces écarts consiste à mesurer le rapport entre la répartition des différents degrés d'accessibilité dans toutes les phrases, *i.e.* les données théoriques[181] (données présentées en VIII.6) et cette même répartition mais uniquement au niveau des phrases présentant une apposition, *i.e.* les données observées.

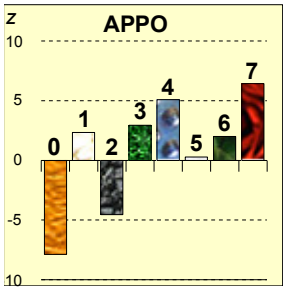
181 La présentation du calcul des données théoriques et de l'hypothèse nulle qu'elles permettent de tester est expliquée en VI.4.2.

Du point de vue de la fréquence des différentes cohabitations, nous retrouvons les DegAccess les plus fréquents au niveau du corpus entier : DegAccess_0, DegAccess_2, DegAccess_3 et DegAccess_7. Il y a cependant de grosses variations.

Comme le montre le graphique X.1, les phrases introduites par une apposition ont significativement plus de sujets de type pronominal ou possessif, nom propre répété et SN défini courts et/ou avec reprise, ce qui correspond respectivement aux DegAccess_7, 4 et 3. Pour comprendre ces variations, nous nous appuyons également sur les graphiques X.3 et X.4 de la page suivante qui indiquent, pour le graphique X.3, les variations par position textuelle et par sous-corpus (graphique) et pour le graphique X.4, la fréquence des différentes cohabitations avec une apposition dans chaque sous-corpus.



Graphique X.1 : Répartition générale des cohabitations [APPO+DegAccess_n]

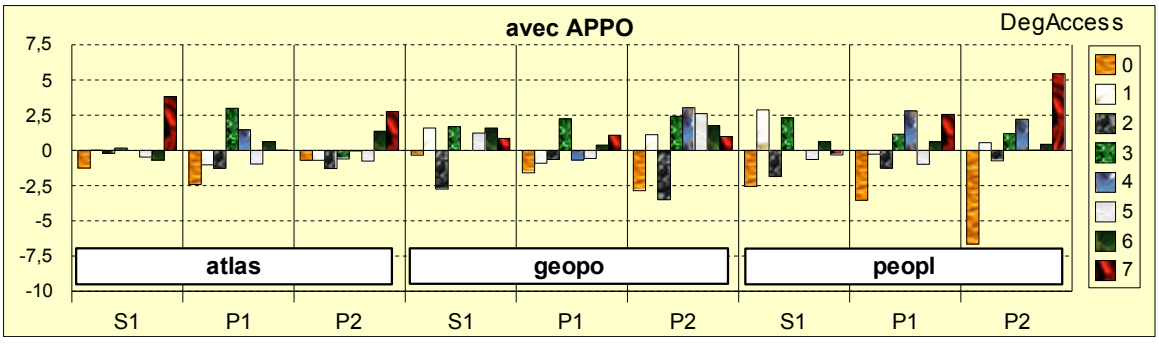


Graphique X.2 : Écarts des DegAccess dans les phrases avec APPO par rapport à toutes les phrases

X.1.1. Des signes de continuité

La cohabitation **[APPO+DegAccess_7]** observée correspond à l'idée que l'on se fait du fonctionnement discursif des appositions, à savoir l'accompagnement d'une continuation idéationnelle (V.4.2). Même en initiale de sections, on peut avoir une forme de DegAccess_7, *i.e.* Un pronom ou un possessif, à la suite d'une apposition. Et justement, c'est l'apposition qui permet une continuité topicale entre le titre et le pronom. Grâce à l'apposition, le référent exprimé dans le titre peut être réalisé en Thème topical sous la forme pronom, marque indéniable de continuité topicale (voir V.4.3.a). L'exemple X.1 illustre ce phénomène de continuité topicale entre un titre et le sujet de la première phrase de section, phénomène visible uniquement dans ATLAS. Nous revenons sur des exemples de ce type en X.5.

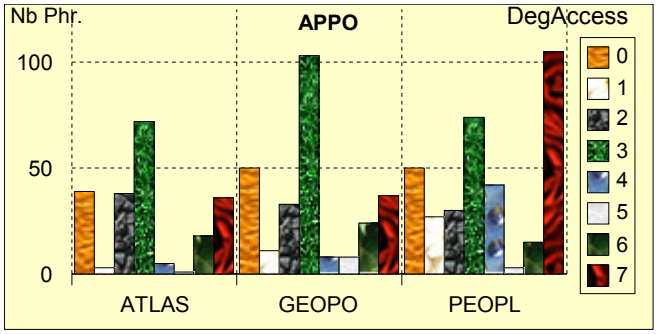
(X.1) ***SeaFrance-Sealink*** *[titre niveau 2]*
Pavillon récent, apparu le 1er janvier 1996, il *est l'héritier de diverses alliances entre la S.N.C.F. et des compagnies étrangères, depuis la privatisation de British Rail en 1984 par le gouvernement britannique. Réduisant progressivement sa présence sur le trafic transmanche, elle a concentré ses prestations sur la ligne Calais-Douvres, son principal but étant d'offrir un service de ferries à la française. Elle propose aussi diverses prestations de voyagiste.* [ATLAS_1]



Graphique X.3 : Écarts des cohabitations [APPO+DegAccess_n] par rapport au modèle général pour chaque PosTxt dans chaque sous-corpus

Si l'on observe les différents sous-corpus (graphique X.3[182]), nous voyons que cette cohabitation [APPO+DegAccess_7] est beaucoup plus fréquente dans PEOPL qu'ailleurs. Rappelons que PEOPL représente plus de 40% des appositions (voir VIII.3.2). Cette forte proportion alliée à la forte proportion de ProPoss dans PEOPL favorise cette cohabitation. Cependant, ATLAS montre également plus de ProPoss après une apposition, que ce soit en initiale de sections, comme dans l'exemple précédent, ou en position intraparagraphique.

Du côté de GEOPO, les cohabitations entre une apposition et des formes de DegAccess_3, 4 et 5 affichent toutes des écarts réduits supérieurs à +2 et cela uniquement en position P2. Nous ne nous attardons pas sur les cohabitations [APPO+DegAccess_4] et [APPO+DegAccess_5], trop peu fréquentes pour que les écarts correspondants puissent être interprétés (à peine 8 occurrences chacune, voir graphique X.4). Il semble clair que, dans GEOPO, une apposition ne peut pas cohabiter avec une forme de DegAccess_2, *i.e.* une description définie longue et sans reprise lexicale, surtout si l'on se trouve en position S1 ou P2. Nous avions vu dans le chapitre IX que ce sont justement les SN longs qui constituent en majorité le Thèmes topicaux des initiales de sections de GEOPO (47,5% des S1 dans GEOPO ont un ThTop de DegAccess_3). L'apposition a donc une incidence importante sur les débuts de sections puisqu'elle change la donne au niveau des initiales de section : ce ne sont plus les descriptions définies longues et sans reprise qui sont le plus fréquentes après une apposition, mais les descriptions définies courtes et/ou avec reprise.



Graphique X.4 : Répartition des cohabitations [APPO+DegAccess_n] dans les trois sous-corpus

Le degré d'accessibilité 4, relatif aux noms propres répétés est très spécifique à PEOPL (voir VIII.2.1). Si l'on tient compte de la fréquence de la cohabitation **[APPO+DegAccess_4]**, seule la variation en position P1 observée dans PEOPL est pertinente (le graphique X.4 montre que seul PEOPL montre un nombre raisonnable d'occurrences de cette cohabitation). Cette variation signifie que lorsqu'un paragraphe commence par une apposition, il y a fréquemment redénomination du participant – en tout cas plus souvent qu'en général (exemple X.2).

(X.2) *Car, pour **lui,** [...], le règne de l'imagination poétique, loin d'être, comme l'avait cru un romantisme superficiel, un abandon, est au contraire la conséquence d'une acuité accrue de la conscience, ce que **Baudelaire** entend lorsque, si souvent, **il** définit le poète comme étant par excellence un"homme spirituel", lorsqu'**il** se donne, comme critère esthétique universel, ce qu'il appelle le"spiritualisme".*
***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": [...] [PEOPL_5]*

Dans ATLAS et GEOPO, les reprises lexicales ne se manifestent pas au niveau des noms propres, mais au niveau des SN définis. Les phrases présentant une cohabitation **[APPO+DegAccess_3]** montrent un écart réduit *z*=+3 dans tout le corpus (graphique X.2). Cet écart se retrouve uniquement en position P1 dans ATLAS (*z*=+2,7). Nous retrouvons ici un type de cohabitation proche de celui observé précédemment dans PEOPL ([APPO+DegAccess_4]), mais dans des textes non caractérisés par une continuité autour d'un même participant et surtout ne comportant

182 Les écarts présentés dans ce graphique sont calculés par rapport aux répartitions générales des DegAccess mesurés pour chaque position textuelle dans chaque sous-corpus. Par exemple, l'écart très élevé pour le DegAccess_7 en P2 dans PEOPL signifie que, dans PEOPL, il y a beaucoup plus de ProPoss après une apposition en position intraparagraphique que dans une phrase intraparagraphique quelconque de PEOPL.

presque jamais de continuité topicale autour d'une personne (ATLAS comporte à peine 309 noms propres sujets contre 961 dans PEOPL[183]). L'exemple X.3 montre un début de paragraphe dans GEOPO où une apposition précède un SN défini court avec reprise (ici une réelle description réduite), forme au DegAccess_3. On y voit un effet de continuité entre la fin du premier paragraphe retranscrit et le début du deuxième paragraphe, introduit par l'apposition.

(X.3) *Pour éviter toute complication, le gouvernement a donc adopté une autre solution pour empêcher la diffusion de l'imagerie commerciale. Un accord commercial a été conclu entre la compagnie Space Imaging et l'agence de renseignement responsable de l'imagerie (la National Imagery and Mapping Agency, NIMA), accordant à cette dernière une exclusivité sur les images prises de l'Afghanistan.*
***Rendu public le 18 octobre, l'accord** a été signé le 5 octobre pour une durée de un mois. Pour 1,9 millions de dollars, Space Imaging s'engage à ne plus vendre d'images de l'Afghanistan et de la région (Pakistan, Ouzbékistan) à d'autres clients que la NIMA. Chaque kilomètre carré imagé est facturé 20 dollars et les commandes ne peuvent porter sur moins de 10.000 km à la fois. L'accord a été renouvelé le 5 novembre pour un second mois. [GEOPO_2]*

Ces exemples suggèrent que la cohabitation **[APPO+{DegAccess_3/4}]** permet de marquer une continuité référentielle dans les situations où la continuité est en danger, comme lorsqu'il y a un changement de paragraphe lié à un déplacement rhétorique (exemple X.3) ou que d'autres référents ont été mis en position sujet des propositions précédentes (exemple X.4).

(X.4) *La prodigieuse vitalité de cette vie aux multiples entreprises et au gigantesque travail littéraire se développe sur le terrain d'une famille bourgeoise représentative des ascensions de ce temps de mutations. La famille du père, né Balssa, est une famille de paysans du Tarn. Le père, Bernard-François, petit clerc de notaire, monte à Paris à vingt ans et finit comme directeur des vivres aux armées. La mère, née Laure Sallembier, appartient à une famille de passementiers-brodeurs parisiens. Quand Balzac naît à Tours le 20 mai 1799, le père a cinquante-trois ans et la mère vingt et un. Balzac est l'aîné de quatre enfants : Laure, la sœur bien-aimée, naît en 1800 ; Laurence en 1802 ; Henri-François en 1807, vraisemblablement fils naturel de M. de Margonne, le châtelain de Saché. **Bachelier en droit, d'abord clerc de notaire et clerc d'avoué à Paris, Balzac** décide, à vingt ans, de se consacrer à la littérature. C'est en effet sa principale occupation de 1820 à 1824, puis de 1829 à 1848, deux ans avant sa mort. Mais, de 1824 à 1828, et pendant tout le reste de sa vie, parallèlement à l'œuvre littéraire, les entreprises de tout ordre se sont succédé. En 1825, l'édition. En 1826, l'imprimerie. En 1827, une société pour l'exploitation d'une fonderie de caractères d'imprimerie. C'est l'échec ; ce sont, déjà, les dettes. Après le retour à la littérature, les années 1829-1833 sont des années d'intense activité journalistique. Des ambitions électorales se manifestent en 1831. En 1836, c'est l'entreprise malheureuse de la Chronique de Paris, revue éphémère. **En 1838, désireux d'exploiter une mine argentifère, Balzac** part pour la Sardaigne, mais, quand il arrive, la place est déjà prise. En 1839, il devient président de la Société des gens de lettres ; il milite pour tenter de sauver le notaire Peytel, accusé du meurtre de sa femme, et qui est condamné à mort par les assises de Bourg. En 1840, il lance la Revue parisienne : c'est un échec. En 1848, il se porte candidat à la députation. Quant à ses candidatures à l'Académie française, elles sont toujours restées sans succès. [PEOPL_7]*

Nous remarquons également ce phénomène à la suite d'un circonstant en INIT1 (premier élément détaché), en considérant que la présence d'un circonstant en INIT1 peut constituer un 'danger' pour la continuité topicale en déplaçant les circonstances d'évaluation du référent topique. Dans ce cas, l'apposition est en INIT2 (dexuième élément détaché), ce qui permet au Thème topical de réaliser une progression thématique constante comme dans l'exemple X.5.

(X.5) *De septembre 1777 à janvier 1779, c'est le grand voyage à Paris. Il [Mozart] part, accompagné seulement de sa mère, et l'aventure sera très décevante sur le plan du sentiment (son amour déçu pour Aloisia Weber), de la famille (sa mère meurt à Paris) et de sa carrière (il est évincé des milieux musicaux de la capitale et lâché par le baron Grimm, son protecteur). Par contre, sur le plan musical, ce voyage sera très fructueux : à l'aller, il s'arrête longuement à Mannheim où il découvre les puissances expressives de l'orchestration romantique moderne. **À Paris, lui qui depuis toujours est hanté par le désir d'écrire des opéras, il** tombe en plein dans la lutte entre piccinnistes et gluckistes ; mais il ne s'y engage pas parce qu'il prend déjà conscience du style de théâtre musical qui sera le sien. Par-dessus tout, ce séjour à Paris aura une importance capitale du fait que Mozart capte de l'esprit français - sans en retenir la sécheresse - le goût*

183 La plupart des noms propres d'ATLAS font référence à une société, comme en X.24. Le dernier texte d'ATLAS portant sur l'évolution politique de l'Ouest ces dernières années comporte des références à des hommes politiques, mais rarement en position topique (les topiques sont plutôt les partis politiques dont le nom n'est pas catégorisé nom propre par Syntex dans la version que nous avons utilisée pour le programme).

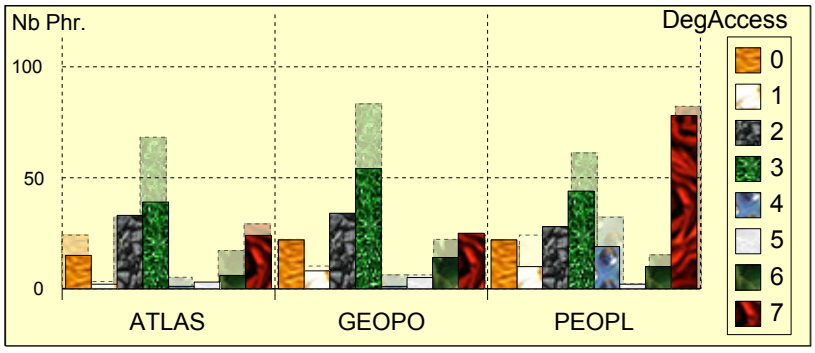
de la pudeur, de l'élégance et de la concision. Il aura dès lors plus que jamais horreur de la longueur et de l'emphase oratoire (ce qu'il appelle le"goût long des Allemands"). [PEOPL_16]

Si on calcule les variations occasionnées par la présence d'un adverbial circonstanciel suivi d'une apposition en INIT (les APPO en INIT2 sont généralement précédées d'un INIT1 à fonction circonstancielle[184], voir VIII.3.4), nous remarquons que seul le DegAccess_4 (nom propre répété) est significativement plus importante (*z*=+3,9). Ce résultat est à mettre en relation avec plusieurs faits : le fait que PEOPL, qui montre le plus d'apposition en INIT2, montre également beaucoup de noms propres répétés et le fait que, dans PEOPL, les adverbiaux circonstanciels cohabitent également avec noms propres répétés (voir partie suivante).

On voit bien que les appositions s'inscrivent dans des contextes de continuité (position intraparagraphique, cohabitation avec des pronoms, des possessifs, des redénominations et des réductions de terme), même si quelques différences entre sous-corpus sont observées. Plus encore, il semble que la présence d'apposition en initiale de phrase permet de maintenir une continuité topicale menacée (par un changement de paragraphe ou la présence d'un circonstant en INIT1) ou fragile (car le référent n'a été introduit que dans le titre de section ou qu'il n'a pas été mentionné depuis un certain temps). Il n'y a là aucun indice de rupture à voir, mais plutôt **un indice très fort de continuité topicale.**

X.1.2. Des contextes de continuité

Nous allons maintenant observer ce qui se passe dans la phrase suivant celle introduite par une apposition (Phr+1). Selon les restrictions décrites en introduction de ce chapitre, on dénombre 499 Phr+1 après une apposition. Le graphique X.5 indique la répartition des degrés d'accessibilité dans la phrase succédant à la phrase d'accueil de l'apposition. Pour mieux se rendre compte de la différence des répartitions entre Phr+1 et Phr0, nous avons fait apparaître en ombre la répartition des degrés d'accessibilité en Phr0. Cette représentation ne permet malheureusement pas de voir les augmentations d'un DegAccess des Phr0 aux Phr+1. Seuls les graphiques indiquant les écarts réduits le permettent (graphique X.6).



Graphique X.5 : Répartition des Degrés d'accessibilité dans les Phr+1 d'APPO

La répartition des différents DegAccess restent approximativement les mêmes qu'au niveau des phrases d'accueil (Phr0). La baisse de fréquence est principalement due au fait que l'on passe de 642 Phr0 à 499 Phr+1, soit 22% de moins. Le seul degré qui ne baisse pas de fréquence est le DegAccess_2, correspondant aux descriptions définies complètes. Il y a donc beaucoup plus de descriptions définies complètes en Phr+1 qu'en Phr0. En mesurant l'écart réduit entre la répartition des DegAccess en Phr0 et en Phr+1, nous obtenons un seul écart significatif correspondant précisément au DegAccess_2 qui montre en Phr+1 un *z*=+6,2. Les degrés d'accessibilité qui baissent le plus fortement

184 Sur les 210 INIT2 de type APPO, 63% sont précédées d'un INIT1 CIRC et 27% d'un INIT1 également APPO.

(au delà de 22%) correspondent aux descriptions définies co-référentielles et au DegAccess_0 (SNindef, ThTop autres et ThSpe). Concernant les ProPoss, il semble y en avoir autant en Phr+1 qu'en PHr0, ce qui signifie qu'il est assez fréquent de voir une continuité topicale marquée par une cohabitation [APPO+DegAccess_7] continuer dans la phrase suivante (presque 10% des phrases avec APPO montrent cette progression). L'exemple X.6 illustre ce type de continuités particulières à PEOPL, où trois phrases se succèdent : les deux premières introduites par une cohabitation [APPO+DegAccess_7(ProPoss)], la dernière composée d'un Thème topical seul de forme pronominale.

(X.6) *D'Ecbatane, en hiver 324, Alexandre organise l'exploration de l'Arabie et celle de la Caspienne. Selon Diodore et les Hypomnemata, conservés par Eumène, Alexandre avait conçu des projets plus vastes encore, la conquête de l'Afrique et de l'Espagne. En novembre, la mort d'Héphaistion l'atteint profondément : il reste plusieurs jours couché près du cadavre, sans prendre de nourriture. Enfin, il lui accorde les honneurs funèbres dus à un héros. Lui-même souffre d'un grand épuisement. Cependant, il se reprend une fois encore grâce à son énergie. De Babylone, où il rentre en février 323, il entreprend de grands travaux, reçoit des ambassadeurs venus de fort loin : de Carthage, d'Italie, de Gaule, peut-être de Rome.* ***Continuant son œuvre de colonisateur, il*** *fonde des villes, futurs centres commerciaux, telle Alexandrie Charax près de l'embouchure du Tigre.* ***Frappé par la malaria, il*** *est emporté en douze jours.* ***Il*** *meurt le 13 juin 323, à l'âge de trente-trois ans. [PEOPL_27]*

Il est évident que dans PEOPL les continuités topicales sont facilitées tout le long du texte. Mais on peut également trouver de telles suites dans GEOPO ou ATLAS, comme le montre l'exemple X.7. Ces cas restent tout de mêmes rares.

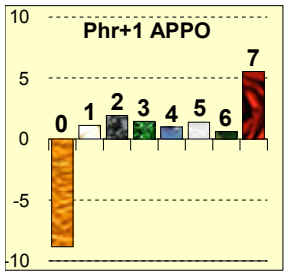
(X.7) *La politique en Irak ne se réduisait donc pas à l'exercice permanent de la violence. En encourageant la renaissance des tribus et les phénomènes de communautarisation en général, le régime s'est placé au coeur d'un jeu dont il possédait seul la capacité d'arbitrage.* ***Unique garantie d'un quelconque intérêt général, il*** *s'est investi d'un rôle modérateur des tensions qu'il a lui-même attisées...* ***Il*** *tirait d'ailleurs l'essentiel de sa légitimité, y compris auprès de ses détracteurs, de cette qualité minimale de rempart contre le chaos. Pourquoi la disparition du pouvoir signifierait-elle forcément l'anarchie? Déjà, parce que trop d'armes circulent au sein de la population irakienne.[...] [GEOPO_9]*

9,5% des 499 Phr+1 correspondent à ce type de continuités (DegAccess(Phr0)=7, DegAccess(Phr+1)=7). L'autre type de séquence fréquemment observée correspond à une succession de DegAccess_3, *i.e.* une succession de phrases ayant en Thème topical un SNdef court et/ou présentant une reprise lexicale (10,5% des cas). Mais si l'on observe les données (54 séquences de phrases), on se rend rapidement compte qu'il s'agit rarement de continuité topicale. En fait, le sujet de la Phr+1 peut tout à fait être une description courte sans lien avec le sujet de la phrase précédente ou un SNdef avec reprise qui ne reprend pas le Thème topical de la phrase précédente (mais un autre élément dans une phrase plus lointaine). Quelques exemples correspondent cependant à des cas de continuité, telle que la progression thématique dérivée observée en (X.8), mais ces cas sont plutôt rares, une dizaine tout au plus.

(X.8) *Malgré les récents votes budgétaires, les États sont encore dans une situation des plus précaires. Si la crise actuelle sert jamais de révélateur, le manque de coordination entre États constitue un handicap lourd. En tous les cas, la crise actuelle permet de relativiser les baisses d'impôt au niveau fédéral, tant vantées par l'équipe actuelle. Leur vote n'a été possible que suite à un désengagement réel de l'État fédéral qui laisse ses partenaires fédérés confrontés à des charges nouvelles, notamment en termes de politique sociale, sans financement.* ***Déjà fragilisés par le contexte économique d'ensemble et leur propre faiblesse institutionnelle, les États*** *sont ainsi placés en position très difficile. Actuellement,* ***les États qui ont une tradition d'activisme public*** *restent les plus menacés : par exemple le Minnesota, New York, le Connecticut, et enfin la Californie, où la crise budgétaire se double d'une crise politique grave. [GEOPO_21]*

Si l'on compare la répartition des degré d'accessibilité dans les Phr+1 d'APPO à la répartition générale observée en position P2, nous obtenons le graphique X.6. Le lien entre la présence d'un apposition et les phénomènes de continuité topicale apparaissent là encore : les phrases succédant une phrase comportant une apposition montrent significativement plus de formes pronominales et possessives que n'importe quelle autre phrase intraparagraphique. Dans le même ordre d'idée, il y a significativement moins de formes au degré d'accessibilité faible comme les SN indéfinis ou les Thèmes spécifiques.

En mesurant pour chaque sous-corpus les variations entre la répartition des DegAccess après une apposition et la répartition des DegAccess en position P2, nous voyons que seul le sous-corpus PEOPL montre un schéma similaire à celui présenté par le graphique X.6. ATLAS et GEOPO ne montrent aucune variation significative.



Graphique X.6 : Écarts des DegAcces des Phr+1 d'APPO par rapport au modèle général (phrases P2)

Nous retrouvons ici l'hypothèse posée par Combettes (2005) concernant la possibilité d'un pouvoir cadratif des appositions (voir V.4.2). D'après nos données, ce pouvoir cadratif ne serait possible que dans certains contextes textuels, ici dans PEOPL. Cependant, nous n'avons trouvé aucun exemple où ce pouvoir cadratif est effectivement réalisé.

°, °, °, °

Que ce soit au niveau du degré d'accessibilité de la phrase contenant l'apposition ou dans la phrase suivante, toutes les données indiquent que l'apposition est un indice fort de continuité topicale. Ce sens instructionnel est stable à travers les différents types de texte considérés. De plus, la force de cet indice semble telle qu'elle permet de maintenir des continuités topicales par delà les différents niveaux de segmentation. Ainsi, les appositions permettent le maintien d'une continuité topicale malgré des déplacements au niveau du fond ou de l'articulcation rhétorique.

En dehors de ces chevauchements de paragraphes ou de sections, l'apposition permet simplement d'attribuer au référent qui la suit le statut de topique.

X.2. Des circonstants aux rôles différents

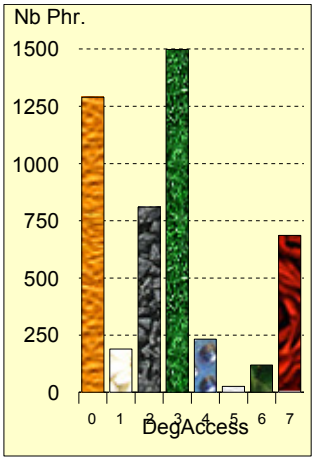
De façon générale, il est ressorti des deux chapitres précédents une association entre la présence d'un adverbial circonstanciel et la position initiale de sections ou de paragraphes. Cette association varie selon les sous-corpus. Ainsi, dans ATLAS les circonstants sont plutôt en P1 alors que dans PEOPL, ils apparaissent plus en S1. Entre ces deux extrêmes, les écarts observés dans GEOPO sont juste en dessous de la limite de signifiance, ceux en S1 égalant ceux en P1. Ce comportement des circonstants ne se retrouve pas toujours au niveau des rôles sémantiques distingués dans notre thèse. Alors que les adverbiaux temporels ne montrent pas de comportement différent dans les sous-corpus, ils affichent des variations significatives selon les niveaux de segmentation, semblant jouer à un niveau plutôt global d'organisation (cf. Bestgen & Vonk 1995, 2000 Bestgen & Costerman 1997). Dans un autre registre, la localisation spatiale par les adverbiaux temporels semble rester à un niveau intraparagraphique et ce, uniquement chez ATLAS. Les adverbiaux notionnels (associés à GEOPO) et les autres types de circonstants (associés à PEOPL) ne varient pas significativement selon les positions textuels et semblent simplement caractéristiques d'un type de texte.

X.2.1. Les circonstants en général : un indice de déplacement

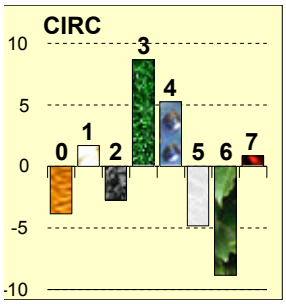
Si l'on mesure la répartition des cohabitations entre la présence d'un adverbial circonstanciel et les différents degrés d'accessibilité, nous obtenons le graphique X.7. Nous y voyons une très grande fréquence de cohabitation

[CIRC+DegAccess_3] (SNdef court et/ou présentant une reprise). Cette cohabitation est deux fois plus grande que celle entre CIRC et DegAccess_2, ce qui semble correspondre davantage à un déplacement qu'à une rupture.

La cohabitation [CIRC+DegAccess_3] est également deux fois plus fréquente que la cohabitation **[CIRC+DegAccess_7]**. Les adverbiaux circonstanciels semblent correspondre davantage à des indices de déplacement que de continuité topicale. La différence de fréquence entre les cohabitations [CIRC+DegAccess_3] et [CIRC+DegAccess_7] est toutefois moins importante qu'entre [CIRC+DegAccess_3] et [CIRC+DegAccess_2], ce que souligne le graphique X.8. On y voit un écart réduit nul au niveau de la cohabitation [CIRC+DegAccess_7], alors que l'écart pour la cohabitation [CIRC+DegAccess_2] est de -2,7.

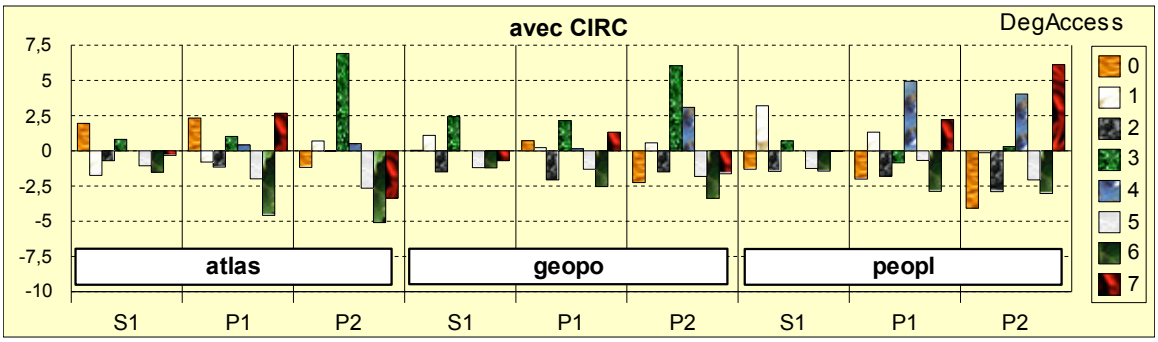


Graphique X.7 : Répartition générale des cohabitations [CIRC+DegAccess_n]



Graphique X.8 : Écarts des DegAccess dans les phrases avec CIRC par rapport à toutes les phrases

Cette corrélation CIRC/déplacement se retrouve bien dans les écarts réduits mesurés au niveau de la répartition des différents DegAccess dans les phrases introduites par un circonstant. La forte fréquence de DegAccess_3 après un adverbial circonstanciel est significativement plus importante que dans le modèle général. Un écart similaire est observé au niveau de la cohabitation **[CIRC+DegAccess_4(nom propre répété)]**, moins fréquente.

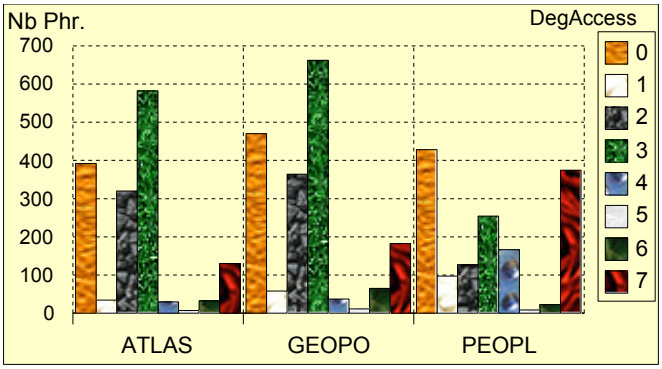


Graphique X.9 : Écarts des cohabitations [CIRC+DegAccess_n] par rapport au modèle général pour chaque PosTxt dans chaque sous-corpus

Au niveau des variations selon les sous-corpus (graphique X.9), un comportement similaire se dessine chez ATLAS et GEOPO. Ce comportement suit ce que nous avons noté plus haut : la présence d'un circonstant en initiale de phrase semble entraîner l'emploi de formes au DegAccess_3 (SNdef courts et/ou avec reprise), ce qui signifie une corrélation CIRC/déplacement. Par contre, PEOPL ne montre pas du tout les mêmes schémas et semble associer à chaque position textuelle une (ou des) cohabitation(s) particulière(s), à savoir : [CIRC+DegAccess_1] en S1, [CIRC+DegAccess_3] en P1 et [CIRC+DegAccess_3/7] en P2. La cohabitation en S1 est à prendre avec beaucoup de précautions. En effet, notre programme d'annotation ne caractérise pas le fait qu'il y ait reprise en initiale de section, alors qu'il est tout à fait possible d'avoir une reprise en S1, surtout dans PEOPL où une redénomination peut très souvent apparaître en première phrase de section, comme l'illustre l'exemple X.9.

(X.9) ***4.2) L'"œuvre en cours"***[titre niveau 2]
En 1920, ***Joyce*** *arrive à Paris, venant de Zurich où Dada avait inauguré, en 1916, une négation culturelle globale. Au même moment, le cubisme s'installe, Gertrude Stein épaulant Picasso. L'agression surréaliste s'affirme. [...]*
4.3) Le rabâchage de la comédie humaine [titre niveau 2]

> ***Joyce*** *écrivit les deux premières pages de Finnegans Wake le 10 mars 1923. Elles devinrent les pages 380 et 382 de l'œuvre terminale. [...]*
> ***4.4) Du particulier à l'universel*** [titre niveau 2]
> *Comme dans Ulysse,* ***Joyce*** *voulut donner à son entreprise une structure propre à la tirer du particulier vers l'universel : c'est ainsi que Finnegans Wake évolue au gré d'une double structure de pensée constituée d'abord par l'affrontement dialectique de couples opposés, où l'on reconnaît une idée clé du philosophe Giordano Bruno, retrouvée par Blake :"Sans l'opposition des contraires, pas de progrès."[...]* *[PEOPL_2]*



Graphique X.10 : Répartition des cohabitations [CIRC+DegAccess_n] dans les trois sous-corpus

Le plus intéressant est ce qui se passe en position P2 et notamment le double écart positif des cohabitations **[CIRC+DegAccess_7(ProPoss)]** et **[CIRC+DegAccess_4(NP_R)]**. La variation de [CIRC+DegAccess_7] se retrouve dans le graphique X.10 où l'on voit, dans PEOPL, le très grand nombre de ces cohabitations comparé aux autres sous-corpus. Cette fréquence élevée est liée au fait que les phrases intraparagraphiques sont plus nombreuses que dans les autres positions textuelles et que, dans PEOPL, en P2, les adverbiaux circonstanciels cohabitent préférentiellement avec des pronoms ou des SN possessifs. Cette forte fréquence de [CIRC+DegAccess_7] signifie que les adverbiaux circonstanciels montrent un fonctionnement différent dans PEOPL. Dans les autres sous-corpus, la présence d'un adverbial circonstanciel va de pair avec d'autres indices de déplacement et plus particulièrement l'indice de reprise lexicale. Dans PEOPL la présence d'un circonstant n'entrave pas toujours la continuité topicale puisqu'il y a préférentiellement une forme au degré d'accessibilité élevée après un tel adverbial. Nous verrons, dans la partie X.4.1 les écarts conséquents de la présence de formes de DegAccess_7 dans PEOPL. Par contre, le fait d'avoir un autre écart positif pour la cohabitation [CIRC+DegAccess_4(nom propre répété)] peut signifier que dans certains cas, la présence d'un adverbial circonstanciel a une incidence sur la continuité topicale. Nous gardons toutefois des réserves quant à cette incidence, puisque comme l'a remarqué Schnedecker (2005), les noms propres répétés peuvent, dans le cas de portraits de personnalités, être employés en alternance aux pronoms de 3e personne.

°, °, °, °

Nous avons remarqué dans les chapitres VIII et IX que certains rôles sémantiques semblaient avoir une certaine influence sur la nature des Thèmes topicaux qui les suivent. Ainsi, les Thèmes topicaux en collocation avec les adverbiaux temporels et les adverbiaux « autres » restent associés significativement au même sous-corpus que celui auxquels ils sont associés sans tenir compte du rôle sémantique ou de la fonction d'INIT (voir partie VIII.7). À l'opposé, les adverbiaux spatiaux et notionnels (ainsi que les modalisateurs et l'absence d'INIT) modifient les variations observées d'un point de vue général. Les sous-sections suivantes vérifient ces premiers résultats en étudiant les variations sur l'environnement des adverbiaux temporels et spatiaux. Nous ne ferons pas cas des adverbiaux notionnels et des adverbiaux circonstanciels « autres ». Pour les premiers, aucune variation significative n'est observée, dans aucun sous-corpus (les adverbiaux notionnels sont une spécificité de GEOPO). Pour les seconds, aucun rôle sémantique n'est déterminé. Les circonstants « autres » peuvent exprimer des conditions, des buts, de la

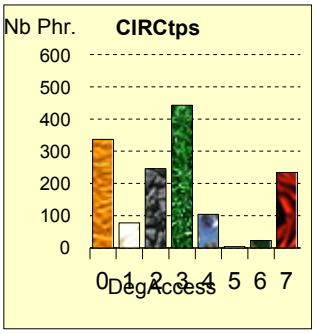
localisation spatiale ou temporelle non caractérisée, etc... De fait, nous n'aurions pas pu interpréter les variations observées et leur analyse aurait été redondante avec l'analyse des circonstants en général.

X.2.2. Les adverbiaux temporels sans influence sur la continuité topicale

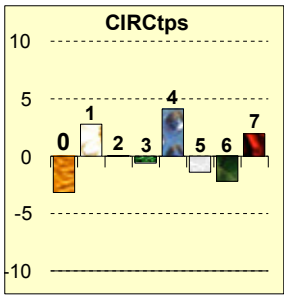
Suite au chapitre IX, nous savons que les adverbiaux temporels apparaissent communément en initiale de sections et de paragraphes, ce qui leur confère une forte chance pour constituer des indices de discontinuité. Cependant, il faut s'interroger sur le poids des adverbiaux temporels dans la mise en place d'une relation de discontinuité. Pour examiner cette question, nous observons l'environnement des adverbiaux temporels en dehors de ces positions textuelles particulièrement associées aux phénomènes de discontinuité. Si en position intra-paragraphique les adverbiaux temporels s'entourent significativement de degrés d'accessibilité faibles et associés à un effet de discontinuité, alors l'idée que les adverbiaux temporels signalent un déplacement ou une rupture sera en partie validée. Par contre, si aucune variation pertinente n'est observée cela signifiera soit que ces adverbiaux ne sont pas structurants, soit que les TSC temporelles (dans lesquelles les adverbiaux temporels ont un pouvoir structurant) sont dominées par un autre mode organisationnel.

Le graphique X.11 montre la répartition des cohabitations entre la présence d'un adverbial temporel et les différents degrés d'accessibilité. Le graphique X.12 indique les écarts réduits observés au niveau de la répartition des degrés d'accessibilité selon qu'il y ait ou non un adverbial temporel en initiale de phrase.

Les adverbiaux temporels constituent le rôle sémantique le plus fréquent des adverbiaux circonstanciels. Comme le montre le graphique X.11, les adverbiaux temporels montrent quasiment les mêmes proportions de cohabitation que les adverbiaux circonstanciels en générale (en comparant avec le graphique X.7, infra). Au niveau du calcul des écarts réduits par contre, les adverbiaux temporels montrent des variations spécifiques. Tout d'abord, il y a moins d'écarts significatifs qu'au niveau des circonstants en général (en comprenant avec le graphique X.8 p.266). Les seuls écarts positifs se situent au niveau des cohabitations avec des noms propres : [CIRCtps+DegAccess_1] et [CIRCtps+DegAccess_4], alors que pour les circonstants en général, il y a en plus un écart positif important au niveau du DegAccess_3 (SN défini court et/ou avec reprise). Ces premières observations vont dans le même sens que les résultats des chapitres précédents : le temps ne semble pas vraiment occasionner de variations sur son environnement.



Graphique X.11 : Répartition générale des cohabitations [CIRCtps+DegAccess_n]

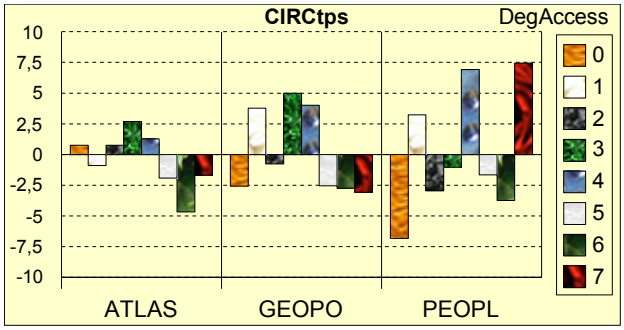


Graphique X.12 : Écarts des DegAccess dans les phrases avec CIRCtps par rapport à toutes les phrases

Les écarts observés au niveau des DegAccess_1 et 4 nous font penser que les seules variations mesurées proviennent en fait du sous-corpus PEOPL, sous-corpus affichant le plus de noms propres. Si l'on observe les écarts selon les différents sous-corpus (graphique X.13), nous voyons effectivement que c'est dans PEOPL que la plupart des écarts sont mesurés. L'écart le plus important concerne la cohabitation **[CIRCtps+DegAccess_7],** ce qui signifie qu'après un adverbial temporel, il y a plus de pronoms et de possessifs que dans le modèle général. Cet écart se situe essentiellement en position intraparagraphique comme le montre le graphique X.14. Ainsi, même en position P2 où l'on a déjà plus de ProPoss, la présence d'un adverbial temporel amène à utiliser davantage de ProPoss. Cela signifie que

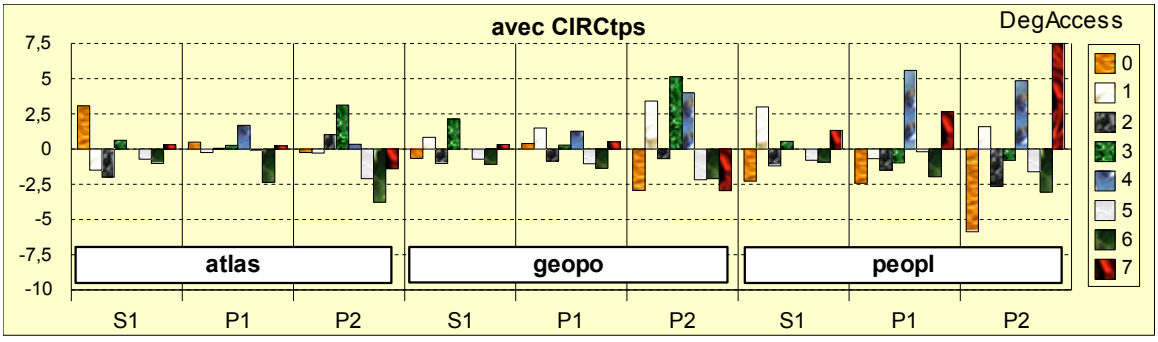
la présence d'un adverbial temporel n'entrave pas une continuité topicale en cours, ce qui va à l'encontre de nos hypothèses sur le pouvoir structurant des adverbiaux temporels (un exemple est donné plus bas en (X.14).

PEOPL montre deux autres cohabitations significatives : les cohabitations [CIRCtps+DegAccess_1 et 4]. Ces cohabitations sont également présentes dans GEOPO qui ajoutent à celles-ci la cohabitation [CIRC+DegAccess_3].



Graphique X.13: Écarts des DegAccess dans les phrases avec CIRCtps par rapport à toutes les phrases pour chaque sous-corpus

Ce qui est étonnant ici, c'est sans doute davantage la cohabitation [CIRC+DegAccess_1] que [CIRC+DegAccess_4]. Dans PEOPL et GEOPO, la présence d'un adverbial temporel peut occasionner la présence d'un nom propre nouveau. Dans PEOPL, cette cohabitation n'est significative qu'en initiale de sections (X.14), ce qui ne nous apprend pas grand chose. Par contre, dans GEOPO, la cohabitation [CIRCtps+DegAccess_1] se rencontre en position P2, ce qui tend à prouver que, dans GEOPO, la présence d'un adverbial temporel implique une redénomination ou l'arrivée d'un nouveau référent Thème. ATLAS ne montre pas vraiment d'écart significatif[185] (celui observé au niveau du DegAccess_4 est juste à *z*=+2,7). Même en distinguant chaque position textuelle (graphique X.14), les écarts restent faibles.

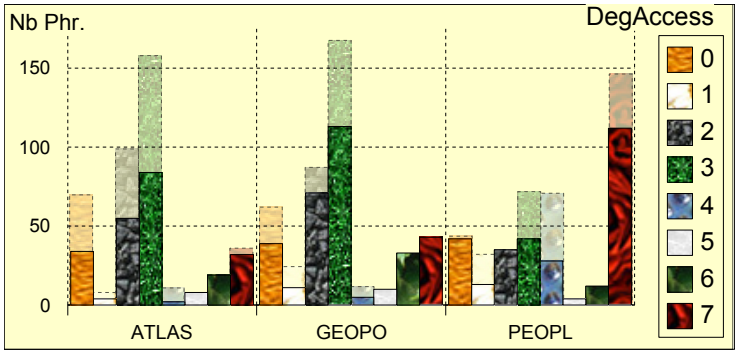


Graphique X.14: Écarts des cohabitations [CIRCtps+DegAccess_n] par rapport au modèle général pour chaque PosTxt dans chaque sous-corpus

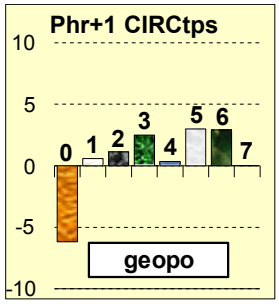
Les écarts observés dans GEOPO vont tous dans le sens d'une corrélation entre la présence d'un circonstant temporel et les phénomènes de discontinuité, le signalement des ruptures étant corrélé à la cohabitation [CIRCtps+DegAccess_1(nom propre nouveau)] et celui des déplacements à la cohabitation [CIRCtps+DegAccess_3(SNdef_R/courts)] et [CIRCtps+DegAccess_4(nom propre nouveau)]. Ces corrélations se confirment si l'on observe le degré d'accessibilité de la Phr+1 des adverbiaux temporels. Le graphique X.16 indique

185 Nous ne commenterons pas l'écart négatif des cohabitations [CIRCtps+DegAccess_6(SN démonstratifs)], les données étant trop faibles dans ATLAS pour être interprétables. On peut d'ailleurs remarquer que les descriptions démonstratives en position Thème sont bien rares dans notre corpus, c'est pourquoi nous ne nous arrêtons pas souvent sur ce type de Thème topical, sauf dans la suite de cette partie...

que dans les phrases suivantes celles introduites par un adverbial temporel les DegAccess_3, 5 et 6 sont significativement plus présents[186].



Graphique X.15: Répartition des DegAccess dans les Phr+1 de CIRCtps dans les trois sous-corpus



Graphique X.16: Écarts des DegAccess dans les Phr+1 de CIRCtps dans GEOPO par rapport au modèle de GEOPO (P2)

Concernant la fréquence de ces cohabitations dans GEOPO (graphique X.15), nous voyons qu'il y a pour une fois un nombre non négligeable de formes au DegAccess_6 (*i.e.* de SN démonstratifs courts et/ou avec reprise). Les SN démonstratifs ont un emploi majoritairement co-référentiel avec la phrase précédente (voir V.4.3.c, notamment les résultats du travail de Dupont 2003), ce qui signifie que le référent exprimé après un adverbial temporel « continue » en Phr+1. Cependant, les SNdem sont également associés à un phénomène de déplacement, ce que soulignent également les travaux de De Mulder (1994, 1997, 2000). Les démonstratifs sont généralement utilisés pour co-référer par reclassification, *i.e.* référer à un référent déjà mentionné mais sans maintien des circonstances qui ont permis d'introduire ce référent. Les exemples trouvés dans GEOPO montrent tout à fait ce type de fonctionnement des SNdem dans les Phr+1 de CIRCtps. Dans l'exemple X.10, le référent « les écoles » est introduit par un adverbial temporel qui pourrait ouvrir un cadre temporel. Cependant, cet adverbial temporel ne sert qu'à introduire le référent sans pour autant introduire un cadre temporel (« *des écoles* » constitue alors le premier Thème d'une progression thématique linéaire).

(X.10) ***UN SCÉNARIO POSSIBLE POUR UNE GUERRE ANNONCÉE*** [titre niveau 1]
La plasticité du régime [de Saddam Hussein] est un facteur ignoré dans toutes les anticipations de la guerre. Constatant que les "options militaires" de l'Irak sont limitées, les analystes n'envisagent comme alternative à ces options classiques que le scénario catastrophe des "armes de destruction massive". [...] Les "frappes" et autres ingérences étrangères l'ont préparé à cette confrontation ultime. Elles lui ont appris à escamoter ses cibles les plus vitales, à savoir la personne physique des hauts responsables, les missiles sol-air de la Défense aérienne et d'éventuelles armes de destruction massive. [...] Le régime escamote parfois jusqu'aux cibles les plus ordinaires. ***Lors des bombardements de 1998, des écoles,*** *ainsi que des installations industrielles et des hangars alimentaires, ont accueilli des dépôts de munitions.* ***Ces écoles*** *abritent actuellement les membres du Parti chargés de maintenir l'ordre dans chaque quartier. Ceux-ci ont quitté leurs locaux officiels, imitant l'ensemble de l'appareil de sécurité. [GEOPO_11]*

Cette fonction d'introduction sans portée des adverbiaux temporels semble très liée à la situation des adverbiaux temporels dans le texte. Un adverbial temporel situé en initiale de sections ou de paragraphes porte généralement sur l'ensemble de l'unité, sauf indication contraire (notamment l'expression d'une nouvelle localisation temporelle) comme le montre l'exemple X.11. Cet exemple montre un adverbial temporel en début de section qui porte jusqu'à l'apparition d'une autre localisation temporelle dans la première phrase du deuxième paragraphe : *en 1995* (complément circonstanciel intégré et situé en position non initiale).

(X.11) ***La Bosnie*** [titre niveau 3]
En 1993, *les débats sur la possibilité d'un engagement militaire en Bosnie se sont accompagnés de nombreuses critiques sur le coût des opérations, la nécessité d'engager les forces armées et la participation des alliés européens. Les parlementaires se sont interrogés sur les pouvoirs de guerre accordés au président Bill Clinton [...]. Certains*

186 ATLAS, dont nous n'avons pas affiché les résultats, ne montre pas d'écart significatif, sauf pour les formes au DegAccess_0.

membres du Congrès se sont montrés sceptiques quant aux prétentions de la Maison-Blanche à jouer les "gendarmes du monde". Ils estimaient que le pays, en intervenant de façon excessive à l'extérieur, risquait de gaspiller ses ressources et de dévoiler trop aisément ses forces à ses adversaires. Le sénateur John McCain (républicain, Arizona) pensait ainsi que, "si nous usons de nos forces et de notre prestige de façon inconsidérée, nous gaspillerons des ressources que nous n'avons pas".
Des remarques du même type ont été exprimées ***en 1995,*** *notamment de la part d'élus démocrates, qui jugeaient que l'Administration avait prêté une trop grande attention aux aspirations de l'aile conservatrice du Congrès en décidant d'envoyer des troupes dans les Balkans. Le sénateur Byron Dorgan (démocrate, Dakota-du-Nord) estimait ainsi que "[...]". Pour sa part, et conscient de l'importance de la participation financière que supposait une intervention armée, le représentant Jerry F. Costello (démocrate, Illinois) considérait que la Bosnie concernait les Européens au premier chef et que ceux-ci devaient en assumer la principale responsabilité, écartant ainsi le principe d'un envoi de troupes américaines sur le terrain. [GEOPO_20]*

Cet exemple montre bien un effet de portée de la référence temporelle *En 1993.* L'adverbial oriente toute l'interprétation du premier paragraphe et sans la présence du SP *en 1995*, il semble bien que la référence temporelle correspondant à l'année 1993 persisterait dans le second paragraphe. En dehors des initiales de sections et de paragraphes, les adverbiaux temporels ne se montrent pas vraiment de pouvoir structurant, sauf lorsque ceux-ci sont impliqués dans une configuration de TSC temporelle locale (*i.e.* une structure énumérative indexée temporellement à l'intérieur d'un paragraphe), ce qui est finalement assez rare. Dans les cas de TSC locale, les adverbiaux temporels ont bien un pouvoir structurant. Cependant, ce n'est pas la seule présence des adverbiaux temporels qui construit la TSC. Nous remarquons qu'en général les TSC locales sont annoncées dès le début du paragraphe par une phrase amorce type[187] ou la présence en initiale de paragraphes d'un adverbial temporel qui, alors, a double fonction : amorcer la TSC et indexer le premier item de TSC (exemple X.12).

(X.12) ***Au début des années 1970,*** *les objectifs de l'intervention publique sur le marché pétrolier changèrent brutalement. Du soutien des prix intérieurs par la réglementation de l'offre et des importations, on passa à la lutte contre les effets de la hausse des prix. On pourrait dire : [...]. On citera en particulier [...].* ***À la fin des années 1970,*** *l'administration Carter souhaitait libéraliser le marché pétrolier mais se heurtait à de fortes résistances au Congrès. Une loi votée en 1978 prévoyait un décontrôle progressif étalé sur 10 ans ; l'administration Reagan le réalisa en un mois. [GEOPO_12]*

En dehors des positions S1 et P1 et des TSC temporelles, les adverbiaux temporels ne montrent pas de portée allant au delà de leur phrase d'accueil et servent plutôt à introduire un nouveau référent. Il semble donc que le cas d'un cadre temporel intraparagraphique isolé soit très rare. En d'autres termes, pour avoir un pouvoir structurant, les adverbiaux circonstanciels de temps doivent apparaître en initiale de sections ou de paragraphes ou dans des TSC temporelles. Autrement, ils servent plutôt à l'introduction d'un nouveau référent (exemples X.10 et X.13 ci-dessous).

(X.13) ***2/ Les débuts de la contestation des mesures gouvernementales.*** [titre niveau 2]
Lannée 2002 n'a pas manqué de renforcer ces initiatives sécuritaires. ***Dès le mois de janvier,*** *et avec le soutien du Président, J. Ashcroft, Ministre de la Justice, a tenté de lancer son projet TIP (Terrorism Information and Prevention System), dont l'idée principale était d'encourager les Américains à rapporter toute activité " suspecte " en téléphonant à un numéro vert.* ***Ce "système d'information et de prévention terroriste"****, partie intégrante du " White House Citizen Corps Program " et destiné à s'appliquer initialement dans 10 villes, a soulevé un tel tollé qu'Ashcroft a dû battre en retraite. Même Dick Armey, un des " durs " de la Chambre des Représentants, était réticent ! [GEOPO_22]*

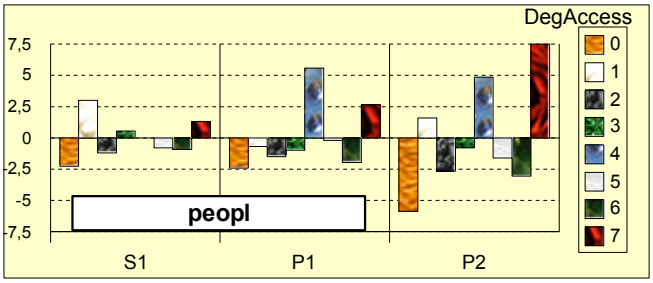
Dans PEOPL, le fonctionnement des adverbiaux temporels est légèrement différent puisqu'il n'y a que rarement introduction d'un nouveau référent (le référent topical étant tout au long du texte la personnalité du portrait). Cependant, nous retrouvons le pouvoir cadratif des adverbiaux temporels en initiale de sections et de paragraphes et dans des TSC temporelles. Nous voyons dans l'exemple X.14 trois circonstants temporels qui structurent le paragraphe entre trois périodes : une première où Léonard de Vinci habite à Florence, une seconde où il loge à la fois à Milan et à Florence et une dernière où il quitte ces deux villes. Cette structuration est amorcée par le titre de section d'une part et la présence d'un adverbial temporel en initiale de paragraphe d'autre part. Nous voyons dans cet exemple

187 Nous ne remettons pas ici le long exemple de structure énumérative temporelle avec amorce et conclusion commenté dans la partie II.3.2.

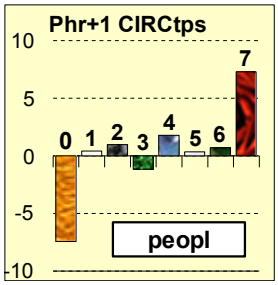
que la référence temporelle exprimée par les adverbiaux temporels (et plus particulièrement par le premier adverbial temporel) ne porte pas forcément sur tout le segment délimité de part et d'autre par un CIRCtps. Les seules références communes aux phrases constitutives des différents segments sont la référence nominale (qui est commune à quasiment toutes les phrases du texte) et la référence spatiale. Cet exemple montre bien la distinction à faire entre portée sémantique et pouvoir cadratif (voir III.3.3.b).

(X.14) ***Florence-Milan, 1500 - 1513*** [titre niveau 2]
En 1500, *Léonard se rend à Mantoue, où il dessine le portrait d'Isabelle d'Este, qui tentera en vain d'obtenir d'autres oeuvres, à Venise, [...], et à Florence, où - [...]- il va rester jusqu'en 1506. Son activité se partage entre des travaux de peinture : [...], et des travaux d'ingénieur militaire dans le val d'Arno et à Piombino. Léonard remet en chantier le Trattato commencé entre 1487 et 1492, et y travaille jusque vers 1513.* ***À partir de 1506,*** *il partage son temps entre Milan où [...], et Florence, où [...]. Il revient au projet de statue équestre, cette fois pour le condottiere Trivulce, donne de petits panneaux (perdus) de Madones pour Louis XII, une seconde version de La Vierge aux rochers, le tableau de la Sainte Anne. Il déploie une grande activité scientifique : anatomie, mathématique, et fournit des projets d'architecture, de décors pour Charles d'Amboise.* ***Mais, en 1513,*** *il quitte définitivement Milan reconquis par la coalition antifrançaise. [PEOPL_11]*

Cet extrait illustre bien comment le temps n'entrave pas la continuité topicale du passage. Il peut y avoir un adverbial temporel, cela n'implique pas nécessaire la redénomination du référent, ce que les résultats du graphique X.14 avaient déjà indiqué : il y a significativement plus de formes au DegAccess_7 après un adverbial temporel en position P2 (les données sont reprises dans le graphique X.17). Cependant, il semble que la corrélation CIRCtps/déplacement soit plus forte dans PEOPL que dans GEOPO. On remarque une cohabitation significative en position P2 entre les adverbiaux temporels et les noms propres répétés correspondant au DegAccess_4. Cette cohabitation peut tout à fait relever d'un effet de déplacement, surtout si ce type de cohabitation donne lieu à des Phr+1 avec DegAccess_7, comme le suggère le graphique X.18.



Graphique X.17: écarts des cohabitations [CIRCtps+DegAccess_n] par rapport au modèle générale pour chaque PosTxt dans PEOPL



Graphique X.18: Écarts des DegAccess dans les Phr+1 de CIRCtps dans PEOPL par rapport au modèle PEOPL (P2)

Mais nos données ne valident pas cette hypothèse. En effet, il semble bien que c'est lorsqu'il y a une cohabitation [CIRCtps+DegAccess_7] que l'on a préférentiellement une Phr+1 au DegAccess_7 comme dans l'exemple X.15 ou même dans l'exemple X.14 et non après une cohabitation [CIRCtps+DegAccess_4], comme l'illustre l'exemple X.16.

(X.15) *On sait que la vocation de Pascal pour la géométrie s'éveilla quand il eut douze ans, à la lecture des Éléments d'Euclide. Mais c'est seulement en 1639 qu'il commença à s'y intéresser de façon sérieuse. Il eut alors connaissance du court ouvrage du géomètre architecte Girard Desargues, Brouillon project d'une atteinte aux événements des rencontres du cône avec un plan, qui venait de paraître. Cette œuvre capitale jetait les bases de la géométrie projective et d'une théorie unitaire des coniques. Le jeune Blaise Pascal fut alors seul à en comprendre toute la richesse. Il en adopta aussitôt les idées fondamentales : introduction des éléments à l'infini, définition des coniques comme sections planes quelconques de cônes à base circulaire, étude de ces courbes comme perspectives du cercle, relation d'involution déterminée sur une droite quelconque par une conique et les côtés d'un quadrilatère inscrit. Mais bientôt Pascal prolonge les idées de Desargues par un apport original.* ***Dès 1639,*** *il démontre son célèbre théorème : les points d'intersection des couples de côtés d'un hexagone inscrit dans une conique sont en ligne droite (cf. CONIQUES, chap. 1). Il rédige alors l'Essay pour les coniques. Il utilise l'œuvre d'Apollonius, qui, dans son ouvrage sur les coniques, avait*

notamment défini les propriétés des diamètres et des tangentes, mais il énonce ces résultats "d'une manière plus universelle qu'à l'ordinaire". [PEOPL_4]

(X.16) ***2.1) Épiphanie*** [titre niveau 2]

Il [Joyce] associait l'épiphanie à une manifestation banale de langage ou autre. [...]. Épiphanie ou non, il avait perçu bien vite que la vie n'est pas faite pour être vécue, mais pour être réinventée par l'écriture. Peu importent dès lors les valeurs vitales : l'insuffisance, l'absence, le manque sont les biens de l'esprit créateur, la base d'un nouveau jeu de rapports ; avoir été jeté par le sort dans un pays tel que l'Irlande, avoir pour père un John Joyce, c'est un destin à cultiver. Il ne faut pas se méprendre, lorsqu'on lit au terme de Dedalus : "Ô vie, je vais pour la millionième fois à la rencontre de la réalité de l'expérience." Ce n'est pas Rastignac toisant Paris. Une seconde phrase éclaire la première : "Je veux façonner dans la forge de mon âme la conscience incréée de ma race." En fait, l'homme dont l'œuvre sera essentiellement parodique commencera par avoir de la vie même - et de " l'expérience " - une vision parodique : c'est une sorte de jeu. Son père, déjà, avait joué à être étudiant en médecine : il allait en faire autant, et améliorer le modèle, en choisissant Paris plutôt que Cork. ***À Paris, en 1902,*** *Joyce ne prit même pas ses inscriptions. Ce simulacre fut un moment décisif par sa nullité même. Joyce passait d'un refus à l'autre, et tous signifiaient la volonté de ne pas se laisser intégrer au social. Ses armes : " le silence, l'exil et la ruse ". L'exil était sa condition. Il ne fit que le matérialiser lorsqu'en octobre 1904, muni de la compagne qu'il s'était donnée ce 16 juin, jour d'Ulysse, il quitta l'Irlande. [PEOPL_2]*

À la lecture de ces deux exemples, il semble se dessiner des configurations assez inattendues : un adverbial temporel intraparagraphique et non intégré dans une TSC locale précédant un thème topical de haut degré d'accessibilité montre une portée alors qu'un même adverbial précédant une redénomination non. Nous ne mettons pas plusieurs exemples par manque de place, mais nous n'avons pas trouvé d'exemples inverses dans nos données. Il serait bien entendu nécessaire d'effectuer de plus amples recherches sur des données plus importantes pour confirmer ces tendances.

°, °, °, °

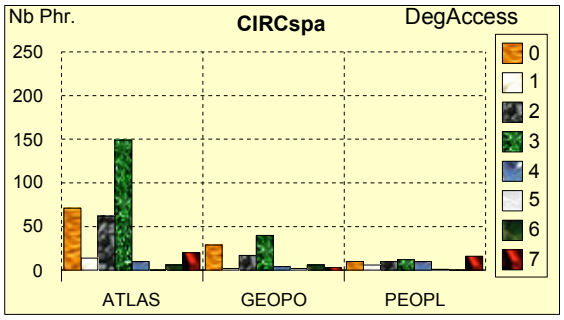
Les adverbiaux temporels sont généralement considérés comme de bons indices dans le signalement des discontinuités dans le discours. Or, nos résultats d'analyse ne vont pas nettement dans ce sens. En effet, si les adverbiaux temporels ont un fort pouvoir cadratif, c'est principalement parce qu'ils se situent préférentiellement en début de paragraphe ou de section (voir partie IX.3.2). Tout comme un circonstant détaché en initiale de phrase oriente l'interprétation de sa phrase d'accueil, un circonstant détaché en initiale de paragraphes oriente l'interprétation de son paragraphe d'accueil. Il semble alors que ce ne sont pas les adverbiaux temporels qui ont un pouvoir cadratif, mais leur position. Dans des contextes de TSC temporelles, le problème est différent. En effet, par un effet de 'solidarité', les adverbiaux temporels acquièrent un pouvoir structurant. Lorsqu'un adverbial temporel apparaît seul au milieu d'un paragraphe, ce pouvoir s'estompe en faveur d'un fonctionnement similaire à celui des appositions. Les adverbiaux temporels isolés correspondraient alors à des appositions locatives (cf. V.4.2). Cette dernière remarque est à nuancer dans le cas de PEOPL où la continuité topicale marquée par les formes de DegAccess_7 (*i.e.* les pronoms et les SN possessifs) permet aux adverbiaux d'étendre une portée sémantique au delà de leur phrase d'accueil. Pris dans une continuité par défaut engendrée par la continuité topicale autour d'un même personnage, la référence temporelle poursuit sa référence aux phrases suivantes. Par contre, si l'adverbial est précédé d'une redénomination, cet effet semble affaibli et alors le fonctionnement discursif des circonstants temporels se rapproche des appositions descriptives.

Enfin, dernière découverte de cette section : le rôle des SN démonstratifs représentés par les DegAccess_5 et 6 dans la délimitation des portées de cadre temporel. Cette découverte s'appuie sur les analyses effectuées dans GEOPO qui constitue le seul sous-corpus a montrer significativement plus de SNdem en Phr+1 de CIRCtps. Elle consiste à poser l'hypothèse que les démonstratifs ferment un cadre de discours, puisqu'ils co-réfèrent à une entité en

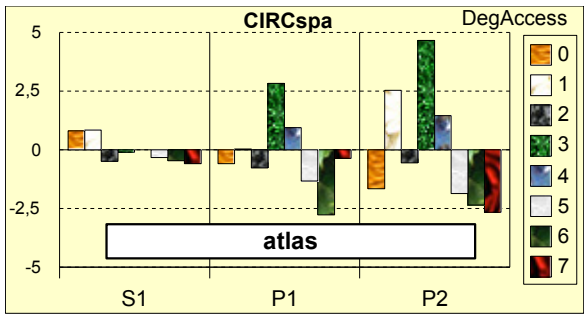
la désignant en dehors des circonstances qui l'ont introduite dans le discours. Cette hypothèse a déjà fait l'objet d'une autre expérience relatée dans Hồ-Đắc (2005). Elle ouvre une piste tout à fait intéressante pour les travaux sur l'encadrement du discours, même si, il ne faut pas l'oublier, les démonstratifs constituent des données relativement rares.

X.2.3. Les localisations spatiales dans ATLAS : un indice de déplacement?

Les adverbiaux spatiaux sont une spécificité d'ATLAS qui représente à lui seul plus de 65% de tous les adverbiaux spatiaux du corpus (334/500). Le graphique X.19 montre bien cette disproportion. Les données dans GEOPO et PEOPL sont d'ailleurs trop faibles pour donner des résultats pertinents pour notre analyse[188]. C'est pourquoi, dans la suite de cette section, nous ne représentons que les données pour ATLAS.



Graphique X.19: Répartitions des cohabitations [CIRCspa+DegAccess_n] dans les trois sous-corpus



Graphique X.20: Écarts des cohabitations [CIRCspa+DegAccess_n] par rapport au modèle générale pour chaque PosTxt dans ATLAS

Le calcul des écarts réduits entre la répartition des DegAccess après un adverbial spatial et celle dans tout le sous-corpus ATLAS montre clairement une cohabitation **[CIRCspa+DegAccess_3(SNdef_R/court)]**, en initiale de paragraphe comme à l'intérieur d'un paragraphe. Nous avons vu que ATLAS emploie fréquemment des reprises lexicales pour réaliser des continuités référentielles intraparagraphiques, notamment des progressions thématiques dérivées construites autour de descriptions longues et dont la tête lexicale reprend celle de l'hyperthème (VIII.2.2, IX.2.3 et IX.2.4). Selon ce constat et les écarts indiqués dans le graphique X.20, on peut supposer que le DegAccess_3 correspond principalement à des SNdef longs avec reprise. Nous en apprenons davantage en observant les environnements des SNdef courts et des SNdef avec reprise. Les SNdef courts ne montrent pas de variations significatives selon le type d'INIT qui les précède, alors que les SNdef avec reprise oui et particulièrement après un adverbial spatial en position P2 (voir X.4.2). Ces résultats signifient que lorsque l'on a un Sndef avec reprise en P2, il est significativement plus souvent précédé d'un adverbial spatial. En recoupant ces résultats avec ceux indiqués dans le graphique X.20, il semble bien que les DegAccess_3 dans ATLAS correspondent davantage à des reprises lexicales qu'à des réductions de termes.

Il apparaît donc que, dans ATLAS, les adverbiaux spatiaux indiquent un déplacement qui empêche l'installation d'une continuité topicale construite par des anaphores pronominales, comme l'illustre l'exemple X.17 où quasiment chaque Thème topical suivant un adverbial spatial présente le nom « enfant » en tête de syntagme.

(X.17) *Les structures sociales de la population d'âge scolaire se rangent en cinq types principaux.* ***Au nord du pays, de la Normandie à l'Alsace, région parisienne exclue,*** *les enfants d'ouvriers forment la moitié de la population scolarisable, devançant de beaucoup ceux des cadres moyens, puis des employés.* ***De la Bretagne au Pays Basque et de la Bourgogne à Marseille*** *les situations sont proches de la moyenne française: la proportion des enfants d'ouvriers*

188 La cohabitation la plus fréquente dans GEOPO ([CIRCspa+DegAccess_3]) ne montre aucune variation significative.

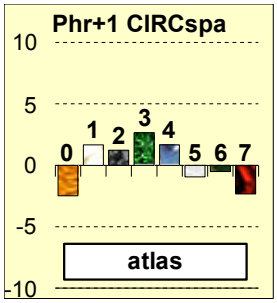
avoisine encore 40% mais les enfants d'agriculteurs et d'artisans et commerçants sont plus nombreux. La part des enfants d'agriculteurs s'accroît dans les départements les plus ruraux de l'Ouest et du Sud-Ouest, les enfants d'ouvriers demeurant cependant de loin les plus nombreux. ***En Provence et dans l'Hérault,*** *les enfants d'ouvriers ne forment plus que le quart d'une population scolarisable dont un enfant sur trois a un père cadre moyen ou supérieur.* ***En Île-de-France ainsi qu'à Toulouse,*** *les enfants des catégories aisées sont aussi nombreux, au moins, que ceux des familles ouvrières.* ***Et à Paris,*** *il y a deux fois plus d'enfants de cadres supérieurs et moyens que d'enfants d'ouvriers.* [ATLAS_2]

Nous retrouvons ici des situations de TSC locales qui confèrent aux adverbiaux, temporels ou spatiaux, un pouvoir cadratif. Comme nous l'avons remarqué dans la partie précédente au niveau des TSC temporelles, la TSC spatiale est annoncée par une amorce qui avertit le lecteur d'une organisation de l'information en *cinq types principaux*. Lorsque l'on a un adverbial isolé *(i.e.* non intégré dans une TSC spatiale), il montre la même absence de portée que celle observée pour les adverbiaux temporels dans GEOPO, comme l'illustre l'exemple X.18 où l'adverbial « *Au nord de l'île* » ne sert qu'à illustrer les faits concernant le « grignotage de la mer » par les habitants de Jersey.

(X.18) *La décision de la Cour internationale de Justice de La Haye en 1953 octroyant définitivement les plateaux des Minquiers et des Ecrehous à la Couronne qui les rétrocéda aussitôt au bailliage de Jersey, est le facteur déclenchant des ambitions insulaires visant à étendre leur contrôle sur une portion de plus en plus importante de la mer commune. Depuis, forts des bastions avancés acquis en 1953, les Jersiais ont entrepris un grignotage de la mer commune en s'appuyant sur le moindre caillou ou banc de sable découvrant aux marées les plus fortes pour étendre la zone maritime dont ils assurent la surveillance (zones roses autour des Minquiers, au Nord et à l'Ouest de Jersey).* ***Au nord de l'île,*** *les Paternosters ou les Dirouilles constituent des points d'appui modestes mais, selon les insulaires, opérant pour projeter leur souveraineté sur une part toujours plus importante de la mer commune définie en 1839. On voit ce que la décision de 1953 portait en germe: Jersey exerce ses contrôles, voire restreindrait ou interdirait l'accès aux pêcheurs normands à des zones pouvant se situer à dix-huit kilomètres de Jersey et à six kilomètres de continent dans le secteur de Carteret. En effet, les propositions insulaires de 1992 visaient [...]* [ATLAS_1]

°, °, °, °

Tout comme le temps, il semble bien que les adverbiaux spatiaux acquièrent un pouvoir cadratif lorsqu'ils sont dans des configurations de changement de paragraphe ou de TSC locale. Cette absence de pouvoir cadratif ne signifie pas pour autant qu'ils ne constituent pas des indices de déplacement. En effet, on voit bien dans l'exemple X.18 que la phrase introduite par « *Au nord de l'île*, » est en discontinuité par rapport au discours précédent. Il n'y a pas de progression thématique ici, ni d'organisation spatiale, mais une articulation rhétorique qui met cette phrase en dehors des continuations idéationnelles du passage. Cet effet est d'ailleurs accentué par l'absence de reprise lexicale dans la description définie en collocation du circonstant spatial : « *les Patermosters ou les Dirouilles* ».

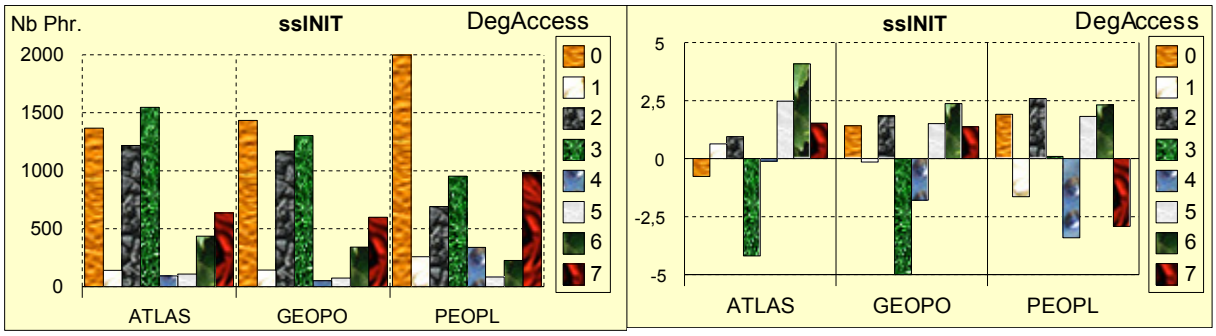


Graphique X.21: Écarts des DegAccess dans les Phr+1 de CIRCspa dans ATLAS par rapport au modèle ATLAS (P2)

À l'inverse des adverbiaux temporels, il semble difficile qu'un adverbial spatial puisse orienter toute l'interprétation d'une section, leur rôle étant certainement plus local que celui des adverbiaux temporels et donc leur portée (quand il y a portée) plus réduite. Nous avions en effet remarqué un écart significatif des adverbiaux cpatiaux en P1 mais pas en S1 (IX.3.2). Ce fonctionnement plus local se retrouve dans l'écart négatif observé pour les DegAccess_7 dans les Phr+1 de CIRCspa (graphique X.21). Cet écart, à la limite du seuil de significativité *(z*=-2,3), ce qui signifie qu'il est très rare d'avoir une continuité topicale après un adverbial spatial (dans GEOPO, les Phr+1 de CIRCtps ne montraient pas d'écart au niveau des DegAccess_7).

X.3. L'absence d'INIT : un indice valable ?

Alors que nous ne considérions pas l'absence d'INIT comme un indice dans le marquage de la séquentialité du discours, il est apparu au fil des chapitres VIII et IX que cette absence pouvait fonctionner comme un indice de continuité. En effet, dans tous les sous-corpus, il y a plus de ssINIT (absence d'INIT) en P2. Il semble effectivement logique que l'absence d'un élément détaché en initiale ne permet pas d'indiquer un déplacement ou une rupture dans la séquentialité du discours. Mais indique-t-elle pour autant une continuité?



Graphiques X.22: Répartition des DegAccess dans les phrases sans INIT et écarts par rapport à la totalité des phrases dans chaque sous-corpus

Les phrases sans INIT sont majoritaires. Elles représentent 70% de toutes les phrases du corpus. Comme le montre le graphique des écarts en X.22, l'absence d'INIT ne montre que deux cohabitations significatives : la cohabitation **[ssINIT+DegAccess_6(SNdem_R/courts)]** dans ATLAS et **[ssINIT+DegAccess_2(SNdef longs et sans reprise)]** dans PEOPL. Nous revenons en fin de partie sur la cohabitation observée dans PEOPL. Concernant la cohabitation observée dans ATLAS, et qui se retrouve également dans les deux autres sous-corpus, mais à la limite de notre seuil de signifiance, nous constatons que la fréquence des SN démonstratifs est pour une fois assez importante pour que cet écart soit pertinent. Les SNdem apparaissent très majoritairement dans les phrases sans INIT. Sur les 1 538 SNdem recensés, 82,5% ne sont pas précédés d'un INIT. Les SNdem sont principalement des SN courts (70%) et/ou des SN avec reprise lexicale (42%), voir partie VIII.2.2. Le DegAccess_6, qui confond description courte et reprise lexicale, représente 80% de tous les SNdem (1 229 occurrences). Les SNdem ont fait l'objet de nombreuses études par rapport à leur fonction de 'reclassifieur'. Nous avons vu qu'ils pouvaient constituer de bons indices de fermeture d'un cadre temporel (X.2.2). La reclassification va nécessairement de pair avec l'absence de l'expression d'une nouvelle circonstance, puisque le SNdem réfère à une entité en dehors de tout cadre, comme dans l'exemple X.19.

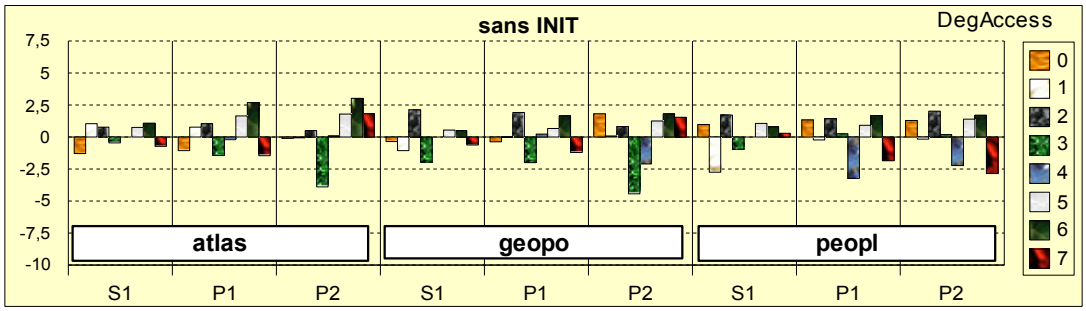
(X.19) *Afin d'assurer l'amélioration continue de la qualité et de la sécurité des soins, tous les établissements de santé, publics et privés, doivent faire l'objet d'une procédure externe d'évaluation, dénommée accréditation.* ***Cette procédure****, conduite par l'Agence Nationale d'Accréditation et d'Evaluation en Santé (ANAES), vise à porter une appréciation indépendante sur la qualité d'un établissement à l'aide d'indicateurs, de critères, et de référentiels portant sur les procédures, les bonnes pratiques cliniques et les résultats des différents services et activités de l'établissement. [ATLAS_1]*

Il est finalement assez logique de voir ici la cohabitation [ssINIT+DegAccess_6]. Comme le montre le graphique X.23, cette cohabitation est significative, dans ATLAS uniquement, en position P1 et en position P2. Dans Hồ-Đắc (2005), nous avons montré que les SNdem pouvaient se situer significativement en dernière phrase de paragraphe. Dans cette position textuelle (la dernière phrase de paragraphe), les SNdem jouent plutôt le rôle d'une anaphore encapsulante et constituent alors le sujet d'une phrase exprimant une sorte conclusion des phrases précédentes,

comme dans l'exemple X.20. Les anaphores encapsulantes sont généralement réalisées par des descriptions courtes construites autour de noms sous-spécifiés (e.g. s*ituations, cas, problèmes, points, idées,* etc.)

(X.20) *La dimension personnelle a pour conséquence, plus que pour toute autre élection, de privilégier la continuité. Ceci valait jusqu'à la fin des années 1980. Un conseiller général, une fois installé, avait plus tendance à se représenter qu'à se retirer et la prime au sortant et à la stabilité s'est révélée importante depuis trente années. Il n'était pas rare de voir se représenter dans les années 1980, des élus qui avaient obtenu leur premier mandat au début des années 1950. La longévité politique des conseillers généraux se mesure à l'aune de celle des sénateurs. Nous sommes là dans le même registre de représentation politique, qui n'est plus exactement le même que celui de la représentation à l'Assemblée Nationale. La conséquence logique de ceci est l'âge moyen élevé des assemblées départementales dans les années 1980. Les conseils généraux de l'Ouest ont une moyenne d'âge supérieure à la moyenne française.* ***Ces traits caractéristiques*** *vont être à l'origine de modifications dans les élections cantonales de la fin des années 1980. [ATLAS_3]*

Malgré leur relative rareté (8% des Thèmes topicaux), les SNdem semblent constituer de bons indices dans le signalement des déplacements. Cette corrélation SNdem/déplacement est renforcée par le fait que la cohabitation [ssINIT+DegAccess_6] apparaît également en initiale de paragraphe.



Graphique X.23 : Écarts des cohabitations [ssINIT+DegAccess_n] par rapport au modèle général pour chaque PosTxt dans chaque sous-corpus

En dehors de la cohabitation [ssINIT+DegAccess_6], on peut remarquer que les marques de co-référence lexicale sont généralement dissociées des phrases sans INIT. Ainsi, dans ATLAS et GEOPO, il y a significativement moins de SNdef au DegAccess_3 dans les phrases sans INIT sans avoir significativement plus de ProPoss. Nous verrons dans la partie X.4 que les SNdef avec reprise et les noms propres répétés cohabitent significativement plus avec des INIT. Cela conforte l'idée de la nécessité d'une reprise lexicale après un déplacement.

Dans PEOPL, nous observons des écarts opposés au modèle général. En effet, alors que les associations DegAccess_1/S1, DegAccess_4/P1 et DegAccess_7/P2 caractérisent PEOPL (IX.2.4), lorsque l'on se trouve dans une phrase sans INIT, les écarts observés s'inversent. Ces résultats semblent dire que la co-référence se fait généralement accompagnée d'un INIT, ce qui est bien surprenant surtout dans le cas du DegAccess_7. Nous ne pouvons malheureusement pas dire davantage de cette observation qui nous laisse quelque peu démunie.

°, °, °, °

L'analyse de l'environnement des phrases sans INIT n'a pas fourni de résultats pertinents nous permettant d'affirmer que l'absence d'INIT constitue un indice de continuité. Cependant, elle nous a permis de remarquer la cohabitation [ssINIT+DegAccess_6] qui nous informe davantage sur le fonctionnement discursif des SN démonstratifs que sur celui lié à l'absence d'INIT. Il semble bien que les SNdem constituent de bons indices de déplacement qui permettent la fermeture de cadres de discours (discontinuité idéationnelle) ou l'encapsulation de plusieurs phrases en une phrase conclusive (discontinuité rhétorique). L'idée d'un tel fonctionnement mériterait d'être plus approfondie,

notamment pour comprendre les phénomènes de fermeture des cadres de discours et des TSC ou autres structures énumératives.

X.4. Des indices de continuité référentielle

Nous avons analysé différents indices de continuité référentielle dans les chapitres VIII et IX : les pronoms et SN possessifs correspondant au plus haut degré d'accessibilité (DegAccess_7), les cas de reprise lexicale dans les SNdef (DegAccess_3) et dans les NP (DegAccess_4). Les trois sous-sections suivantes analysent les variations observées au niveau des éléments détachés lorsque ce type d'élément apparaît en Thème topical.

Nous avons observé dans les chapitres précédents un autre indice de continuité référentielle : les SNdef courts qui apparaissaient significativement plus dans les phrases intraparagraphiques qu'ailleurs (IX.2.3.b), ce qui laissait présager leur rôle dans le signalement des continuités, notamment dans les textes où l'emploi des ProPoss est peu fréquent comme dans les textes d'ATLAS et de GEOPO. À l'inverse, les SN longs se situaient davantage en initiale de sections et de paragraphes. Cette différence de fonction liée à la taille des SN est vraisemblablement plus importante pour les SN démonstratifs et les SN définis, c'est pourquoi nos analyses ne portent que sur les SN de ces catégories. Cependant, si l'on observe les variations occasionnées par la présence d'un SNdef ou SNdem long – SNdef/dem long – sur la nature des INIT en collocation, aucun écart significatif n'apparaît, ce qui signifie que la présence d'un SN long ne change pas les répartitions observées d'un point de vue général.

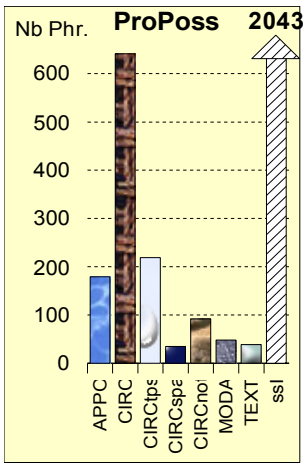
Les SNdef/dem courts semblent, eux, avoir quelques influences sur leur environnement, mais celles-ci restent très faibles et spécifiques aux phrases intraparagraphiques dans GEOPO : les SNdef/dem courts apparaissent significativement plus après un INIT de nature appositive ou circonstancielle. Nous avons déjà remarqué ces cohabitations lors de l'étude de l'influence des appositions et des adverbiaux temporels sur leur environnement. Nous ne reviendrons pas sur ce point, aucune information nouvelle ne venant compléter le tableau.

Pour plus de légèreté, nous ne présenterons pas les données pour les catégories indéfinies : les éléments détaché à la fonction indéfinie par notre programme (Iautres) et les circonstants aux rôles sémantiques autres que le temps, le lieu et les notions ou au rôle indéfini par notre programme (CIRCautres). Ces deux types d'INIT ne correspondent pas à une fonction ou un rôle sémantique bien défini, ce qui rend l'interprétation de leurs variations ou de leurs cohabitations peu informatives.

X.4.1. Le plus haut degré d'accessibilité : les Pronoms et les Possessifs

X.4.1.a) Un indice de continuité topicale

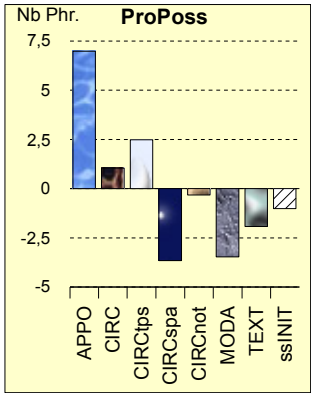
Les pronoms et les SN possessifs sont des indices forts de continuité topicale. Nous avons vu précédemment que la présence d'un adverbial circonstanciel n'entrave pas cette continuité dans PEOPL. Plus encore, la présence d'un adverbial temporel favorise les formes de la catégorie ProPoss. Ce phénomène est proche de celui remarqué au niveau des appositions qui, elles aussi, cohabitent significativement plus avec un ProPoss, mais cela dans tous les sous-corpus (voir notre remarque finale de la partie X.2.2).



Graphique X.24 : Répartition générale des cohabitations [INIT+ProPoss]

Si l'on observe les répartitions des cohabitations [INIT+ProPoss] (graphique X.24), nous retrouvons ces deux types d'INIT. Le calcul des écarts réduits montre que les cohabitations **[APPO+ProPoss]** et **[CIRCtps+ProPoss]** sont significatives (graphique X.25) et surtout la cohabitation [APPO+ProPoss]. Nous avons déjà commenté ces cohabitations dans les parties X.1 et X.2.2. Les phrases sans INIT sont également très présentes, ce qui ne signifie pas pour autant une cohabitation significative entre ProPoss et l'absence d'INIT, ce qu'indique le graphique X.25 et ce que nous venons de voir dans la partie X.3.

Les différentes cohabitations montrent des écarts significatifs plus importants dans ATLAS et PEOPL que dans GEOPO (graphique X.26, page suivante). Les écarts significatifs observés dans GEOPO sont tous négatifs et recoupent avec l'absence significative de cohabitation [CIRCtps+ProPoss] notée dans la partie X.2.2. L'écart au niveau des circonstants spatiaux (CIRCspa) indique qu'il y a une même absence significative au niveau de la cohabitation [CIRCtps+ProPoss]. Il semble que les localisations spatiales et temporelles dans GEOPO ne permettent pas la continuité topicale (ce qui ne signifie pas que ces localisations portent, voir X.2.2).



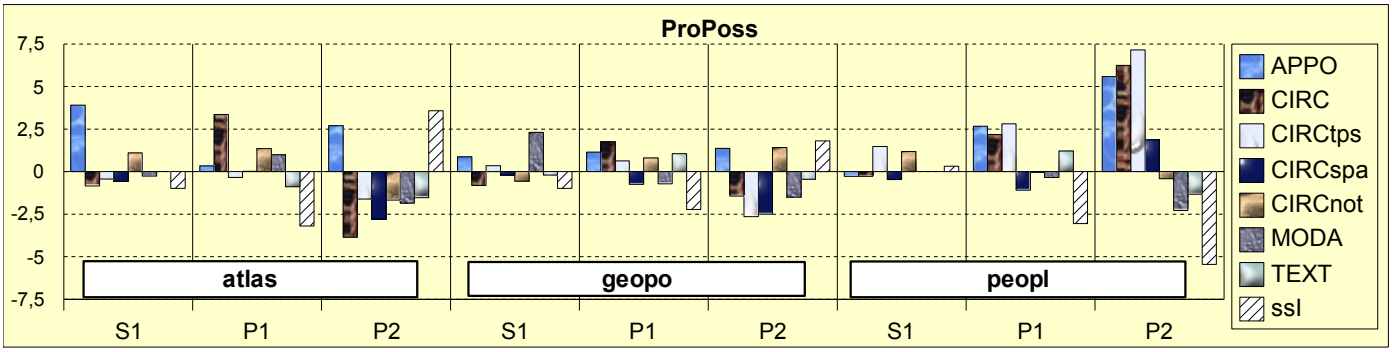
Graphique X.25 : Écarts des INIT dans les phrases avec ProPoss par rapport à toutes les phrases

Dans ATLAS, la cohabitation [APPO+ProPoss] réalise une progression de la position S1 à la position P2 : les ProPoss en S1 ne semblent possibles qu'après une APPO (nous en avons déjà parlé en X.1); en P1, les ProPoss cohabitent significativement avec un CIRCautre ou un CIRCnot ; en P2, nous avons préférentiellement des phrases sans INIT ou, dans une moindre mesure, des appositions. Cette répartition des rôles entre APPO et CIRC semble indiquer que l'apposition n'est généralement pas utilisée au niveau des initiales de paragraphe. Les exemples (X.21) et montrent des débuts de paragraphes dans ATLAS où cohabitent un CIRCnot/autre et un ProPoss. Le pronom de 3e personne en fonction sujet y instruit généralement une reprise anaphorique du référent introduit en INIT.

(X.21) *Si les relations de Caen avec Cherbourg, au niveau du flux de véhicules légers, sont relativement élevées il n'en est pas de même au niveau du flux des poids lourds, car en calculant la proportion du nombre de poids lourds par rapport au nombre de véhicules légers, nous remarquons, qu'en règle générale, celle-ci est entre 2 à 3,6 fois moins élevée. Or pour Cherbourg, le nombre de poids lourds est 7 fois moins important que celui des véhicules légers, ce qui dénote un certain manque de relation entre les deux agglomérations au niveau du transport de marchandises et donc au niveau commercial.*
***En ce qui concerne les relations de Cherbourg avec Le Havre et Rouen, elles** sont faibles, en effet elles s'élèvent respectivement à 30 PL/j, soit 450 tonnes, et 65 PL/j, soit 975 tonnes. [ATLAS_1]*

(X.22) *Plus des trois quarts des élèves entrés en 6e vont jusqu'en 3e, avec éventuellement une ou deux années de redoublement; la proportion n'était que des deux tiers vers 1975. Le rapport entre le nombre d'élèves en 6e et celui des 3e donne une idée de l'érosion des effectifs en cours de collège; au milieu des années 1980, elle laissait en route plus de 200 000 élèves de chaque classe d'âge. Cette érosion des effectifs tend à s'atténuer: en 1985-86, on comptait trois élèves de 3e pour quatre en 6e; en 1991-92, la proportion approche neuf sur dix: les redoublements sont de moins en moins nombreux, les sorties en cours de collège de moins en moins fréquentes.*
***Même si l'indicateur de l'érosion des effectifs scolaires au cours du cycle d'enseignement en collège est grossier, il** met en évidence avec force que les scolarités en collège vont plus souvent à leur terme dans le Midi, en Bretagne et en Île-de-France; et qu'au contraire, dans la plupart des départements situés au nord d'une ligne Bordeaux-Genève la scolarité en collège tourne plus souvent court. [ATLAS_2]*

Dans PEOPL, les cohabitations entre un ProPoss et un adverbial circonstanciel, notamment temporel sont significatives. C'est cette signifiance qui nous a permis d'affirmer dans la partie X.2.2 qu'un changement de référence temporelle n'entrave pas les continuités topicales dans PEOPL. Les résultats renforcent cette affirmation puisque les ProPoss cohabitent significativement plus avec des adverbiaux temporels qu'avec n'importe quel autre type d'INIT. Cela va jusqu'à entraîner un écart réduit négatif de l'absence d'INIT devant un ProPoss (alors que, intuitivement, nous aurions associé la présence d'un ProPoss avec l'absence d'INIT). Cette dissociation entre ProPoss et l'absence d'INIT avait déjà été remarquée en X.3.

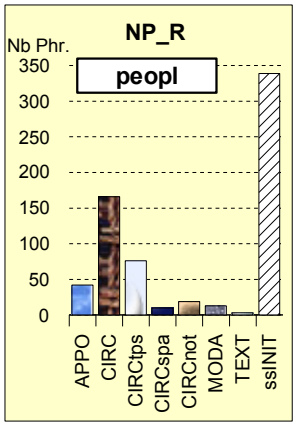


Graphique X.26 : Écarts des cohabitations [INIT+ProPoss] par rapport au modèle général pour chaque PosTxt dans chaque sous-corpus

On retrouve également une cohabitation, moins forte, entre les adverbiaux temporels et les noms propres répétés. On est alors en droit de se demander si les noms propres répétés sont une simple alternative aux pronoms de 3e personne ou s'ils ont un autre fonctionnement discursif.

X.4.1.b) Les noms propres répétés – NP_R – dans PEOPL : une alternative aux pronoms?

Les noms propres répétés sont particulièrement associés au sous-corpus PEOPL. 71% d'entre eux proviennent de PEOPL. Et il y en a certainement plus, si l'on mesurait les redénominations en première phrase section (voir la remarque de la partie IX.2.3.b). Suite à l'article de Schnedecker (2005) et à nos observations dans les chapitres VIII et IX, il semble que, dans des textes de type portrait, les noms propres répétés collaborent avec les PRO3 dans le marquage des continuités topicales. Notre question est de savoir si la présence d'un nom propre répété est totalement aléatoire dans ce type de collaboration ou si les redénominations y sont utilisées dans des environnements particuliers.



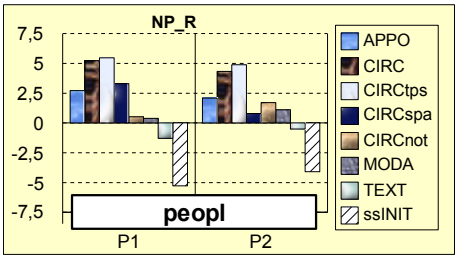
Graphique X.27 : Répartition des cohabitations [INIT+NP_R] dans PEOPL

Comme le montrent les graphiques X.27 et X.28, même si la cohabitation la plus fréquente associe les noms propres répétés aux phrases sans INIT (graphique X.27), il y a significativement moins de noms propres répétés dans les phrases sans INIT et cela, que ce soit en première phrase de paragraphe ou à l'intérieur d'un paragraphe (graphique X.28).

Ce sont les appositions et encore plus les circonstants temporels qui cohabitent significativement avec une redénomination et cela, dans les positions P1 et P2. Ces cohabitations avaient déjà été remarquées dans les parties X.1.1 et X.2.2, où elles s'affichaient à côté des cohabitations [APPO+ProPoss] et [CIRCtps+ProPoss]. La comparaison des graphiques X.26 et X.28 nous apprend que les écarts occasionnés par la présence d'un nom propre répété sont très proches de ceux occasionnés par un pronom ou un SN possessif. Cela suggère une influence identique de ces formes sur leur environnement. Toutefois, contrairement aux écarts

occasionnés par les ProPoss, les schémas observés en P1 et P2 pour les noms propres répétés sont très semblables, ce qui signifie un fonctionnement identique dans ces deux positions textuelles.

Cette similarité d'écarts en initiale et à l'intérieur d'un paragraphe laisse à penser que les noms propres répétés ne sont pas placées par hasard dans le texte mais en accompagnement d'un déplacement, ce qui est à l'origine la définition même du fonctionnement des redénominations (V.4.3.b). Nous avions déjà notée cette corrélation dans la partie IX.2.3.a, corrélation illustrée par l'exemple X.23 dans lequel on voit bien que la présence d'une redénomination n'est jamais le fruit d'une redondance.



Graphique X.28 : Écarts des cohabitations [INIT+NP_R] par rapport au modèle général pour chaque PosTxt dans PEOPL

(X.23) ***Léonard**, quittant l'atelier de Verrocchio, a pu être quelque temps au service de Laurent de Médicis; c'est ce qu'affirme l'Anonyme Gaddiano, [...]*
*En 1482-1483, **Léonard** est au service de Ludovic le More qui vient de s'emparer du duché de Milan (1480). **Il** devient le grand animateur de la cour. Après 1499, **il** cherche un autre protecteur princier : [...]. **Il** intéresse César Borgia pour la guerre en Romagne (1502), Charles d'Amboise pour l'architecture (projet de villa au bord du le Noviglio) et la décoration des demeures (à partir de 1506). Aux yeux des princes français comme à ceux de César Borgia, **Léonard**, si célèbre qu'il soit comme peintre, compte pour ses autres capacités. On est frappé aussi par la facilité avec laquelle l'artiste-ingénieur passe du service d'un protecteur à celui de son adversaire. **Il** revient à Milan avec les princes français qui ont chassé Ludovic; à la fin de 1504, il est à Piombino, auprès de Jacoppo IV d'Appiano, qui, l'année précédente, avait été chassé par César Borgia, le patron de Léonard. Les grands esprits n'ont pas de camp. **Léonard** appartient à qui se l'attache et lui laisse un loisir pour l'étude. Paul Jove a été frappé de ses capacités comme organisateur de fêtes, musicien, etc., et conclut que ces aptitudes "l'ont rendu cher à tous les princes qui l'ont connu." En dehors des décors de théâtre ou de parade, **Léonard** a composé des rébus, constitué des recueils de devinettes et de fables, rédigé des devises, des [...]* [PEOPL_11]

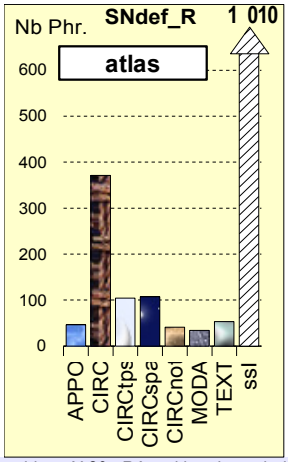
°, °, °, °

Ce regard croisé sur le fonctionnement, dans PEOPL, des pronoms et des noms propres répétés confirme l'idée d'un fonctionnement différent entre ces deux expressions co-référentielles. La présence d'un pronom de 3e personne est l'indicateur même d'une continuité topicale dominante, *i.e.* qui organise la globalité d'un segment en regroupant autour du même topique les propositions contenues dans ce segment. Ainsi, comme nous l'avons vu au niveau des adverbiaux temporels, il peut y avoir une TSC temporelle ou spatiale dans un segment organisé essentiellement par des pronoms, mais la TSC temporelle ou spatiale sera gouvernée par la continuité topicale et la portée des circonstants sera totalement dépendante de la progression topicale. Par contre, lorsque l'on a un déplacement fort, qui touche à la fois un *setting* et une étape dans l'articulation rhétorique de cette continuité topicale, il semble que les noms propres répétés soient préférentiellement utilisés. C'est pourquoi nous avons significativement plus de cohabitations [CIRCtps+NP_R] en initiale de paragraphe.

X.4.2. SNdef avec reprise lexicale dans ATLAS : indice de déplacement

Nous avons observé dans le chapitre précédent que, dans ATLAS, les SNdef_R apparaissaient significativement plus en initiale de paragraphes (voir IX.2.3.b), ce qui nous permettait de croire en leur rôle dans le signalement des déplacements, et plus précisément dans le signalement d'une continuité référentielle et d'un changement de *setting* ou une articulation rhétorique. Les SNdef sans reprise ne montraient pas de variation significative selon la position textuelle. Notre analyse de l'influence des SNdef sans reprise sur leur environnement ne montre pas non plus de cohabitations significatives (nous n'exposons pas les résultats).

Les résultats exposés dans la partie X.2.3 ont indiqué que dans ATLAS les adverbiaux spatiaux cohabitent significativement plus avec les SNdef_R. Notre conclusion de cette partie évoquait l'idée suivante : un adverbial spatial isolé et suivi d'un SNdef sans reprise lexicale est plus susceptible d'indiquer un déplacement au niveau des continuations rhétoriques et semble se situer en dehors des continuations idéationnelles. Nous appuyons cette suggestion sur l'exemple suivant.



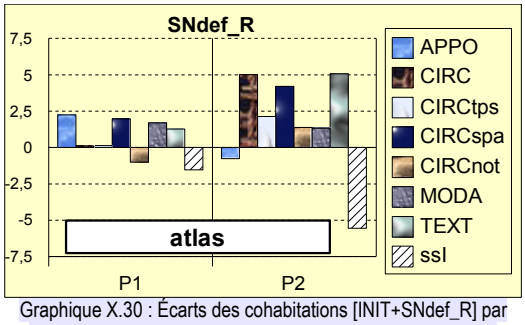
Graphique X.29 : Répartition des cohabitations [INIT+SNdef_R] dans ATLAS

(X.24) *La décision de la Cour internationale de Justice de La Haye en 1953 octroyant définitivement les plateaux des Minquiers et des Ecrehous à la Couronne qui les rétrocéda aussitôt au bailliage de Jersey, est le facteur déclenchant des ambitions insulaires visant à étendre leur contrôle sur une portion de plus en plus importante de la mer commune. Depuis, forts des bastions avancés acquis en 1953, les Jersiais ont entrepris un grignotage de la mer commune en s'appuyant sur le moindre caillou ou banc de sable découvrant aux marées les plus fortes pour étendre la zone maritime dont ils assurent la surveillance (zones roses autour des Minquiers, au Nord et à l'Ouest de Jersey).* ***Au nord de l'île,*** *les Paternosters ou les Dirouilles constituent des points d'appui modestes mais, selon les insulaires, opérant pour projeter leur souveraineté sur une part toujours plus importante de la mer commune définie en 1839. On voit ce que la décision de 1953 portait en germe: Jersey exerce ses contrôles, voire restreindrait ou interdirait l'accès aux pêcheurs normands à des zones pouvant se situer à dix-huit kilomètres de Jersey et à six kilomètres de continent dans le secteur de Carteret. En effet, les propositions insulaires de 1992 visaient [...] [ATLAS_1]*

Cet effet de parenthèse dans les continuations idéationelles ne se retrouve pas lorsqu'il y a une cohabitation [CIRCspa+PRO3] comme dans l'exemple X.25 où l'on a une même phrase illustrative (la fonction illustrative est ici accentuée par la présence de l'expression « *par exemple* ») où le pronom force le maintien de la continuité topicale.

(X.25) *Dans les départements les plus ouvriers, de l'école élémentaire à la 4e de collège, la proportion d'enfants d'ouvriers passe souvent de la moitié au tiers environ.* ***Dans le Pas-de-Calais par exemple,*** *elle tombe de 53 à 37%; par contre celle des enfants de cadres ou d'employés augmente de moitié. Le phénomène est général, en sorte que les différences régionales des origines sociales de la population scolarisable, bien qu'atténuées, subsistent: les 4e des collèges de la moitié nord de la France, à l'exception de ceux de l'agglomération parisienne et de la Bretagne, comptent plus [...] [ATLAS_1]*

Pour en revenir aux SNdef avec reprise, nous voyons dans ATLAS qu'ils cohabitent significativement plus avec des circonstants spatiaux ou temporels et des adverbiaux textuels, ce qui renforce l'idée d'une corrélation SNdef_R/déplacement. Ces cohabitations ne sont significatives qu'en P2, ce qui recoupe avec les remarques faites ci-dessus : les cohabitations entre un INIT (à l'exception des appositions) et un SNdef_R constitueraient, dans ATLAS, des indices d'une discontinuité soit à un niveau idéationnel (un changement de *setting*) dans le cas d'une organisation spatiale du paragraphe ou de la section; soit à un niveau plus rhétorique dans le cas d'un adverbial isolé. Ce comportement explique également pourquoi les SNdef_R apparaissent significativement moins dans les phrases sans INIT.



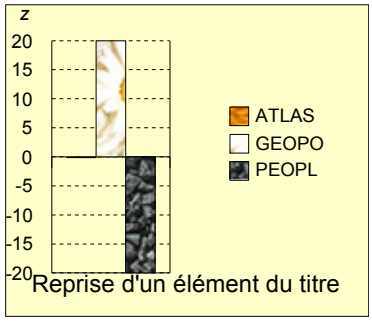
Graphique X.30 : Écarts des cohabitations [INIT+SNdef_R] par rapport au modèle général pour chaque PosTxt dans ATLAS

X.5. Co-référence en initiale de section : reprise des éléments du titre

Au fur et à mesure de notre analyse, la première phrase de section nous a plusieurs fois intriguée par son haut degré d'accessibilité. Ce haut degré d'accessibilité est vraisemblablement lié au titre de section, mais ce n'est pas le cas dans tous les sous-corpus. Le dernier indice pris en compte pour comprendre l'organisation des différents sous-corpus correspond à la reprise d'un élément du titre de section.

Près de 30% des phrases du corpus (6 309 précisément) comportent la reprise d'au moins un élément du titre de section en cours d'activation. Les éléments du titre pris en compte sont les formes étiquetées « NOM » ou « VINF » (verbe à l'infinitif) par Syntex ainsi que les SN tels que délimités par Syntex. Par exemple, les éléments (dans leur forme lemmatisée) retenus du titre de section en cours d'activation à ce moment même sont : [co-référence] [initiale] [section] [initiale de section] [co-référence en initiale de section] [reprise] [élément] [titre] [éléments du titre] [reprise des éléments du titre]. Le programme cherche ensuite si les phrases de la section contiennent un ou plusieurs de ces éléments. Parallèlement à ce calcul, le nombre de reprises repérées pour chaque titre de section est enregistré.

La proportion de phrases présentant une reprise de titre est très variable selon le type de texte (les écarts réduits observés dépassent les +20 et -20). Nous observons un schéma des écarts de dispersion dans lequel seul ATLAS correspond à la moyenne, *i.e.* il présente effectivement 30% (en réalité 27%) de phrases reprenant un élément du titre de section en cours. Aux deux pôles de ces variations, GEOPO affichent beaucoup plus de reprises que les deux autres sous-corpus, avec 38% des phrases; et PEOPL beaucoup moins, avec 16% de phrases reprenant un élément du titre. La faiblesse de reprise dans PEOPL peut en partie s'expliquer par la nature même des titres. En effet, les titres de PEOPL sont beaucoup plus 'simples' que dans ATLAS ou GEOPO. Ces deux derniers affichent des titres plus complexes et notamment plus fournis en formes nominales.



Graphique X.31 : variations selon les sous-corpus des reprises lexicales d'éléments du titre de section

La mise entre parenthèses temporaire des intérêts minoritaires dans la définition de la politique étrangère [GEOPO_27]
Face à l'urgence : les premières décisions de l'administration Bush [GEOPO_1]
Des indications qui donnent la mesure des difficultés scolaires [ATLAS_2]
Prévisions démographiques dans la zone transmanche [ATLAS_1]
La maturité (1779-1788) [PEOPL_16]
Face à la barbarie [PEOPL_15]
L'épopée [PEOPL_27]

Si l'on mesure la taille des titres, cette différence saute aux yeux : PEOPL présente des titres de 27 caractères en moyenne alors que les titres font entre 43 et 44 caractères en moyenne dans GEOPO et ATLAS. Mais cette grande faiblesse de reprise dans PEOPL est essentiellement due au fonctionnement des titres dans ce sous-corpus.

Comme nous l'avions remarqué dans la partie X.2.1, les premières phrases de sections dans PEOPL montrent très fréquemment une redénomination du personnage principal, comme l'illustre l'exemple suivant.

(X.26) ***4.2) L'"œuvre en cours"****[titre niveau 2]*
En 1920, ***Joyce*** *arrive à Paris, venant de Zurich où Dada avait inauguré, en 1916, une négation culturelle globale. Au même moment, le cubisme s'installe, Gertrude Stein épaulant Picasso. L'agression surréaliste s'affirme. [...]*
4.3) Le rabâchage de la comédie humaine [titre niveau 2]
Joyce *écrivit les deux premières pages de Finnegans Wake le 10 mars 1923. Elles devinrent les pages 380 et 382 de l'œuvre terminale. [...]*

4.4) Du particulier à l'universel [titre niveau 2]
Comme dans Ulysse, ***Joyce*** *voulut donner à son entreprise une structure propre à la tirer du particulier vers l'universel : c'est ainsi que Finnegans Wake évolue au gré d'une double structure de pensée constituée d'abord par l'affrontement dialectique de couples opposés, où l'on reconnaît une idée clé du philosophe Giordano Bruno, retrouvée par Blake :"Sans l'opposition des contraires, pas de progrès."[...]* *[PEOPL_2]*

Comme on le voit, les titres dans PEOPL indiquent véritablement des étapes dans le portrait qui est fait. Ces titres sont essentiellement thématiques : ils orientent la section en précisant la caractéristique du personnage ou la période de sa vie qui fait l'objet de la section. Ils n'indiquent pas une rupture dans la séquentialité du texte, mais un véritable déplacement. C'est pourquoi, il n'est pas rare de trouver un PRO3 en S1 qui ne reprend pas un élément du titre mais le personnage principal.

Il est très rare que le titre fasse l'objet d'une reprise pronominale. Hồ-Đắc *et al.* 2004 dénombre 6 titres sur 1 041 donnant lieu à une reprise pronominale. Parmi ces 6 titres, 5 apparaissent dans un corpus constitué de documents de travail. Notre corpus nous fournit 12 titres (sur 1 300) faisant l'objet d'une reprise pronominale. 11 se trouvent dans ATLAS, dont les textes peuvent être apparentés à des documents de travail (voir VII.1.1). Ces reprises se retrouvent dans des structures titrées de bas niveau composées de sections de très petite taille, comme le montre l'exemple X.27.

(X.27) **DES STRUCTURES D'ÂGE CONTRASTÉES** [titre de niveau 1]
Classification en deux classes d'âges [titre de niveau 2]
[...]
Classification en six classes d'âges
Ce type de classification permet d'affiner la première lecture des structures démographiques des zones d'emploi françaises et britanniques réalisée précédemment; six classes sont individualisées et correspondent à des profils bien distincts. Pour faciliter la lecture, les profils et leur localisation géographique seront analysés un par un.
Le type 1 (titre de niveau 3)
Il correspond à une situation démographique ou dominent deux classes d'âges : [...]
Le type 2
La seconde classe correspond à une variante de la précédente : [...]
Le type 3
Dominant en France, il ne se retrouve pas en Angleterre. Il correspond aux zones rurales [...]
Le type 4
Il individualise une à une les principales zones d'emploi [...]
Le type 5
Il correspond aux zones d'emploi affichant un vieillissement de leurs structures [...]
Le type 6
Variante de la précédente, il correspond à un vieillissement plus important [...] *[ATLAS_1]*

Les exemples de ce type ne se trouvent que dans ATLAS et uniquement dans le texte 1 : l'Atlas Transmanche. Nous y trouvons une amorce (dans le titre de plus haut niveau et dans la plupart des énumérations. La pronominalisation est souvent précédée d'une apposition, comme dans les sous-sections titrées « le type 3 » et « le type 6 ». L'exemple X.28 montre une structure similaire : des sous-sous-sections très courtes, des titres référentiels, des reprises par pronominalisations en initiale de section, précédées une fois sur deux par une apposition. Ces titres créent un effet de rupture, les sections n'ayant pas vraiment d'autre lien entre elles que le titre de niveau supérieur.

(X.28) ***DIVERSITÉ DES COMPAGNIES DE FERRIES*** *[titre de niveau 1]*
Brittany-Ferries *[titre de niveau 2]*
Compagnie bretonne née en 1972 pour pallier le manque de moyens de transport des marchandises agricoles entre le Finistère-Nord et le Royaume-Uni, la S.A."Bretagne-Angleterre-Irlande"(B.A.I.), prend en 1974 son nom actuel de Britanny-Ferries, avec la diversification de ses activités vers le trafic passager. Basée au départ à Roscoff, elle a ouvert des lignes sur Saint-Malo, Cherbourg et Caen. Grâce à une diversification de ses activités (transport roulier et passager, produits vacances. ..), la Britanny-Ferries couvre tous les ports en Manche-Ouest, sauf le Havre, avec des liaisons à destination de la France, de La Grande-Bretagne, de l' Irlande et même de l' Espagne
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. Réduisant progressivement sa présence sur le trafic transmanche, elle a concentré ses prestations sur la ligne Calais-Douvres, son principal but étant d' offrir un service de ferries à la française. Elle propose aussi diverses prestations de voyagiste.
***Les compagnies étrangères** [titre de niveau 2]*
***P & O-European Ferries** [titre de niveau 3]*
Fondée en 1837, la Peninsular and Oriental Steam Navigation Company est une des plus anciennes compagnies de navigation britannique. Grand groupe multinational aux activités très diversifiées (transport,immobilier,services. ..), on la trouve par l'intermédiaire de filiales dans tous les domaines de l' armement maritime, des croisières aux ferries.
Elle est aujourd'hui la première compagnie assurant des liaisons transmanche, principalement sur la ligne Calais-Douvres. Elle tente de s'imposer sur le marché de la Manche-Ouest, avec des départs de ferries de Portsmouth vers Le Havre et Cherbourg. Elle développe aussi des services vers l' Irlande. Elle est en train de se redéployer, par des accords avec d' autres compagnies comme la Stena Ligne (effectif au 10 Mars 1998) ou Irish Ferries. Elle dispose ainsi d' un quasi monopole sur certaines liaisons transmanche.
***Stena-Line** [titre de niveau 3]*
Arrivée en Manche en 1994, à l'occasion de la privatisation de British Ferries et de la création de la SeaFrance-Sealink, elle se sépare de cette dernière pour devenir en 1996 la Stena-Line. Appartenant au groupe suédois Stena, elle est aujourd'hui une des plus grosses compagnies de ferries du monde. Elle joue la carte de la modernité et de la rapidité, en mettant à la disposition de sa clientèle transmanche, des catamarans et des navires de moyenne capacité. Malgré des restructurations sur le marché transmanche - suppression de la ligne Cherbourg-Southampton, accords commerciaux avec la P & O-European Ferries - elle reste néanmoins une compagnie maritime parmi les plus puissantes.
***Sally-Ferries** [titre de niveau 3]*
La Sally-Ferries est une filiale britannique du groupe finno-suèdois Effjonh. Positionnée uniquement sur la ligne Dunkerque - Ramsgate, sa situation sur le marché transmanche est quelque peu précaire. Elle a du supprimer des emplois et des rotations en 1996, et les difficultés semblent persister en 1997.
***Hoverspeed** [titre de niveau 3]*
Appartenant au groupe britannique Sea Containers, la compagnie Hoverspeed est présente sur les lignes Calais-Douvres et Boulogne-Folkstone. Elle a concentré sa cible uniquement sur le trafic passager rapide, disposant pour cela de catamarans et d' aéroglisseurs. Pour concurrencer le Tunnel sous la Manche, elle offre désormais le service de navettes au départ de Paris comme de Londres. C' est une des rares compagnies transmanche à voir son chiffre d' affaire en progression. [ATLAS_1]

Dans PEOPL, c'est tout à fait autre chose. Si l'on observe les données en fonction de la position textuelle, nous voyons que c'est PEOPL qui montre le plus de pronoms de 3e personne en initiale de sections. Ces pronoms ne sont pas des reprises du titre de section actif, mais des reprises du titre de l'article, en l'occurrence le personnage dont on fait le portrait. Ces reprises soulignent la très forte structuration de ces textes autour d'une continuité topicale (le topique étant le personnage du portrait). Dans l'exemple X.29 issu du portrait de Pascal, nous remarquons également un indice de cohésion temporelle (*alors*) qui, en plus de la continuité topicale autour de Pascal, ajoute un effet de continuité temporelle par rapport à la section précédente.

(X.29) ***1. AUX CONFINS DE LA CONNAISSANCE ET DE LA FOI***
1.5) Le projet d'"Apologie"
Il se consacre alors complètement à son projet d'Apologie de la religion chrétienne... [PEOPL_4]

La force du titre de l'article est telle que nous trouvons un texte (le portrait de Diderot) commençant directement par un pronom personnel : « *Il naît à Langres le 5 octobre 1713. De sa mère, il ne parlera que par occasions. En revanche, son père, un petit industriel coutelier, garde sur lui une influence décisive...* ».

Il semble bien que dans PEOPL, les titres n'entravent absolument pas la continuité topicale (cela rappelle notre observation du fonctionnement des adverbiaux temporels dans PEOPL). Ce constat est éclatant dans l'exemple X.30, cas rarissime d'une continuité topicale – autre que celle entourant le personnage du portrait – passant par dessus un changement de section.

(X.30) *Ce n'est pas que l'amour romanesque soit mis en accusation dans ces comédies de la belle période, de Peines d'amour perdues à La Nuit des rois (Twelfth Night), car l'ironie qui tempère ses extravagances n'est ni mordante ni sarcastique, elle baigne dans le climat général de jeunesse et de gaieté qui en fait l'arme idéale pour désamorcer les*

prétentions du sentiment. La plus significative de ces héroïnes, c'est sans doute Rosalinde qui a assez d'esprit pour échapper aux envoûtements de l'amour, assez d'amour pour ne pas céder à la tentation du dénigrement. Chez elle, l'équilibre est parfait : le sourire d'Éros se pare des reflets de l'ironie. Un rien d'ironie de trop, et la comédie pourrait se prendre à grincer.

5.3) Grincements

Elle grincera, effectivement, pour des raisons que le biographe ne pourra jamais élucider, lorsque l'inquiétude et le désarroi envahiront l'univers idéalisé où jusqu'ici le dramaturge les avait relégués au niveau inférieur de la farce ou du burlesque. Déjà, sur la fin du siècle, Le Marchand de Venise, tout poétisé qu'il soit par le triomphe de l'amour et le clair de lune de la nuit lyrique à Belmont, avait des résonances peu rassurantes. La mélancolie... [PEOPL_8]

Dans PEOPL, les titres de section ne s'associent pas vraiment à un effet de rupture dans la séquentialité du texte, mais plus à un effet de déplacement et certainement plus lorsque le niveau du titre est bas. Cet effet de déplacement ne se retrouve pas vraiment dans ATLAS et GEOPO, même si on peut poser l'hypothèse que plus le niveau du titre est bas, moins le titre marque une rupture dans la séquentialité du discours.

Les ruptures réalisées par des titres dans GEOPO et PEOPL ne sont pour autant pas identiques. Le graphique X.31 montrait significativement plus de reprises d'élément du titre dans GEOPO, alors que ATLAS restait dans la moyenne générale. La différence entre ATLAS et GEOPO est d'un tout autre ordre que celle entre PEOPL et GEOPO. Elle ne semble pas mettre en cause une différence au niveau de la complexité des titres ou au niveau de la dominance hégémonique d'une continuité topicale. Les deux sous-corpus présentent des titres relativement longs et complexes comme le montrent les exemples précédents. Par contre, ATLAS présente beaucoup plus de titres de sections – ATLAS comporte 532 titres de sections, contre 377 dans GEOPO (et 391 dans PEOPL) – et notamment beaucoup plus de titres de niveau inférieur comme le montre le tableau X.2.

Corpus	Niveau du titre	Nb titre	Longueur du titre (Nb de caractères)		Longueur de la section (Nb de phrases)	
			Moyenne	Mediane	Moyenne	Mediane
ATLAS	1	91	*50*	45	*132*	45
	2	189	*45*	40	*34*	17
	3	241	*42*	37	*19*	11
	4	11	*32*	18	*7*	6
	Total	532	*44*		*43*	
GEOPO	1	272	*40*	35	*37*	23
	2	49	*54*	51	*34*	28
	3	51	*52*	49	*16*	15
	4	5	*50*	29	*14*	9
	Total	377	*43*		*34*	
PEOPL	1	131	*25*	22	*61*	51
	2	239	*28*	28	*21*	20
	3	21	*16*	15	*9*	7
	Total	391	*27*		*33*	

Tableau X.2: Longueur des titres de section et des sections selon le sous-corpus et le niveau du titre

Le fait que GEOPO présente plus de titres de niveau 1 que de niveau 2 ou 3 signifie que la plupart des textes qui le composent ne comportent pas de titraille complexe et hiérarchisée. Il n'y a généralement que des titres de même niveau. À l'inverse, ATLAS et PEOPL montre une titraille plus complexe, ce qu'illustre l'augmentation du nombre de titre plus on descend de niveau. En effet, si toutes les sections de niveau 1 sont divisées en sections de niveau 2, il y a au minimum deux fois plus de titres de niveau 2 que de titres de niveau 1 – ce qu'on observe dans PEOPL et ATLAS. PEOPL s'arrête à une titraille de complexité 2 (*i.e.* allant jusqu'au niveau 2) alors qu'ATLAS va jusqu'au niveau 3 (le niveau 4 est très rare). La proportion moyenne de reprise lexicale mesurée dans ATLAS peut être mise en relation avec

le découpage en section des textes d'ATLAS. Nous avons décrit les textes composants ATLAS comme des textes aux sections courtes alors que GEOPO ou PEOPL présentent des textes aux sections plus longues (voir partie VII.1.1.a). Cette différence de découpage en sections peut certainement expliquer les écarts observés.

X.6. Récapitulatif des configurations d'indices découvertes

Ce chapitre a permis d'observer en détails la cohabitation de tous les indices émergeant des variations textuelles et positionnelles. De nombreuses configurations d'indices ont été découvertes et leur définition, plus ou moins complexe, implique généralement des facteurs de nature diverse : le type de texte, la position textuelle, l'environnement discursif, la complexité de la position initiale, les relations section/titre.

Le tableau X.3 récapitule toutes les configurations pertinentes signalant la séquentialité du discours. Les dernières colonnes indiquent le type de signalement associé à ces configurations. Nous y distinguons le niveau idéationnel des autres niveaux d'organisation : lorsqu'une configuration a été analysée comme signalant tel ou tel type de (dis)continuité au niveau idétaionnel, nous mettons une croix dans la colonne correspondante (C = continuité, D = déplacement et R = rupture)[189]. La dernière colonne indique les instructions concernant les articulations rhétoriques, la portée des CIRC et l'implication des titres de section dans l'organisation discursive.

Configuration d'indices				idéationnel			Autre instruction
indice	type de texte	position textuelle	environnement	C	D	R	
APPO	PEOPL	S1	+{ProPoss,_R}		X		titre thématique
	ATLAS	S1				X	titre référentiel
	-	P1	+{ProPoss, _R, SNcourt}	X			articulation rhétorique
	-	P2		X			
CIRCtps	-	S1	-		X	X?	portée
	-	P1	-		X		portée
	ATLAS GEOPO	P2	dans une TSC		X		portée
	PEOPL	P2	+ProPoss dans une TSC	X			Portée + articulation rhétorique
	PEOPL	P2	dans une TSC		X		portée
CIRCspa	ATLAS	P1	-		X		portée
	-	P2	dans une TSC		X		portée
	-	P2	+ProPoss	X			
PRO3	-	S1 ou P2	-	X			
	-	P1	+CIRC en INIT		X		articulation rhétorique
	tous	P1	+APPO en INIT	X			articulation rhétorique
NP_R	PEOPL	-	-		X		articulation rhétorique
SNdef_R	ATLAS	-	-				

Tableau X.3 : Récapitulatif des configurations d'indices signalant une (dis)continuité dans la séquentialité du discours

189 L'absence de croix ne signifie pas que la configuration d'indices ne joue pas au niveau de la (dis)continuité concernée, mais que nous n'avons aucune information sur ce type de (dis)continuité relative à cette configuration.

Ce tableau est à mettre en regard de celui établi à la fin du chapitre V (partie V.6) dans lequel nous indiquions nos hypothèses quant au rôle des différents indices dans le signalement de la séquentialité. Le tableau X.4 reprend ces différents indices en y ajoutant les informations obtenues par notre exploration en corpus.

	dénomination	***indice de...***	***Résultats des tests***
mise en forme	changement de section	discontinuité (rupture)	rupture ou déplacement (selon l'organisation topicale du texte)
	changement de paragraphes	discontinuité globale (déplacement)	oui
	titre	discontinuité (rupture)	rupture ou déplacement (selon l'organisation topicale du texte)
	puce	discontinuité locale (déplacement)	non testé
connecteurs		continuité	certainement, mais uniquement au niveau idéationnel
INIT *élément détaché en initiale*	adverbial circonstanciel	discontinuité	non (voir tableau X.3)
	adverbial modalisateur	discontinuité ?	pas de résultats probants
	adverbial textuel	discontinuité	pas de résultats probants
	apposition	continuité	oui
	argument inversé	continuité	pas de résultats probants
ThTop *Thème topical*	pronominalisation	continuité	oui
	reprise lexicale	continuité ou déplacement	oui
	détermination indéfinie	rupture	non
ThSpe *Construction Spéciale*	construction clivée	discontinuité	non
	construction présentationnelle	discontinuité	pas de résultats probants
	Construction à sujet inversé	continuité	pas de résultats probants
	Dislocation à gauche	continuité	pas de résultats probants
	Construction avec On/Nous...	discontinuité ?	pas de résultats probants
	Construction impersonnelle	discontinuité ?	pas de résultats probants

Tableau X.4: Regard croisé de nos hypothèses quant au signalement de l'organisation discursive en position initiale et nos résultats

Chapitre XI
Conclusion

Nous avons présenté dans cette thèse le résultat de notre exploration en corpus dont le but est de découvrir des configurations d'indices pertinentes pour la caractérisation de l'organisation discursive de différents types de textes. Cette thèse s'appuie sur l'hypothèse forte que la position initiale donne accès tant au contenu d'un texte qu'à son mode organisationnel. L'accès au contenu relève de la notion de Thème au sens sémantico-pragmatique d'« à propos », l'accès au mode organisationnel relève du Thème au sens cognitif : « le Thème constitue le point de départ qui oriente la suite du discours ». Ainsi, ce qui se situe en position initiale est à même d'exprimer les acteurs et les circonstances des procès et de structurer ces procès.

Pour comprendre la réalité de ce rôle de la position initiale dans l'organisation du discours, nous avons mis en place et expérimenté une méthodologie nouvelle pour l'étude de l'organisation discursive. Nous avons tout d'abord effectué une grande collecte de données. Ces données ont été délimitées et caractérisées de façon exhaustive, ce qui était nécessaire afin de considérer tous les éléments susceptibles de participer au signalement de la séquentialité du discours. L'analyse des données est basée sur une mesure statistique très simple : le test de l'écart réduit. Ce test est apparu comme étant le mieux adapté à notre capacité de statisticienne débutante. Sa facilité d'utilisation nous a permis d'explorer une multitude de pistes en manipulant à volonté nos données et nos facteurs de variation.

L'exhaustivité de l'annotation nous a permis de réaliser une approche guidée par les données. Ainsi, les éléments analysés correspondent aussi bien à des éléments par rapport auxquels nous avions des attentes quant à leur implication dans le signalement de l'organisation discursive (comme les adverbiaux circonstanciels ou les redénominations) qu'à des éléments dans lesquels nous mettions peu d'espoirs (les appositions en sont un bel exemple). Cette méthodologie opérationnelle nous permet ainsi de porter un regard nouveau sur l'organisation discursive. Elle nous permet ainsi de valider ou de remettre en cause certaines hypothèses mais également de découvrir de nouveaux fonctionnements discursifs.

XI.1. Validation d'hypothèses ...

Nos analyses nous ont permis de valider quatre hypothèses à l'origine de notre étude : une certaine corrélation entre la structuration de la représentation mentale et la structuration du discours; l'importance des facteurs de position textuelle et de la variation textuelle dans l'étude de l'organisation du discours; une définition large du marquage

discursif par configurations d'indices plutôt que par des formes dédiées; et le sens instructionnel porté par certaines expressions référentielles.

Les modèles cognitivistes proposés par Werth (1999) ou Gernsbacher (1990) semblent tout à fait adaptés pour représenter les organisations discursives observées. Le fait de localiser les adverbiaux temporels et spatiaux en initiale d'unités textuelles globales va de pair avec la notion de fondations (les « *world-builders* » de Werth), ainsi qu'avec la notion de première mention de Gernsbacher. Il semble effectivement que les éléments qui commencent l'unité textuelle orientent textuellement et idéationnellement toute son interprétation. Ces résultats indiquent une certaine corrélation entre la **structuration du « *text-world* »** (en sous-mondes ou sous-structures selon les auteurs) et la **structuration du texte**. Cette corrélation est particulièrement remarquable au niveau de la fonction discursive des titres de section et du découpage en section. Les titres montrent un rôle dans l'organisation discursive d'autant plus fort que le texte présente une structure titrée complexe et des sections à texte propre (i.*e*. telles que visualisées par le lecteur) courtes.

Nos analyses ont montré l'importance de prendre en compte les **facteurs de la position textuelle et de la variation textuelle**. La position textuelle et le type de texte constituent des indices à part entière dans les configurations discursives observées. Les éléments positionnés en première phrase de section ou de paragraphe ont une fonction dans l'organisation discursive qui dépasse leur fonctionnement propre. Le cas des adverbiaux circonstanciels est exemplaire. En effet, alors que nous avions de grandes attentes concernant les circonstants détachés en initiale et leur capacité à indiquer un déplacement, il semble que leur pouvoir structurant et leur portée soient tout à fait conditionnés par leur situation textuelle (voir section suivante). Concernant l'indice de variation textuelle, nos résultats ont bien montré que la structure des textes ainsi que leur organisation idéationnelle globale (en particulier leur caractère mono/pluri-référentiel) change considérablement le sens instructionnel associé aux éléments lexicaux-syntaxiques mais aussi aux différentes positions textuelles.

Le **marquage de l'organisation discursive** apparaît donc comme le fruit d'une interaction d'indices de différentes natures. Peu d'expressions montrent un comportement stable c'est-à-dire similaire dans nos trois sous-corpus et dans les trois positions textuelles distinguées. Par exemple, au niveau des connecteurs 'purs', nous avons obtenu des résultats qui mériteraient une étude plus fine. L'association entre certains types de texte et certaines formes de connecteur permettrait certainement de comprendre mieux le caractère polysémique des connecteurs. Si l'on ajoute à cela la prise en compte de la position textuelle, l'idée de polysémie s'agrandit : il est très probable qu'un *mais* en initiale de paragraphes n'exprime pas la même relation de discours qu'un *mais* inter-propositionnel ou du moins, qu'il exprime plus qu'une simple relation de discours.

Une troisième validation nous semble importante est celle relative au **sens instructionnel associé aux différentes expressions référentielles sujets** (les Thèmes topicaux). D'après nos résultats, il semble que le pronom *il* marque effectivement une continuité topicale, et cela indépendamment de sa position textuelle. Cependant, l'apparition de ce pronom en initiale de sections ou de paragraphes est plus difficile qu'en position intraparagraphique (à nuancer selon le type de texte). Ainsi, d'autres éléments peuvent faciliter l'emploi du pronom *il* en première phrase de section ou de paragraphe et notamment les appositions, nous y revenons plus loin lors de la présentation des découvertes issues de nos analyses. Dernière validation concernant les instructions relatives aux expressions co-référentielles : le cas des reprises lexicales. Une reprise lexicale semble effectivement employée lorsqu'il y a un déplacement dans la continuité référentielle (*i.e.* un changement de *setting* ou une articulation rhétorique). Cette corrélation doit néanmoins être considérée de façon plus nuancée dans des textes au caractère mono-référentiel (de type portraits) où la redénomination peut marquer des déplacements moins importants que dans des textes au caractère pluri-référentiel. Nous avons obtenus moins de résultats probants pour les descriptions courtes. Il semble que

l'emploi des descriptions courtes ne soit pas spécialement associé à un signalement de la séquentialité du discours, mais plutôt au tissage de la texture du texte en permettant de nombreuses continuités référentielles, comme nous l'avons observé dans GEOPO.

XI.2. ... et remises en cause

Nos résultats nous ont poussée à remettre en cause certaines hypothèses portant notamment sur la portée des adverbiaux circonstanciels, sur le fonctionnement discursif des constructions spéciales ou encore sur le signalement des discontinuités par les modalisateurs ou les adverbiaux textuels. Ils nous ont également amenée à nous questionner sur notre utilisation de certaines théories comme la théorie de l'accessibilité d'Ariel (1990).

La plus importante remise en cause concerne l'**hypothèse de l'encadrement du discours**. Alors que nous avions tendance à croire en une corrélation forte entre la présence d'un adverbial en initiale de phrase et sa capacité à étendre une portée sémantique au delà de sa phrase d'accueil, il semble que cette corrélation soit fortement contrainte. La simple présence en initiale d'un adverbial ne constitue pas un indice valide d'introduction d'un nouveau cadre de discours. Le phénomène de portée associé aux adverbiaux détachés en initiale apparaît être la conséquence de trois phénomènes discursifs : le pouvoir structurant alloué aux index d'une structure énumérative, la continuité topicale et le changement de paragraphe ou de section.

Selon nos résultats, il semble que la portée sémantique des adverbiaux soit une conséquence d'un certain pouvoir structurant alloué aux adverbiaux dans les situations de TSC. Le terme « pouvoir structurant » nous paraît ici préférable à celui de « portée cadrative ». En effet, le pouvoir structurant que présentent certains adverbiaux antéposés et qui leur permet de délimiter les différents segments-item d'une structure énumérative nous semble bien différent d'un quelconque effet de portée (il serait étrange de penser le rôle des puces dans une structure énumérative en terme de portée !) Dans ces structures énumératives, l'important n'est pas la persistance du critère d'interprétation, mais l'effet organisationnel de certains parallélismes comme l'utilisation successive d'adverbiaux circonstanciels de même rôle sémantique. Ensuite, la portée sémantique peut être fortement associée à une continuité topicale dominante comme dans PEOPL. Comme nous l'avons vu dans le chapitre précédent, la référence temporelle exprimée par un adverbial isolé semble être 'emmenée' par la continuité topicale plus qu'elle ne semble porter d'elle même. Ainsi, les circonstances dans lesquelles un référent topique est exprimé peuvent persister tant que ce référent reste le topique et ne change pas de circonstances ou n'est pas reclassifié. Notons que dans ces situations, les circonstances n'ont aucune raison d'apparaître en position initiale. Enfin, une référence circonstancielle exprimée en première phrase de paragraphe ou de section (et pas nécessairement en position antéposée) donne l'impression de 'profiter' du découpage textuel pour étendre sa portée.

Ainsi, les adverbiaux spatiaux et temporels n'ont de portée cadrative qu'en situation d'initiale de sections, d'initiale de paragraphes ou de participation à une continuité texto-stratégique particulière de type structure énumérative. Dans les autres situations discursives*, i.e.* en situation d'isolement, un adverbial spatial ou temporel ne porte pas réellement, sauf si une continuité topicale 'emmène' sa référence avec elle. Le cas des adverbiaux circonstanciels montre bien à quel point nous avons affaire à des interactions entre indices et non à des corrélations 'absolues' entre un marqueur (lexical) et une fonction.

Les différentes remarques sur l'emploi des expressions nous amène également à certaines remises en cause, notamment sur la question de l'**utilisation du degré d'accessibilité pour représenter la séquentialité du discours**. Nous avons à plusieurs fois remarqué la non adéquation entre le degré d'accessibilité d'une expression et sa capacité

à indiquer une continuité référentielle. Ainsi, les reprises lexicales dans ATLAS construisent l'identité référentielle de nombreuses portions de texte organisées autour de progressions thématiques dérivées. Dans GEOPO, ce sont les descriptions courtes sans reprise lexicale qui tissent dans de nombreux cas les liens référentiels entre segments. Enfin, dans PEOPL, l'emploi massif des noms propres répétés nous pousse à revoir notre échelle d'accessibilité. Cependant, tout ce que nous disons là n'a que très peu à voir avec la notion d'accessibilité. Les différentes formes de Thème topical nous intéressent pour leur capacité à instruire d'une continuité, d'un déplacement ou d'une rupture. Or, il semble que des expressions dénotant des entités moyennement accessibles peuvent instruire d'une relation de continuité (comme les SN ou les noms propres avec reprise) et que des expressions associées à des degrés d'accessibilité très élevés peuvent, à l'inverse, forcer la continuité topicale dans des contextes de discontinuité (les pronoms de 3e personne dans PEOPL ou en première phrase de section dans ATLAS). Ces résultats montrent clairement que les notions d'accessibilité et de continuité diffèrent et que leur relation est très liée au type de texte considéré. Nous avons d'un côté une modélisation des processus cognitifs impliqués dans la représentation mentale d'un référent et de l'autre, une modélisation des techniques linguistiques utilisées pour répartir les informations référentielles à la surface du texte. Nous retrouvons alors une distinction semblable à celle posée entre la segmentation textuelle et la segmentation conceptuelle. Ce constat ne dévalue pas nos analyses. Nous avons toujours gardé les deux notions en parallèle. Nous espérions peut-être naïvement une corrélation forte entre degré d'accessibilité et signalement des différents degrés de (dis)continuité, or une telle corrélation n'est absolument pas révélée par nos résultats.

XI.3. Découvertes

Dernier aspect de nos résultats : la découverte de nouveaux fonctionnements discursifs et notamment du fonctionnement discursif des appositions. **Les appositions** ne sont généralement pas associées à l'organisation du discours, ni à la position initiale détachée. Combettes (2005) est certainement un des premiers et rares travaux supposant l'idée d'un potentiel cadratif lié aux appositions. Nos analyses nous montrent que les appositions détachées en initiale sont nombreuses dans les textes expositifs, que ces textes soient mono-référentiels ou pas. L'apposition permet au Thème topical qui la suit d'acquérir le statut de topique quelle que soit sa catégorie morpho-syntaxique. Mais plus encore, l'apposition constitue un indice pour caractériser la force de la continuité topicale. Ainsi, des continuités topicales peuvent dominer tous les modes organisationnels d'un texte, et même le découpage en sections. Notre petite étude des reprises des titres dans les différents sous-corpus nous a bien montré que l'apposition alliée à la présence d'un pronom en Thème topical pouvait constituer un indice dans l'identification des titres référentiels ou thématiques mais surtout dans la distinction entre des titres réalisant une rupture et des titres réalisant un déplacement dans la séquentialité du discours.

Toutes les configurations découvertes dans ce travail pourraient avec profit faire l'objet d'études plus poussées. En optant pour une méthode exploratoire, nous ne pouvions pas étudier dans le détail tous les phénomènes rencontrés. De ce fait, des analyses plus qualitatives ou portant un regard plus précis sur l'organisation discursive manquent à notre thèse.

XI.4. Perspectives

Notre étude du fonctionnement de la position initiale serait beaucoup plus complète si elle était accompagnée d'une étude comparative avec ce qui se passe en position non initiale. Cette comparaison entre position Thème et

position Rhème a fait l'objet de quelques études qui ont montré l'importance du rôle discursif des éléments détachés en position initiale (Thompson 1985, Le Draoulec & Péry-Woodley 2001). Cependant, comme nous l'avons vu dans différents exemples de continuités texto-stratégiques, les circonstants détachés en finale ou intégrés en tant que compléments du verbe peuvent signaler une étape et alors avoir une fonction similaire à celle des introducteurs de cadre. Il en va de même pour les localisations spatiales exprimées en position sujet. Ces quelques remarques constituent une remise en cause partielle de notre méthodologie. En effet, en associant l'expression des localisations spatiales et temporelles aux seuls circonstants détachés en initiale de phrase, nous avons quelque peu biaisé notre point de vue sur le phénomène. Seule une analyse qualitative permettrait de comprendre comment se construisent les TSC et sur la base de quels indices. Mais malgré ce manque, nous avons tout de même appris un peu plus sur le fonctionnement des TSC et notamment sur leur rapprochement avec les structures énumératives et leur influence sur le fonctionnement discursif des adverbiaux circonstanciels.

Deux autres éléments d'analyses sont également nécessaires pour compléter ce travail : la prise en compte du lexique et la prise en compte de la variation intra-textuelle.

Les éléments présents en position initiale constituent les points d'entrée dans les unités textuelles (le texte, les sections, les paragraphes, les phrases). Selon le principe de l'information cruciale en premier, les éléments perçus en premier orientent notre interprétation. Si ce principe est valide, ce que nous défendons, nous pourrions distinguer le lexique de la position initiale d'un texte du lexique général d'un texte. Ce lexique serait alors représentatif de la thématique et du domaine dans lesquels s'inscrit un texte. Nous avons pris en compte la récurrence lexicale pour caractériser les différents sous-corpus. Une analyse plus poussée aurait distingué parmi ces noms récurrents ceux qui apparaissaient significativement plus en position initiale. D'un autre point de vue, il aurait été certainement très pertinent de considérer la présence d'un nom récurrent en initiale de phrase comme un indice potentiel de séquentialité. Nous aurions pu alors mesurer les variations de la répartition des noms récurrents selon les différentes positions textuelles.

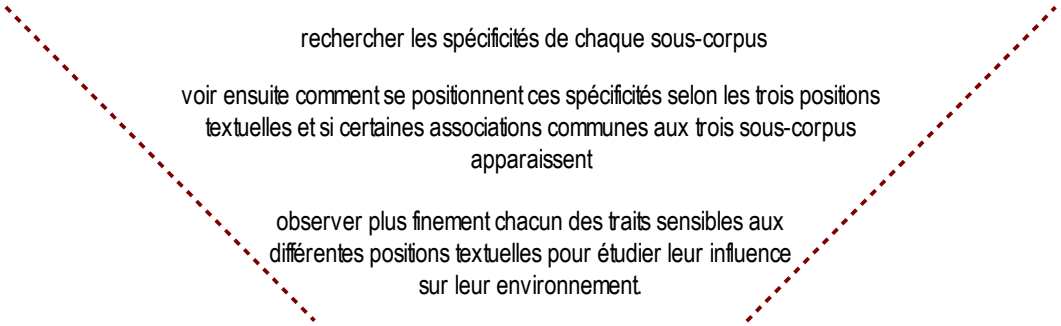
La prise en compte du lexique nous aurait appris davantage sur la gestion des continuités référentielles dans les différents sous-corpus et aurait permis de considérer davantage le contenu propositionnel du texte. Dans le cas des continuités référentielles, les liens entre propositions peuvent autant se construire par l'utilisation de pronominalisations, de reprises lexicales ou de réduction de termes que par des liens lexicaux. Comme le souligne Mahlberg (2006:363) dans une étude plus lexicale à la méthodologie très proche de la nôtre :

"Cohesion is created by interlocking lexico-grammatical patterns and overlapping lexical items"

Dans l'optique d'une utilisation applicative de notre travail, cette prise en compte du lexique est essentielle. Nous avons plusieurs fois mentionné l'intérêt grandissant en linguistique du discours pour les titres, les éléments détachés en initiale et les Thèmes topicaux. Ces éléments ont la particularité de jouer à la fois au niveau de la composante textuelle et au niveau de la composante idéationnelle. La compréhension des interactions entre ces trois types d'éléments nous apprend beaucoup sur l'organisation des textes tout en nous donnant un accès direct au contenu, *i.e.* aux éléments constitutifs du *text-world* véhiculé par le texte. Cet accès au contenu peut certainement être utile pour des tâches de TAL telles que la recherche d'information ou la représentation du contenu.

Autre point manquant à notre travail : la prise en compte de la variation intra-textuelle. Nous soutenons fortement l'implication du facteur de variation textuelle dans l'organisation du discours. Et pourtant, nous n'avons pas considéré un aspect très important de cette variation : le fait qu'un texte n'est pas un 'sac' de sections, de paragraphes, de phrases. Un texte peut répondre à plusieurs stratégies textuelles et ces stratégies peuvent être totalement différentes si

l'on se situe en introduction, en conclusion ou dans le développement du texte. Le travail de Biber & Finegan (1994) a mis en évidence cette variation intra-textuelle. Le travail de Teufel (1996) se base justement sur les variations de surface occasionnées par l'organisation globale d'un texte. Nous aurions aimé étudier ce facteur de variation, mais nous avons préféré nous concentrer sur la variation entre types de texte et entre positions textuelles. Deux raisons peuvent expliquer ce choix. Tout d'abord, comme l'ont conclu Biber & Finegan (1994), la variation intra-textuelle est beaucoup moins grande que celle observée entre textes de types différents. Ensuite, le rapport entre l'ajout de complexité et l'apport d'information ne nous paraissait pas satisfaisant. La combinatoire entre le facteur de variation intra-textuelle et le facteur de position textuelle n'est pas évidente. Soit nous aurions dû analyser les variations de positions textuelles à l'intérieur des 'débuts de texte', 'milieu de texte' et 'fin de texte', soit nous aurions eu deux facteurs de variation non corrélés qui offraient deux points de vue différents mais difficilement reliables. Nous avons préféré un parcours d'analyse où chaque étape d'analyse servait de base pour les étapes suivantes :



Malgré l'absence de comparaison avec la répartition des éléments en position non-initiale, la non prise en compte d'éléments lexicaux et de la variation intra-textuelle, cette thèse apporte la confirmation que la position initiale constitue une point de vue pertinent pour l'étude de l'organisation discursive. La position initiale en tant que point d'entrée dans des unités textuelles plus ou moins grandes est a priori universelle, car elle relève plus de la perception que d'une quelconque compétence linguistique. Il est en effet évident que ce que l'on perçoit en premier influence énormément notre interprétation de ce qui suit (que ce soit un texte, un film, une musique, un lieu, une personne, etc.). Par contre, ce que l'on met en position initiale des unités textuelles relève de la compétence linguistique. Sur ce point, il n'y a certainement pas d'universalité. Les langues à contraintes V2 par exemple laissent sans doute moins de liberté dans la composition de la position initiale. Les langues sans obligation de sujet donnent peut-être plus de liberté. Les fonctions des différents éléments détachés en initiale peuvent également être révélateurs du système d'une langue. Il semble y avoir peu de constructions appositives détachées en initiale dans les productions en langue anglaise. Notre conception des appositions comme indices de continuité topicale dominante est alors difficilement transposable à l'anglais. Certaines langues ne montrent pas tous les types d'éléments détachés en initiale que nous avons distingués. Ainsi, en japonais, les modalités d'énonciation n'apparaissent pas sous forme d'adverbiaux mais sont intégrées dans la sémantique du verbe. Elles ne peuvent donc pas acquérir le statut d'adverbial de phrase et de Thème interpersonnel.

La comparaison de la composition initiale à travers différents systèmes linguistiques apporterait des informations pertinentes sur le rôle de la position initiale dans l'organisation du discours mais surtout sur la diversité conceptuelle des langues. La méthodologie que nous avons mis en place s'adapterait tout à fait à la réalisation d'études contrastives exploratoires, permettant de découvrir les points communs et les spécificités de différentes langue dans leur utilisation de la position initiale dans l'organisation du discours.

Index des notions

Index des auteurs

Bibliographie

Adam J.-M. (1992) *Les Textes : types et prototypes*. Nathan : Paris

Adam J.-M. (1997) 'Genres, textes, discours : pour une reconception du concept de genre'. *Revue Belge de philologie et d'histoire* n°75 vol.3 (665-681)

Adam J.-M. (1999) *Linguistique textuelle. Des genres de discours aux textes*. Nathan : Paris

Allwood J. (1996) 'Some Comments on Wallace Chafe's "How consciousness Shapes Language"'. *Pragmatics and Cognition* n°4 vol.1 (55-64)

Alvarez de Mon y Rego I. (2001) 'Encapsulations and prospection in written scientific English'. *Estudios ingleses de la Universidad Complutense* n°9 (81-102)

Ariel M. (1990) *Accessing noun phrase antecedents*. Routledge: London

Ariel M. (2001) 'Accessibility Theory: an overview'. In T. Sanders, J. Schilperoord & W. Spooren (eds) *Text Representation: Linguistic and Psycholinguistic aspects*. John Benjamins: Amsterdam/Philadelphia

Asher N. (1993) *Reference to Abstract Objects in discourse*. Kluwer Academic Publishers: Dordrecht

Asher N. & Vieu L. (2005) 'Subordinating and Coordinating discours relations'. *Lingua* n°115 vol.4 (591-610)

Austin J. (1970) *Quand dire c'est faire*. Seuil : Paris

Berman R.A. & Nir-Sagiv B. (2007) 'Comparing Narrative and Expository text construction across adolescence: a developmental paradox'. *Discourse processes* n°43 vol.2 (79-120)

Bestgen Y. (1998) 'Segmentation markers as trace and signal of discourse structure'. *Journal of Pragmatics* n°29 (753-763)

Bestgen Y. & Costermans J. (1997) 'Temporal markers of narrative structure: studies in production'. In J. Costermans & M. Fayol (eds) *Processing interclausal relationships: studies in the production and comprehension of text*. Lawrence Erlbaum Associates: Mahwah, New Jersey (201-218)

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*. Neerlandistiek VU & Nodus Publikationen: Amsterdam/Münster (179-188)

Bestgen Y., Degand L. & Spooren W. (2006) 'Toward Automatic Determination of the Semantics of Connectives in Large Newspaper Corpora'. *Discourse Processes* n°41 vol.2 (175-193)

Bestgen Y. & Vonk W. (1995) 'The role of temporal segmentation markers in discourse processing'. *Discourse Processes* n°19 (385-406)

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

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

Biber D. (1992) 'On the complexity of Discourse complexity: a Multidimensional Analysis'. *Discourse Processes* n°15 (133-163)

Biber D. (1995) *Dimensions of register variation*. Cambridge University Press: Cambridge, Massachusset

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

Biber D. & Finegan E. (1994) 'Intra-textual variation within medical research articles'. In N. Oostdijk & P. de Haan (eds) *Corpus-based research into language. In honour to Jan Aarts*. Rodopi: Amsterdam (201-221)

Biber D., Johansson S., Leech G., Conrad S. & Finegan E. (1999) *Grammar of spoken and written english*. Longman: London

Bilhaut F. (2006) *Analyse automatique de structures thématiques discursives, application à la recherche d'information*, Thèse de doctorat en informatique, Université de Caen, France

Bilhaut F. & Widlöcher A. (2006) 'LinguaStream: an integrated environment for Computaional Linguistics Experimentation'. Proceedings of the 11th Conference of the European Chapter of the Association of Computational Linguistics (EACL'06), Trento, Italy

Bolinger D. (1972) 'Linear Modification'. In F.W. Householder (ed) *Publications of the Modern Language Association of America Harmondsworth*, (firstprinted in Publications of the Modern Language Association of America n° 67 (1952)). Penguin: Harmondsworth (31-50)

Borillo A., Bras M. & Le Draoulec A. (2004) 'Marqueurs temporels et aspectuels du français'. In F. Corblin & H. de Swart (eds) *Sémantique formelle et données du français*. CSLI : Stanford

Bosredon B. (1997) *Les titres de tableaux. Une pragmatique de l'identification*. Presses Universitaires de France : Paris

Bouchard D. (2007) 'The origin of Language, and why it is as it is'. The Syntax Circle, avril 2007, Universiteit Leiden, Netherlands

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

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

Branca-Rosoff S. (1999) 'Types, modes et genres : entre langue et discours'. *Langage et société* n°87 (5-24)

Bras M., Le Draoulec A. & Vieu L. (2001) 'French Adverbial "Puis" between Temporal Structure and Discourse Structure'. *Semantic and Pragmatic Issues in Discourse and Dialogue: Experimenting with Current Theories*, CRiSPI series, vol.9. Elsevier: Amsterdam (109-146)

Bras M. & Le Draoulec A. (2007) '"Alors" as a possible temporal Connective in Discourse'. *Cahiers Chronos* n°17

Britton L. (1996) *Pragmatics markers in English: grammaticalization and Discourse Functions*. Mouton de Gruyter: Berlin/New-York

Bruner J. (1986) *Actual minds, possible worlds*. Harvard University Press: Cambridge, Massachusset

Carlson L., Marcu D. & Okurowski M.E. (2003) '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*. Kluwer Academic Publishers: Dordrecht (85-112)

Chafe W.L. (1976) 'Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View'. In C.N.Li (ed) *Subject and Topic*. Academic Press: New York/San Francisco/London (25-55)

Chafe W.L. (1994) *Discourse Consciousness and Time: The flow and displacement of conscious experience in speaking and writing*. University of Chicago Press: Chicago

Charaudeau P. (2006) 'Discours journalistique et positionnements énonciatifs. Frontières et dérives'. *Semen : énonciation et responsabilité dans les médias*, http://semen.revues.org/document2793.htmlm, n°22

Charolles M. (1983) 'Coherence as a principle in the interpretation of discourse'. *Text* n°3 (71-97)

Charolles M. (1988) 'Les plans d'organisation textuelle : périodes, chaînes, portées et séquences'. *Pratiques* n°57 (3-13)

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

Charolles M. (1997) 'L'encadrement du discours; univers champs domaines et espaces'. *Cahier de Recherche Linguistique*, LanDisCo université Nancy2, n°6

Charolles M. (2002) *La référence et les expressions référentielles en français*. Orphys : Paris

Charolles M. (2003) 'De la topicalité des adverbiaux détachés en tête de phrase'. *Travaux de linguistique* n°47 (11-51)

Charolles M., Le Draoulec A., Péry-Woodley M.-P. & Sarda L. (2005) 'Temporal and spatial dimensions of discourse organisation'. *Journal of French Language Studies* n°15 vol.2 (203-218)

Charolles M. & Péry-Wooldey M.-P. (2005) 'introduction'. *Langue Française* n°148 (3-8)

Charolles M. & Vigier D. (2005) 'Les adverbiaux en position préverbale : portée cadrative et organisation des discours'. *Langue Française* n°148 (9-30)

Combettes B. (1998) *Les constructions détachées en Français*. Orphys : Paris

Combettes B. (2005) 'les constructions détachées comme cadres de discours'. *Langue française* n°148 (31-44)

Combettes B. & Tomassone R. (1988) *Le Texte Informatif, Aspects Linguistiques*. DeBoeck-Duculot : Louvain

Condamines A. (2003) *Sémantique et corpus spécialisés : Consitution de bases de connaissances terminologiques*, mémoire d'HDR en sciences du langage, Carnets de Grammaire, ERSS, Toulouse, France

Condamines A. (ed) (2005) *Sémantique et corpus*. Hermès : Paris

Conte M.-E. (1996) 'Anaphoric encapsulation'. *Belgian Journal of Linguistics: Coherence & Anaphora* n°10 (1-9)

Corblin F. (1985) 'Les chaînes de référence : analyse linguistique et traitement automatique'. *Intellectica* n°5 vol.1 (123-143)

Corblin F. (1987) *Indéfini, défini et démonstratif. Constructions linguistiques de la référence*. Droz : Genève
Corblin F. (1995) *Les formes de reprise dans le discours : Anaphores et chaînes de référence*. Presses Universitaires de Rennes : Rennes
Cornish F. (1996) 'Coherence: the lifeblood of anaphora'. *Belgian Journal of Linguistics* n°10 (37-54)
Cornish F. (1998) 'Les "chaînes topicales" : leur rôle dans la gestion et la structuration du discours'. *Cahiers de Grammaire* n°23 (19-40)
Cornish F. (2000) 'L'accessibilité cognitive des référents le Centrage d'attention et la structuration du discours : une vue d'ensemble'. *Verbum* n°22 vol.1 (7-30)
Cornish F. (2001a) 'L'inversion locative en français, italien et anglais : propriétés syntaxiques, sémantiques et discursives'. *Cahiers de grammaire* n°26 (101-113)
Cornish F. (2001b) 'Anaphora Text and the Construction of Discourse: A Practical Application'. In L.Degand, Y.Bestgen, W.Spooren & L.Van Waes (eds) *Multidisciplinary Approaches to Discourse*. Stichting Neerlandistiek & Nodus Publikationen: Amsterdam/Münster (111-122)
Cornish F. (2002) 'Anaphora: lexico-textual structure or means for utterance integration within a discourse ? A critique of the Functional Grammar account'. *Linguistics* n°40 vol.3 (469-493)
Cornish F. (2003) 'The roles of (written) text and anaphor-type distribution in the construction of discourse'. *Text* n°23 vol.1 (1-26)
Cornish F. (2005) 'Une approche pragmatico-discursive des phrases "thétiques"'. In H. Nølke & F. Lambert (eds) *Mélanges en l'honneur de Claude Muller : la syntaxe au cœur de la grammaire*. Presses Universitaires de Rennes : Rennes (75-84)
Crompton P. (2006) 'The effect of position on the discourse scope of adverbials'. *Text and Talk* n°26 vol.3 (245-279)
Daneš F. (1974) 'Functional sentence perspective and the organisation of the text: Different types of Thematic Progression'. In F.Daneš (ed) *Papers on functional sentence perspective*. Mouton de Gruyter: La Hague (106-128)
Danlos L. (2004) 'Discourse Dependency Structures as Constrained DAGs'. Proceedings of the 5th SIGdial Workshop on Discours and Dialogue, Cambridge, Massachussets (127-135)
Danlos L. (2007) 'D-STAG : un formalisme pour le discours inspiré des TAG synchrones'. Actes de TALN'2007, Toulouse, France
De Mulder W. (1994) 'Déterminants cohérence et raisonnement par défaut'. *Travaux de linguistique 'La cohérence textuelle : cohésion et rupture'* n°29 (21-37)
De Mulder W. (1997) 'Les démonstratifs : des indices de changement de contexte'. In N. Flaux, D. Van de Velde, W. de Mulder (ed) *Entre général et particulier : les Déterminants*. Artois Press Université : Besançon (137-200)
De Mulder W. (2000) 'Démonstratifs et accessibilité'. In (ed) *Verbum* (103-125)
Degand L. & Bestgen Y. (2004) 'Connecteurs et analyses de corpus : de l'analyse manuelle à l'analyse automatisée'. In S. Porhiel & D. Klingler (eds) *l'unité texte*. Perspectives : Pleyben (49-74)
Delin J. (1989) *Cleft Constructions in Discourse*, Ph.D. thesis in Cognitive Science, University of Edinburgh, UK
Dik S.C. (1997) *Theory of Funtional Grammar Complex and Derived Constructions*. Mouton de Gruyter: Berlin/New-York
Do-Hurinville D.T. (2004) *Temps et aspect en vietnamien : étude comparative avec le français*, Thèse de doctorat en sciences du langage, Université Paris VII-Denis Diderot, France
Downing A. (2001) 'Thematic progression as a functional resource in analysing texts'. *CLAC (Circulo de Lingüistica Aplicada a la Communicacion)*, http://www.ucm.es/info/circulo/no5/downing.htm, n°5
Dubois B.L. (1987) 'A reformulation of thematic progression typology'. *Text* n°7 vol.2 (89-116)
Dupont M. (2003) *Une approche cognitive du calcul de la référence*, Thèse de doctorat en informatique, Université de Caen, France
Enjalbert P. & Gaio M. (eds) (2004) *Approches sémantiques du document numérique*, Actes du 7e Colloque International sur le Document Électronique (CIDE.7), La Rochelle, France
Enjalbert P. (ed) (2005) *Sémantique et TAL*. Hermès : Paris
Enkvist N.E. (1973) '"Theme Dynamics" and style: an experiment'. *Studia Anglica Posnaniensia* n°5 (127-135)
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*. Publications of the research institute of the Abo Akademi Foundation: Abo (51-74)
Enkvist N.E. (1978) 'Coherence, Pseudo-Coherence, and Non-Coherence'. In J.-O. Östman (ed) *Reports on Text Linguistics: 'Cohesion and Semantics'*. Publications of the research institute of the Abo Akademi Foundation: Abo (109-127)
Enkvist N.E. (1981) 'Experiential iconicism in text strategy'. In (ed) *Text* (97-111)
Enkvist N.E. (1985) 'A parametic view of word order'. In E.Sözer (ed) *Text Connexity Text Coherence: Aspects Methods Results*. Helmut Buske: Hamburg (320-336)

Enkvist N.E. (1989) 'Connexity, Interpretability, Universes of Discourse, and Text Worlds'. In J. Allén (ed) *Possible Worlds in Humanities, Arts and Sciences*. Walter de Gruyter: Berlin/New-York (162-186)

Erteschik-Shir N. (1997) *The dynamics of focus structure*. Cambridge University Press: Cambridge, Massachusset

Fauconnier G. (1984) *Espaces mentaux. Aspects de la construction du sens en langue naturelle*. Editions de Minuit : Paris

Fauconnier G. (1985) *Mental Spaces. Aspects of meaning in natural language*. MIT Press: Cambridge, Massachusset

Fayol M. (1997) 'On Acquiring and using punctuation: A study of written french'. In J. Costermans & M. Fayol (eds) *Processing interclausal relationships: studies in the production and comprehension of text*. Lawrence Erlbaum Associates: Mahwah, New Jersey (157-178)

Ferret O. (2002) 'Segmenter et structurer thématiquement des textes par l'utilisation conjointe de collocations et de la récurrence lexicale'. *TALN 2002* (155-165)

Ferret O. & Grau B. (1998) 'Structuration d'un réseau de cooccurrences lexicales en domaines sémantiques par analyse de textes'. Actes NLP+IA, Moncton, Canada (220-226)

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

Ferret O. & Grau B. (2000) 'A Topic Segmentation of Texts based on Semantic Domains'. Proceedings of ECAI 2000, Berlin, Allemagne (426-430)

Firbas J. (1959) 'Thoughts on the Communicative Function of the Verb in English, German and Czech'. *Brno Studies in English* vol.1 (39-63)

Firbas J. (1964) 'On defining Theme in functional sentence analysis'. *Travaux linguistiques de Prague* n°2 (239-256)

Firbas J. (1992) *Functional Sentence Perspective in written and spoken communication*. Cambridge University Press: Cambridge, Massachusset

Foucambert D. (2003) *syntaxe, vision parafovéale et processus de lecture*, Thèse de doctorat en sciences de l'éducation, Université Grenoble-II, France

Fournier N. & Fuchs C. (1998) 'Place du sujet nominale et opérations de thématisation'. *Cahiers de praxématique* n°30 (55-88)

Francis G. (1989) 'Thematic selection and distribution in written discourse'. *Word* n°40 (201-221)

Frérot C. (2005) *Construction et évaluation en corpus variés de lexiques syntaxiques pour la résolution des ambiguités de rattachement prépositionnel*, Thèse de doctorat en sciences du langage, Université Toulouse-leMirail, France

Fries P. (1995a) 'A personal view of Theme'. In M.Ghadessy (ed) *Thematic Development in English texts*. Pinter Publishers: London (1-19)

Fries P. (1995b) 'Patterns of information in initial position in english'. In P.Fries & M.Gregory (eds) *Meaning and choice in language: studies for Michael Halliday* (47-66)

Fries P. (1995c) 'Themes Method of Development and texts'. In R.Hasan & P.Fries (eds) *On Subject and Theme: A Discourse Functional Perspective*. John Benjamins: Amsterdam/Philadelphia (317-359)

Fuchs C. & Fournier N (2003) 'Du rôle cadratif des compléments localisants initiaux selon la position du sujet grammatical'. *Travaux de linguistique* n°47 (79-109)

Fuchs C. (ed) (1997) *La place du sujet en français contemporain*. Duculot : Louvain

Gaddy M.L., Van Den Broek P. & Sung Y.-C. (2001) 'The influence of text cues on the allocation of attention during reading'. In T.Sanders, J.Schilperoord & W.Spooren (eds) *Text representation: Linguistic and Psycholinguistic aspects*. John Benjamins: Amsterdam/Philadelphia (89-109)

Gaume B. (2004) 'Balades aléatoires dans les petits mondes lexicaux'. *I3 Information, Interaction, Intelligence*, http://www.revue-i3.org/volume04/numero02/revue_i3_04_02_02.pdf, n°2 vol.4

Gayral F., Jacques M.-P., Poibeau T. & Zimina M. (2007) 'Typologie textuelle : état de l'art et applications'. rapport du projet RNTL TEXTCOOP, LIPN-Paris 13, France

Gernsbacher M.A. (1990) *Language comprehension as structure building*. Lawrence Erlbaum Associates: Mahwah, New Jersey

Gernsbacher M.A. (1995) 'The Structure Building Framework: What it is, what it might also be, and why'. In B. K. Britton & A. C. Graesser (eds) *Models of text understanding*. Lawrence Erlbaum Associates: Mahwah, New Jersey (289-311)

Gernsbacher M.A. (1997) 'Coherence Cues Mapping During Comprehension'. In J. Costermans & M. Fayol (eds) *Processing interclausal relationships: studies in the production and comprehension of text*. Lawrence Erlbaum Associates: Mahwah, New Jersey (3-21)

Gernsbacher M.A. & Robertson R.R.W. (2002) 'The definite article the as a cue to map thematic information'. In M. Louwerse & W. van Peer (eds) *Thematics, interdisciplinary studies*. John Benjamins: Amsterdam/Philadelphia (119-136)

Gil P.O. (2001) 'Extended thematic progression'. *Miscelanea: A journal of English and American studies* n°23 (1-19)

Givón T. (1983) 'Topic continuity in discourse: an introduction'. In T.Givon (ed) *Topic continuity in discourse: a quantitative cross-language study*. John Benjamins: Amsterdam/Philadelphia (1-42)
Givón T. (1988) 'the pragmatics of word order: predicability, importance and attention'. In M. Hammond, E. Moravcsik & J. Wirth (eds) *Studies in Syntactic Typology*. John Benjamins: Amsterdam/Philadelphia (243-284)
Givón T. (1990) *Syntax. A Functional-Typological approach, Vol. 2*. John Benjamins: Amsterdam/Philadelphia
Givón T. (1995) *Functionalism and Grammar*. John Benjamins: Amsterdam/Philadelphia
Gómez-González M.A. (2001) *The Theme-Topic Interface. Evidence form English*. John Benjamins: Amsterdam/Philadelphia
Gosselin L. (1990) 'Les circonstanciels : de la phrase au texte'. *Langue Française* n°86 (37-45)
Goutsos D. (1996) 'A model of sequential relations in expository test'. *Text* n°16(4) (501-533)
Grévisse M. (1993) *Le bon usage*. Duculot : Louvain
Grobet A. (2002) *L'identification des topiques dans les dialogues*. Duculot : Louvain
Grosz B.J., Joshi A. & Weinstein S. (1995) 'Centering: A framework for modelling the local coherence of dicourse'. *Computational Linguistics* n°21 vol.2 (203-225)
Grosz B.J. & Sidner C.L. (1986) 'Attention Intentions and the structure of discourse'. *Computational Linguistics* n°12 vol.3 (175-204)
Guimier C. (1993) 'L'établissement d'un corpus de circonstants'. In C. Guimier (ed) *1001 circonstants*. Presses Universitaires de Caen : Caen (11-45)
Guimier C. (1996) *Les adverbes du français. Le cas des adverbes en -ment*. Orphys : Paris
Gundel J.K. (2002) 'It-clefts in English and Norwegian'. In B. Behrens, C. Fabricius-Hansen, H. Hasselgård, & S. Johansson (eds) *Information structure in a cross-linguistic perspective*. Rodopi: Amsterdam (113-128)
Gundel J.K., Hedberg N. & Zacharski R. (1993) 'Cognitive status and the form of referring expressions in discourse'. In (ed) *Language* (274-307)
Gundel J.K., Hedberg N. & Zacharski R. (2000) 'Statut cognitif et forme des anaphoriques indirects'. *Verbum* n°22 vol.1 (79-102)
Habert B., Nazarenko A. & Salem A. (1997) *Les linguistiques de corpus*. Armand Colin : Paris
Haladi H., Ermosilla H. & Eyrolle H. (2002) 'Titres et focalisation attentionnelle dans la lecture et la recherche d'information dans un texte expositif'. actes du colloque 'Inscription spatiale du langage : structures et processus' (ISLsp'02), Janvier 2002, Toulouse, France. Prescot : Toulouse (115-121)
Halliday M.A.K. (1970) 'Language structure and language function'. In R.Hasan & P.Fries (eds) *New horizons in Linguistics*. Penguin: Harmondsworth (140-164)
Halliday M.A.K. (1971/2002) 'Linguistic Function and Literary Style: An Inquiry into the Language of William Golding's The Inheritors'. In J. Webster (ed) *Linguistic Studies of Text and Discourse (Last publication)*. Works of MAK Halliday Continuum: London (88-125)
Halliday M.A.K. (1985) *An introduction to Functional Grammar*. Edward Arnold: London
Halliday M.A.K. & Hasan R. (1976) *Cohesion in English*. Longman: London
Hasselgård H. (1996) *Where and When: Positional and functional conventions for sequences of time and space adverbials in present-day English*. Scandinavian University Press, Acta Humaniora: Oslo
Hasselgård H. (2004a) 'Adverbials in it-cleft constructions'. In K.Aijmer & B. Altenberg (eds) *Language and Computers*, Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23) 22-26 May 2002, Göteborg, Sverige. Rodopi: Amsterdam (195-211)
Hasselgård H. (2004b) 'The role of multiple themes in cohesion'. In K.Aijmer & A.-B.Stenstöm (eds) *Discourse Patterns in Spoken and Written Corpora*. John Benjamins: Amsterdam/Philadelphia (65-87)
Hearst M.A. (1994) 'Multi-paragraph segmentation of expository texts'. 32e meeting of the association for computational linguistics (ACL 94), http://www.sims.berkeley.edu/~hearst/papers/tiling-acl94/acl94.html (9-16)
Hearst M.A. (1997) 'TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages'. *Computational Linguistics* n°23 vol.1 (33-64)
Hernandez N. (2004) *Description et Détection Automatique de Structures de texte*, Thèse de doctorat en informatique, Université d'Orsay - Paris XI, France
Heurley L. (1994) *Traitement de textes proceduraux: étude psycholinguistique cognitive des processus de production et de compréhension chez des adultes non experts*, Thèse de doctorat en psychologie, Université de Bourgogne, France
Heurley L. (1997) 'Processing units in written texts: paragraphs or information blocks?'. In J. Costermans & M. Fayol (eds) *Processing interclausal relationships: studies in the production and comprehension of text*. Lawrence Erlbaum Associates: Mahwah, New Jersey (179-200)

Hồ-Đắc L.-M. (1999) *Au carrefour des cadres et des registres*, Mémoire de Master1 en sciences du langage, Université Toulouse-Le Mirail, France
Hồ-Đắc L.-M. (2000) *Méthode d'analyse et de représentation de l'encadrement du discours*, Mémoire de Master2 en sciences du langage, Université Toulouse-Le Mirail, France
Hồ-Đắc L.-M. (2005) 'Deux modes de segmentation textuelle : univers de discours et chaînes de référence'. *Verbum* n°27 vol.3 (231-248)
Hồ-Đắc L.-M. & Frérot C. (2004) 'Approche discursive et approche syntaxique des circonstants en corpus'. journées ATALA, La Rochelle-France
Hồ-Đắc L.-M., Jacques M.-P. & Rebeyrolle J. (2004) 'Sur la fonction discursive des titres'. In S. Porhiel & D. Klinger (eds) *L'unité texte*. Perspectives : Pleyben (125-152)
Hồ-Đắc L.-M. & Laignelet M. (2005) 'Temporal Structure and Thematic Progression: A Case Study on French Corpora'. Symposium on the Exploration and Modelling of Meaning (SEM-05), Connectives, discourse framing and discourse structure: from corpus-based and experimental analyses to discourse theories, 14-15 nov 2005, Biarritz, France
Hồ-Đắc L.-M., Le Draoulec A. & Péry-Woodley M.-P. (2001) 'Cohabitation des dimensions temps espace et 'phénomènes''. *Cahiers de Grammaire* n°26 (125-142)
Hoek L.H. (1981) *La marque du titre*. Mouton de Gruyter : Berlin/New-York
Jackiewicz A. (2005) 'Les séries linéaires dans le discours'. *Langue Française* n°148 (95-110)
Jacques M.-P. (2003) *Approche en discours de la réduction des termes comlexes dans les textes spécialisés*, Thèse de doctorat en sciences du langage, Université Toulouse-le Mirail, France
Jacques M.-P., Rebeyrolle J. & Hồ-Đắc L.-M. (2004) 'Quelques aspects méthodologiques d'une étude de la fonction discursive des titres en corpus'. journées ATALA, La Rochelle, France
Jacques M.-P. & Rebeyrolle J. (2006) 'Titres et structuration des documents'. Actes du Colloque international 'Discours et Document' (ISDD'06), Juin 2006, Caen, France (1-12)
Johnson-Laird P. (1983) *Mental Models*. Cambridge University Press: Cambridge, Massachusset
Kennedy G. (1998) *An introduction to Corpus Linguistics*. Longman: London
Kintsch W. (1988) 'The use of knowledge in discourse processing: A construction-incrementation model'. *Psychological Review* n°85 (363-394)
Kintsch W. (1995) 'How Readers Construct Situation Models for Stories: The role of syntactic cues and causal inferences'. In M.A. Gernsbacher & T. Givón (eds) *Coherence in Spontaneous Text*. John Benjamins: Amsterdam/Philadelphia (139-160)
Kintsch W. (2002) 'On the notions of theme and topic in psychological process models of text comprehension'. In M. Louwerse & W. van Peer (eds) *Thematics, interdisciplinary studies*. John Benjamins: Amsterdam/Philadelphia (157-170)
Kintsch W. & Van Dijk T.A. (1983) *Strategies of Discourse Comprehension*. New-York Academic Press: New-York
Kleiber G. (1981) *Problèmes de référence : descriptions définies et noms propres*. Klincksieck : Paris
Kleiber G. (1984) 'Sur la sémantique des descriptions démonstratives'. *Linguisticae Investigationes* n°VIII vol.1 (63-85)
Kleiber G. (1989) 'Reprise(s). Travaux sur les processus référentiels anaphoriques'. Publication du groupe "Anaphore et Deixis" Université de Sciences humaines de Strasbourg
Kleiber G. (1990a) 'Marqueurs référentiels et processus interprétatifs : pour une approche plus sémantique'. *Cahiers de Linguistique Française* n°11 (241-258)
Kleiber G. (1990b) 'Article défini et démonstatif : approche sémantique versus approche cognitive'. In G.Kleiber & J.-E.Tyvaert (eds) *L'anaphore et ses domaines*. Klincksieck : Paris (199-227)
Kleiber G. (1992) 'Anaphore-Deixis : deux approches concurrentes'. In M.-A.Morel & L.Danon-Boileau (eds) *La deixis* (613-626)
Kleiber G. (1994) *Anaphores et pronoms*. Duculot : Louvain
Klein W. (1994) *Time in Language*. Routledge: London
Knott A. (1996) ". *A Data-driven Methodology for motivating a set of coherence relations*, Ph.D. thesis in Artificial Intelligence, University of Edinburgh, UK,
Lahousse K. (2003a) 'La complexité de la notion de topique et l'inversion du sujet nominal'. *Travaux de Linguistique* n°47 (111-136)
Lahousse K. (2003b) *The distribution of postverbal nominal subject in French. A syntactic, semantic and pragmatic analysis*, Ph.D. in Linguistics, Katholieke universiteit Leuven, Belgium
Laignelet M. (2003) *Les cadres de discours spatiaux et temporels dans les documents géographiques : interactions et croisements*, Mémoire de Master2, Université de Toulouse-Le Mirail, France
Lambrecht K. (1994) *Information structure and sentence form. Topic focus and the mental representation of discourse referents*. Cambridge University Press: Cambridge, Massachusset

Lambrecht K. (2001) 'Dislocation'. In M.Haspelmath (ed) *Language Typology and Language Universals*. Walter de Gruyter: Berlin/New-York

Landauer T.K., Foltz P.W. & Laham D. (1998) 'Introduction to Latent Semantic Analysis'. *Discourse Processes* n°25 (259-284)

Langaker R. (1991) *Foundations of Cognitive Grammar vol. II: Descriptive Application*. Stanford University Press: Stanford

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) Proceedings of the Corpus Linguistics 2001 Conference, avril 2001, Lancaster, UK (159-166)

Le Draoulec A. & Péry-Woodley M.-P. (2003) 'Time in Travel: Temporal framing in narratives and non-narratives'. In L. Lagerwerf, W. Spporen & L. Degand (eds) *Determination of Information and Tenor in Texts: Multidisciplinary Approaches to Discourse 2003* (267-275)

Le Draoulec A. & Péry-Woodley M.-P. (2005) 'Encadrement temporel et relations de discours'. *Langue française* n°148 (45-60)

Le Querler N. (1993) 'Les circonstants et la position initiale'. In C. Guimier (ed) *1001 circonstants*. Presses Universitaires de Caen : Caen (159-184)

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

Legallois D. (2006) 'Quand le texte signale sa structure : la fonction textuelle des noms sous-spécifiés'. *Corela : Organisation des textes et cohérence des discours*, http://edel.univ-poitiers.fr/corela,

Lemarié J. (2006) *La compréhension des textes visuellement structurés. Le cas des énumérations*, Thèse de Doctorat, Université Toulouse-le Mirail, France

Lemarié J., Eyrolle H. & Cellier J.-M. (2004) 'Compréhension et mémorisation d'un texte transposé automatiquement à l'oral – le rôle de la typo-disposition'. Regards croisés sur l'unité texte, 18-20 mars, Nicosie, Chypre

Levelt W.J.M. (1981) 'The speaker's linearization problem'. *Philosophical Transactions Royal Society London* n°B295 (305-315)

Longacre R.E. (1976) 'An anatomy of speech notions'. In (ed) *Peter de Ridder Publications in Tagmemics*. Peter de Ridder: Lisse

Longacre R.E. (1979) 'the paragraph as a grammatical unit'. In T.Givon (ed) *Syntax and Semantics Discourse and syntax*. Academic Press: New York/San Francisco/London (115-134)

Lorch R.F., Jr. (1989) 'Text-signaling devices and their effects on reading and memory processes'. *Educational Psychology Review* n°1 (209-234)

Lorentz G. (1999) 'Learning to Cohere: causal links in native vs. non-native argumentation writing'. In W.Bublitz, U.Lenk & E.Ventola (eds) *Coherence in spoken and written discourse*, (selected papers from the international workshop on coherence AUGSBURG 24-27 April 1997). John Benjamins: Amsterdam/Philadelphia (55-75)

Luc C. (2000) *Représentation et composition des structures visuelles et rhétoriques du texte*, Thèse de Doctorat, Université Toulouse - Paul Sabatier, France

Luc C., Mojahid M. & Virbel J. (2001) 'Système notionnel de l'architecture textuelle par image de page'. In M. Mojahid & J. Virbel (eds) CIDE, 24-26 octobre 2001, IRIT-Toulouse, France (263-272)

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

Mahlberg M. (2006) 'Exploring the textual positions and functions of lexical items in hard news stories'. Corpus Linguistics 2007, 27-30 July 2007, University of Birmingham, UK

Malrieu D. & Rastier F. (2001) 'genres et variations morphosyntaxiques'. *TAL* n°42 vol.2 (548-577)

Mann W.C. & Thompson S.A. (1986) 'Rhetorical Structure Theory: a theory of texte organization'. In L.Polanyi & N.J.Norwood (eds) *The Structure of Discourse*. Ablex: Norwood

Mann W.C. & Thompson S.A. (1999) 'Introduction à la Théorie de la Structure Rhétorique (Rhetorical Structure Thory: RST)'. http://www.sfu.ca/rst/07french/introduction.html

Manning C. & Schütze H. (1999) *Foundations of statistical natural language processing*. MIT Press : Cambridge, Massachusset

Manuéalian H. (2004) *Descriptions définies et démonstratives : analyses de corpus pour la génération de textes*, Thèse de Doctorat, Université Nancy 2, France

Martin J.R. (2001) 'Cohesion and texture'. In D Schiffrin, D Tannen & H Hamilton (eds) *Handbook of Discourse Analysis*. Blackwell: Oxford (35-53)

Masson N. (1995) 'An Automatic Method for Document Structuring'. Proceedings of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA (372-373)

Mauger G. (1968) *Grammaire pratique du français : langue parlée, langue écrite*. Hachette : Paris

Maurel F., Lemarié J. & Vigouroux N. (2003) 'Oralisation de structures visuelles : de la lexico-syntaxe à la prosodie'. In A. Mettouchi & G. Ferré (eds) *Interface Prosodique 2003 (IP2003)*, 27-29 Mars 2003, Nantes, France (137-142)

Maurel F., Luc C, Vigouroux N, Mojahid M., Virbel J. & Nespoulous J.-L. (2002) 'Transposition à l'oral des structures énumératives à partir de leurs paramètres formels'. actes du colloque 'Inscription spatiale du langage : structures et processus' (ISLsp'02), Janvier 2002, Toulouse, France. Prescot : Toulouse (179-194)

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

Mélis L. (1983) *Les circonstants et la phrase. Étude sur la classification et la systématique des compléments circonstanciels en français moderne*. Presses Universitaires de Louvain : Leuven

Miltsakaki E., Prasad R., Joshi A. & Webber B. (2004) 'The Penn Discourse Treebank'. Proceedings of the Language Resources and Evaluation Conference, Lisbon, Portugal

Mourad G. (2001) *Analyse informatique des signes typographiques pour la segmenation des textes et l'extraction automatique des citations. Réalisation des Applications informatiques : SegATex et CitaRE*, Thèse de Doctorat, Université Paris-Sorbonne, France

Muller C. (1968) *initiation à la statistique linguistique*. Larousse : Paris

Nadau R. (1999) *Vocabulaire technique et analytique de l'épistémologie*. Presses Universitaires de France : Paris

Neveu F. (1998) *Etudes sur l'apposition. Aspects du détachement nominal et adjectival en français contemporain, dans un corpus de textes de J.-P. Sartre*. Honoré Champion : Paris

Östman J.-O. & Virtanen T. (1999) 'Theme, comment and newness as figures in information structure'. In K. van Hoek, A. A. Kibrik & L. Noordman (eds) *Discourse Studies in Cognitive Linguistics*, (selected papers from the international workshop on coherence AUGSBURG 24-27 April 1997). John Benjamins: Amsterdam/Philadelphia (91-110)

Pascual E. (1991) *Représentation de l'architecture textuelle et génération de texte*, Thèse de Doctorat, Université Toulouse 3 - Paul Sabatier, France

Péry-Woodley M.-P. (1989) *Textual design: signalling coherence in first and second language academic writing*, PhD, Universty of Lancaster, UK

Péry-Woodley M.-P. (2000a) 'Cadrer ou centrer son discours ? Introducteurs de cadres et Centrage'. *Verbum* n°22 vol.1 (59-78)

Péry-Woodley M.-P. (2000b) *Une pragmatique à fleur de texte: approche en corpus de l'organisation textuelle*, Mémoire d'HDR, Université Toulouse-Le Mirail, France

Péry-Woodley M.-P. (2001) 'présentation'. *Verbum*, (présentation du numéro intitulé 'cohérence et relations de discours à l'écrit'), n°23 vol.1 (3-8)

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

Piérard S. & Bestgen Y. (2005) 'Deux indices pour l'étude de la continuité thématique dans de grands corpus'. 4e journées de Linguistique de Corpus, Septembre 2005, Université de Bretagne-Sud

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

Pimm C. (2006) 'Quelle plus-value linguistique pour la segmentation automatique de texte?'. Actes du Colloque international 'Discours et Document' (ISDD'06), Juin 2006, Caen, France (85-94)

Porhiel S. (2001) 'Linguistic exrpressions as a tool to extract thematic information'. Corpus Linguistic 2001, Lancaster

Porhiel S. (2003) 'Les indicateurs d'intérêt dans l'organisation textuelle'. In B. Combettes, C. Schnedecker & A. Theissen (eds) *Ordre et distinction dans la langue et le discours*. Honoré Champion : Paris (425-441)

Prince E. (1981) 'Toward a Taxonomy of Given-New Information'. In P. Cole (ed) *radical Pragmatics*. New-York Academic Press: New-York

Quirk R., Greenbaum S., Leech G. & Svartvik J. (1972) *A Grammar of Contemporary English*

Quirk R., Greenbaum S., Leech G. & Svartvik J. (1985) *A comprehensive Grammar of the English language*. Longman: London

Quirk R. & Greenbaum S. (1993) *A university grammar of English*, (http://grammar.ccc.commnet.edu/grammar/). Longman: London

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

Reboul A. (1997) 'What (if anything) is accessible? A relevance-oriented criticism of Ariel's Accessibility Theory of referring expressions'. In J.H.Connolly, R.M.Vismans, C.S.Butler & R.A.Gatward (eds) *Discourse and pragmatics in functional grammar*. Mouton de Gruyter: Berlin/New-York (91-108)

Riegel M., Pellat J.-C. & Rioul R. (1994) *Grammaire méthodique du français*. Presses Universitaires de France : Paris

Romera M. (2004) *Discourse Functional Units*. LINCOM EUROPA: München

Sampson J. (1994) 'Susanne: a domesday book of english grammar'. In N. Oostdijk & P. De Haan (eds) *Corpus Based Research into Language*. Rodopi: Amsterdam (169-187)

Sanders T. & Gernsbacher M.A. (2004) 'Accessibilty in text and Discourse Processing'. *Discourse Processes* n°37 vol.2 (79-89)
Sanders T. & Spooren W. (2001) 'Text representation as an interface between language and its users'. In T.Sanders, J.Schilperoord & W.Spooren (eds) *Text representation: Linguistic and Psycholinguistic aspects*. John Benjamins: Amsterdam/Philadelphia (1-26)
Schiffrin D. (1987) *Discourse Markers*. Cambridge University Press: Cambridge, Massachusset
Schnedecker C. (1997) *Noms Propres et Chaînes de référence*. Université de Metz : Metz
Schnedecker C. (2003) 'La question du nom propre répété dans la théorie dite du centrage et ses problèmes'. *French Language Studies* n°13 (105-134)
Schnedecker C. (2005) 'Les chaînes de référence dans les portraits journalistiques : éléments de description'. *Travaux de Linguistique* n°51 vol.2 (85-133)
Schneuwly B. (1997) 'Textual organizers and text types: Ontogenetic aspects in writing'. In J. Costermans & M. Fayol (eds) *Processing interclausal relationships: studies in the production and comprehension of text*. Lawrence Erlbaum Associates: Mahwah, New Jersey (245-263)
Siepmann D. (2005) *Discourse Markers across languages: A contrastive study of second-level discourse markers in native and non-native text with implications for general and pedagogic lexicography*. Routledge: London
Sperber D. & Wilson D. (1986) *Relevance: Communication and cognition*. Blackwell: Oxford
Stark H.A. (1988) 'What do paragraph markings do?'. *Discourse Processes* n°11 (275-303)
Sullet-Nylander F. (1998) *Le titre de presse. Analyses syntaxique, pragmatique et rhétorique*, Doktorsavhandling, Stockholms universitet, Sverige
Sweetser E. & G. Fauconnier (1996) 'Cognitive Links and Domains: Basic Aspects of Mental Space Theory'. In G. Fauconnier & E. Sweetser (eds) *Spaces, Worlds and Grammar*. University of Chicago Press: Chicago (1-28)
Teufel S. (1998) 'Meta-discourse markers and problem-structuring in scientific articles'. Workshop on Discourse Structure and Discourse Markers, ACL 1998, Montreal
Teufel S. (1999) *Argumentative Zoning: Information Extraction from Scientific Text*, Ph.D. thesis, School of Cognitive Science, University of Edinburgh, Edinburgh, UK
Teufel S. & Moens M. (2002) 'Summarizing Scientific Articles -- Experiments with Relevance and Rhetorical Status'. *Computational Linguistics* n°28 vol.4 (409-445)
Thompson G. (2004) *Introducing Functional Grammar (2nd edition)*. Oxford University Press: Oxford
Thompson S. (1985) 'Grammar and Written Discourse: Initial vs. Final Purpose Clauses in English'. *Text* n°5 (55-84)
Thompson S. & Longacre R. (1985) 'Adverbial clauses'. In T.Shopen (ed) *Language typology and Syntactic Description*. Cambridge University Press: Cambridge, Massachusset (171-234)
Tognini-Bonelli E. (2001) *Corpus Linguistics at Work*. John Benjamins: Amsterdam/Philadelphia
Trosborg A. (1997) 'Text typology: Register, Genre and Text type'. In A. Trosborg (ed) *Text Typology and Translation*. John Benjamins: Amsterdam/Philadelphia (3-23)
Van den Broek P., Risden K., Fletcher C.R. & Thurlow R. (1996) 'A 'landscape' view of reading: Fluctuating patterns of activation and the construction of a stable memory representation'. In B.K. Britton & A.C. Graesser (eds) *Models of understanding text*. Lawrence Erlbaum Associates: Mahwah, New Jersey (165-187)
Vergez-Couret M. (2006) *Repérage en corpus des relations de discours et de leurs marquages : le cas particulier de la relation d'élaboration*, Mémoire de Master, Université Toulouse le Mirail, France
Vergez-Couret M. (2007) 'Repérer la relation d'élaboration : étude de l'interaction entre le participe présent et l'adverbe notamment'. Colloque AFLS, Atelier Doctorants, 3-5 septembre 2007, Boulogne-sur-mer, France
Vigier D. (2003) 'Les syntagmes prépositionnels en « en N » détachés en tête de phrase référant à des domaines d'activité'. *Linguisticae Investigaciones* n°26 vol.1 (97-122)
Vigier D. (2004) *Les groupes prépositionnels en "en N" : de la phrase au discours*, Thèse de Doctorat, Université Paris III, France
Viprey (2005) 'Corpus et sémantique discursive : éléments de méthode pour la lecture des corpus'. In A. Condamines (ed) *Sémantique et corpus*. Hermès : Paris (245-276)
Virbel J. (1986) 'Langage et Métalangage dans le texte du point de vue de l'édition en informatique textuelle'. *Cahiers de Grammaire* n°10 (1-72)
Virbel J. (2002) 'Elements d'analyse du titre'. actes du colloque 'Inscription spatiale du langage : structures et processus' (ISLsp'02), Janvier 2002, Toulouse, France. Prescot : Toulouse (123-134)
Virtanen T. (1992) *Discourse Functions of Adverbial Placement in English: Clause-Initial Adverbials of Time and Place in Narratives and Procedural Place Descriptions*. Abo Akademi University Press: Abo

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)

Walker M.A. (2000) 'Vers un modèle de l'interaction du Centrage avec la structure globale du discours'. *Verbum* n°22 vol.1 (31-58)

Walker M.A., Joshi A. & Prince E. (1998) 'Centering in Naturally Occuring Discourse:An Overview'. In M. Walker, A. Joshi & E. Prince (eds) *Centering Theory of Discourse*. Calendron Press: Oxford (1-28)

Werlich E. (1976) *A text grammar for English*. Quelle & Meyer: Heidelberg

Werth P. (1999) *Text worlds: Representing conceptual space in discourse*. Longman: London

Zerida N., Lucas N. & Crémilleux B. (2006) 'Combinaison de descripteurs linguistiques et de structure pour la fouille d'articles biomédicaux'. Actes du Colloque international 'Discours et Document' (ISDD'06), Juin 2006, Caen, France (69-78)

Annexes

Annexe A. Caractéristiques des textes du corpus

N°	Auteur Titre du ou des article(s)	Nombre de			
		Titres	para.	phr.	mots
	auteurs multiples et souvent anonymes				
Atlas_1	Atlas Transmanche	289	948	2827	79075
	Robert HERIN & Rémi ROUAULT				
Atlas_2	Atlas de la France scolaire : de la maternelle au lycée	176	559	1709	64200
	Pascal BULEON				
Atlas_3	Quarante années d'évolution politique de l'Ouest de la France	66	715	3056	92295
	Atlas_TOTAL	531	2222	7592	235570

	Laurence NARDON				
GeoPo_2	Le contrôle de l'imagerie commerciale : après la campagne d'Afghanistan	16	71	207	5366
GeoPo_3	L'administration Bush et l'espace : Militarisation, gestion et coopération	25	115	363	8933
	May CHARTOUNI-DUBARRY				
GeoPo_4	Le triangle syro-libano-israélien : scénarios de crise	7	40	281	11630
	Volker PERTHES*				
GeoPo_5	Scénarios syriens : processus de paix, changements internes et relations avec le Liban	14	37	233	8829
	Joseph BAHOUT*				
GeoPo_6	Le Liban et le couple syro-libanais dans le processus de paix Horizons incertains	6	31	221	9361
	Eddy FOUGIER				
GeoPo_7	La contestation de la mondialisation : une nouvelle exception française ?	2	10	61	2584
GeoPo_24	Perceptions de la mondialisation en France et aux États-Unis	7	39	154	5865
	François VERGNIOLLE DE CHANTAL[190]				
GeoPo_1	La lutte contre le terrorisme : essai de bilan institutionnel	3	20	136	4173
GeoPo_8	Les élections de mi-mandat aux Etats-Unis (5 novembre 2002)	4	13	161	4542
GeoPo_21	La crise budgétaire des Etats fédérés américains	4	25	213	6588
GeoPo_22	Libertés civiles et lutte anti-terroriste aux Etats-Unis	4	19	179	6072
	David BARAN[191]				
GeoPo_9	Dans l'" après-Saddam ", il y a encore " Saddam "	2	8	122	3413
GeoPo_11	L'adversaire irakien	11	22	232	6218
	Barry R. POSEN				

190 François Vergniolle de Chantal est docteur en Sciences Politiques à l'IEP de Paris et Maître de Conférences en civilisation américaine à l'Université de Bourgogne.

191 David Baran est le pseudonyme d'un journaliste indépendant, ancien consultant en relations internationales pour le Moyen-Orient.

N°	Auteur Titre du ou des article(s)	Nombre de			
		Titres	para.	phr.	mots
GeoPo_10	La maîtrise des espaces, fondement de l'hégémonie militaire des Etats-Unis	9	14	206	6189
	Pierre NOËL[192]				
GeoPo_12	Les Etats-Unis et le pétrole § De Rockefeller à la Guerre du Golfe	9	26	120	3829
GeoPo_14	Les Etats-Unis face à leur dépendance pétrolière	37	100	437	15349
	Thierry DE MONTBRIAL[193]				
GeoPo_13	Perspectives	10	60	397	12934
	Steven C. CLEMONS				
GeoPo_15	La réglementation des lobbies aux Etats-Unis et son impact sur les think tanks spécialisés dans les politiques publiques	7	50	231	10532
	Alan PHILIP				
GeoPo_16	Le lobbying dans l'Union européenne : les intérêts des entreprises concordent-ils avec la politique étrangère ?	15	59	360	13278
	Paul-Henri RAVIER[194]				
GeoPo_17	De Doha à Cancun: les enjeux du cycle de négociations	12	42	168	6032
	Frédérique SACHWALD[195]				
GeoPo_18	Du bon usage de la mondialisation	7	29	136	4321
	Guillaume PARMENTIER				
GeoPo_19	Force, faiblesse, puissance ?	8	28	186	6674
	Barthélémy COURMONT				
GeoPo_20	les pouvoirs de guerre en débat à Washington	27	122	610	25060
	Jean KLEIN				
GeoPo_23	Les chances et la signification d'une politique européenne de sécurité et de défense dans le nouveau conPeopl_international	5	39	206	11315
	Zaki LAÏDI				
GeoPo_25	Mondialisation et démocratie	3	37	238	6134
	Dominique DAVID				
GeoPo_26	La guerre dans le siècle	14	36	183	5026
GeoPo_28	11 septembre : premières leçons stratégiques	10	29	130	4216
	Jacques BELTRAN et Guillaume PARMENTIER				
GeoPo_27	Les États-Unis à l épreuve de la vulnérabilité	6	37	163	6217
	Gaël RABALLAND				
GeoPo_29	Géoéconomie du bassin caspien	24	79	310	8046
	Etienne DE DURAND				

192 Pierre Noël est économiste du pétrole ; Docteur en science politique. Chercheur au Centre français sur les États-Unis à l'Ifri ; chercheur associé au LEPIIEPE, université de Grenoble.

193 Thierry de Montbrial est directeur de l'Ifri, membre de l'Académie des sciences morales et politiques

194 Paul-Henri Ravier est ancien directeur général adjoint de l'OMC. Les opinions exprimées dans cet article n'engagent que la responsabilité de l'auteur.

195 Frédérique Sachwald est responsable des études économiques à l'IFRI.

N°	Auteur Titre du ou des article(s)	Nombre de			
		Titres	para.	phr.	mots
GeoPo_30	Les transformations de l'US Army	40	152	704	36090
	Yves-Marie PÉRÉON[196]				
GeoPo_31	Après Enron. Wall Street et le gouvernement d'entreprise	8	56	291	9501
	Jean-Marie PAUGAM[197]				
GeoPo_32	Pour une relance du cycle du développement : refonder le consensus multilatéral après Cancun	23	87	263	10017
	GeoPo_TOTAL	379	1532	7902	284334

Peopl_1	DON JUAN	12	45	163	6042
Peopl_2	JOYCE (J.) 13	14	44	264	8569
Peopl_3	DOSTOÏEVSKI (F. M.)	15	38	213	5436
Peopl_4	PASCAL (B.)	13	43	228	8597
Peopl_5	BAUDELAIRE (C.)	10	37	152	7770
Peopl_6	CÉSAR	12	30	235	8111
Peopl_7	BALZAC (H. de)	7	59	483	16361
Peopl_8	SHAKESPEARE (W.)	11	60	285	8832
Peopl_9	HUGO (V.)	36	57	328	14669
Peopl_10	MALRAUX (A.)	9	31	223	9627
Peopl_11	LÉONARD DE VINCI	19	53	349	12857
Peopl_12	BOSCH (J.)	16	65	415	9771
Peopl_13	BOTTICELLI (S.)	16	24	123	4502
Peopl_14	APOLLINAIRE (G.)	14	51	146	4660
Peopl_15	BARTÓK (B.)	17	37	266	8094
Peopl_16	MOZART (W. A.)	5	35	150	5004
Peopl_17	BEETHOVEN (L. van)	8	33	159	4490
Peopl_18	BACH (J.-S.)	8	35	313	7889
Peopl_19	PLATON	20	31	441	14904
Peopl_20	DIDEROT (D.)	11	34	217	7100
Peopl_21	HEGEL (G. W. F.)	14	68	364	10523
Peopl_22	LEIBNIZ (G. W.)	12	47	282	14084
Peopl_23	NIETZSCHE (F.)	28	62	289	10234
Peopl_24	HUSSERL (E.)	8	35	160	9114
Peopl_25	SARTRE (J.-P.)	9	31	257	9666

196 Yves-Marie Péréon, Chartered Financial Analyst, est diplômé de l'École supérieure de Commerce de Paris (1989). Depuis 1995, il travaille à New York pour une banque française. Titulaire d'un DEA sur "La France vue par la presse américaine entre 1936 et 1947" (Université de Franche-Comté, 2003), il prépare actuellement une thèse de doctorat sur le même sujet à l'Université Paris I.

197 Jean-Marie Paugam est chercheur à l'IFRI.

N°	Auteur Titre du ou des article(s)	Nombre de			
		Titres	para.	phr.	mots
Peopl_26	NAPOLÉON Ier	6	36	142	3917
Peopl_27	ALEXANDRE LE GRAND	13	68	287	6567
Peopl_28	TCHEKHOV (A. P.)	7	46	209	5586
Peopl_29	LOUIS XIV	15	54	344	10910
Peopl_30	MALLARMÉ (S.)	4	47	237	6826
	Peopl_TOTAL	389	1336	7724	260712

Annexe B. Titres des articles de l'Atlas Transmanche dans le sous-corpus ATLAS

Prévisions démographiques dans la zone transmanche
Les établissements de santé du secteur Dinard - Saint-Malo - Dinan
Les ordonnances d'Avril 1996
Territoires du système scolaire
L'Arc Atlantique : un accord multilatéral de coopération transfrontalière
Contexte d'émergence de la coopération interrégionale : définition
La diversité statutaire des accords de coopération interrégionale
L'heure de l'Union Européenne est aussi l'heure des prérogatives locales accrues
Mouvement, flux de passagers et liaisons aériennes à partir des aéroports français
L'implantation des entreprises britanniques dans les régions littorales françaises de la Manche
Les coutures de l'Europe sont aussi maritimes; des régions frontalières aux régions transfrontalières.
Les régions de la Zone Transmanche
Le trafic maritime passager transmanche en 1996
Synthèse du vieillissement transmanche
Le vieillissement : un terme polysémique
Le trafic passager transmanche
Le trafic roulier transmanche
Diversité des compagnies de ferries
La population des îles Anglo-Normandes
La plaisance
Les jumelages
Pression démographique et foncière
Une place financière d'importance
Le trafic maritime des îles Anglo-Normandes
L'énergie, facteur d'intégration économique
Des structures d'âge contrastées
La pêche
Un trafic transmanche en constante évolution 1990-1998
Une comparaison hiérarchique et théorique de la croissance urbaine de part et d'autre de la Manche (1960-1990)
Les Français en Grande-Bretagne
Importance grandissante du courant Transmanche
Les ports de plaisance du littoral français des mers de la Manche et du Nord
Analyse des relations par les flux routiers entre Cherbourg, Caen, Rouen et Le Havre en 1996
Les liaisons maritimes sur les Iles Anglo-Normandes pendant l'été 1999
Emergence d'un nouveau contexte pour le Transmanche maritime ?

Montée en puissance du Tunnel
La population étrangère côté français
La population britannique côté français
Proximités croissantes. Les sociétés littorales de la Manche, changements : entre facteurs de rapprochement, freins et décalages
Le tourisme en 1998
Les estuaires de la Seine et du Solent : une approche comparée
Les sites protégés par des accords internationaux
La mise en place des systèmes actuels de protection de la nature
Les mesures nationales de protection du littoral
Les mesures de protection des grands ensembles naturels
Les mesures ponctuelles de protection du milieu naturel
Le milieu naturel
South Coast Metropole : Une démographie sous influence économique
South Coast Metropole : La maritimité comme support économique et culturel
South Coast Metropole : Des infrastructures en continuité des liaisons maritimes
South Coast Metropole : Un pôle économique majeur, entre haute technologie, tourisme et activités maritimes
La notion de maritimité
Les évolutions spatio-temporelles des zones portuaires
Renforcement et développement des outils économiques "traditionnels"
Les régions transmanche au dernier recensement de la population française.
Les implications territoriales des accords de juillet 2000 dans le golfe normand-breton
les liaisons aériennes dans les Iles Anglo-normandes en 1999
Valorisation du patrimoine et de la culture maritime
Répartition des activités (1998-1999)

Annexe C. Récurrence lexicale au fil des textes

Les tableaux suivants présentent les noms les plus fréquents dans les différents sous-corpus. Les noms jugés les plus fréquents sont ceux dont la fréquence est supérieure à 1% de tous les noms répertoriés. Nous parlons alors de recouvrement (%R). Parallèlement à ce recouvrement, nous avons mesuré pour chaque nom sa proportion d'apparition en position initial (%_INIT).

Atlas_1 : Atlas Transmanche (20972 noms répertoriés)

NOM	Nb	%init	%R
région	263	35	1,25
zone	259	36	1,23
population	218	39	1,04

Atlas_2 : Atlas de la France scolaire, de la maternelle au lycée (15816 noms répertoriés)

NOM	Nb	%init	%R
élève	459	38	2,90
%	398	18	2,52
enseignement	357	46	2,26
enfant	333	32	2,11
département	312	28	1,97
classe	302	31	1,91
année	298	38	1,88
lycée	271	31	1,71
collège	232	29	1,47
formation	232	29	1,47
an	210	28	1,33
école	209	41	1,32
baccalauréat	187	36	1,18
nombre	160	51	1,01

Atlas_3 : Quarante années d'évolution politique de l'ouest de la France (21694 noms répertoriés)

NOM	Nb	%init	%R
%	522	15	2,41
ouest	516	48	2,38
année	370	44	1,71
droite	343	40	1,58
vote	295	37	1,36
élection	279	42	1,29
département	225	42	1,04
parti	224	40	1,03

Geopo_1 : La lutte contre le terrorisme : essai de bilan institutionnel (1007 noms répertoriés)

NOM	Nb	%init	%R
état	34	0	3,38
terrorisme	21	48	2,09
lutte	19	37	1,89
autorité	19	26	1,89
administration	15	53	1,49
sécurité	15	40	1,49
bush	11	82	1,09
pouvoir	11	45	1,09
l[198]	11	9	1,09
milliard	11	9	1,09

Geopo_2 : Le contrôle de l'imagerie commerciale : après la campagne d'Afghanistan (1479 noms répertoriés)

NOM	Nb	%init	%R
imagerie	54	43	3,65
image	45	27	3,04
résolution	29	28	1,96
Nima	28	0	1,89
contrôle	18	28	1,22
système	18	22	1,22
entreprise	17	59	1,15
administration	17	59	1,15
satellite	17	53	1,15
gouvernement	17	53	1,15
agence	16	50	1,08
mètre	15	27	1,01

198 Dans ce texte, la forme élidée de l'article défini (*l'*) est remplacée 11 fois par la forme numérale 1 suivie d'une apostrophe typographique (erreur survenue lors de l'écriture ou de l'édition du document). Cette forme numérale est alors considérée comme un NomXXNum par le TreeTagger. Par exemple, dans la phrase « *L'administration doit étudier une nouvelle méthode de contrôle de la diffusion de l'imagerie* », « *de l'imagerie* » est étiqueté : Prep, NomXXNum (gouverné par la préposition), Typo, NomFS (sans gouverneur ni dépendances).

Geopo_3 : L'administration Bush et l'espace : Militarisation, gestion et coopération (2540 noms répertoriés)			
NOM	**Nb**	**%init**	**%R**
programme	53	45	2,09
satellite	49	24	1,93
exportation	39	38	1,54
système	37	30	1,46
air	32	53	1,26
espace	32	50	1,26
département	30	33	1,18
force	29	55	1,14
Nasa	26	42	1,02

Geopo_4 : Le triangle syro-libano-israélien : scénarios de crise (2650 noms répertoriés)			
NOM	**Nb**	**%init**	**%R**
paix	67	28	2,53
Syrie	47	23	1,77
Liban	45	56	1,70
Israël	45	18	1,70
sécurité	42	29	1,58
état	40	40	1,51
Golan	40	22	1,51
négociation	39	38	1,47
retrait	35	40	1,32
Liban-sud	34	32	1,28
option	31	35	1,17

Geopo_5 : Scénarios syriens : processus de paix, changements internes et relations avec le Liban (1954 noms répertoriés)			
NOM	**Nb**	**%init**	**%R**
Syrie	92	43	4,71
paix	54	41	2,76
état	34	41	1,74
Israël	34	18	1,74
Liban	32	31	1,64
régime	30	43	1,54
sécurité	27	30	1,38
Damas	27	19	1,38
pays	27	15	1,38
négociation	23	43	1,18
processus	22	23	1,13
pouvoir	20	45	1,02

Geopo_6 : Le Liban et le couple syro-libanais dans le processus de paix Horizons incertains (1952 noms répertoriés)			
NOM	**Nb**	**%init**	**%R**
Liban	70	40	3,59
Syrie	30	33	1,54
paix	28	43	1,43
négociation	26	46	1,33
armée	22	59	1,13
forces	22	32	1,13

Geopo_7 : la contestation de la mondialisation : une nouvelle exception française ? (602 noms répertoriés)			
NOM	**Nb**	**%init**	**%R**
groupe	18	1	2,99
mondialisation	18	0	2,99
contestation	14	4	2,33
France	12	0	1,99
mouvement	9	1	1,50
contestataire	9	0	1,50
Attac	7	0	1,16
campagne	7	0	1,16

Geopo_8 : Les élections de mi-mandat aux États-Unis (5 novembre 2002) (1092 noms répertoriés)			
NOM	**Nb**	**%init**	**%R**
président	29	31	2,66
élection	25	60	2,29
démocrate	16	112	1,47
bush	16	56	1,47
parti	16	38	1,47
%	16	19	1,47
sénat	15	53	1,37
républicain	13	54	1,19
congrès	13	31	1,19
majorité	11	45	1,01

Geopo_9 : Dans l' « après-Saddam », il y a encore « Saddam » (769 noms répertoriés)			
NOM	**Nb**	**%init**	**%R**
régime	25	52	3,25
Irak	13	62	1,69
Saddam	12	83	1,56
pouvoir	12	25	1,56
état	11	9	1,43
tribu	10	70	1,30
arme	10	50	1,30
population	10	40	1,30
Hussein	9	78	1,17

guerre	9	78	1,17
violence	8	62	1,04

Geopo_10 : La maîtrise des espaces, fondement de l'hégémonie militaire des États-Unis (1328 noms répertoriés)

NOM	Nb	%init	%R
États-unis	76	45	5,72
maîtrise	33	76	2,48
forces	30	30	2,26
espace	27	48	2,03
adversaire	23	35	1,73
capacité	18	39	1,36
combat	18	33	1,36
pays	18	22	1,36
défense	16	31	1,20
puissance	14	21	1,05

Geopo_11 : L'adversaire irakien (1431 noms répertoriés)

NOM	Nb	%init	%R
armée	47	47	3,28
régime	35	63	2,45
sécurité	33	48	2,31
Saddam	31	45	2,17
garde	26	42	1,82
Hussein	24	50	1,68
guerre	19	58	1,33
appareil	16	31	1,12

Geopo_12 : Les États-Unis et le pétrole § De Rockefeller à la Guerre du Golfe (1046 noms répertoriés)

NOM	Nb	%init	%R
marché	27	52	2,58
Etats-Unis	17	29	1,63
politique	16	62	1,53
Reagan	15	80	1,43
prix	15	27	1,43
année	13	77	1,24
importation	13	46	1,24
pétrole	13	46	1,24

Geopo_13 : Perspectives (2689 noms répertoriés)

NOM	Nb	%init	%R
Etats-Unis	45	0	1,67
pays	32	22	1,19
politique	28	61	1,04
état	27	111	1,00
11 septembre	27	56	1,00

Geopo_14 : Les États-Unis face à leur dépendance pétrolière (3936 noms répertoriés)

NOM	Nb	%init	%R
marché	110	29	2,79
prix	88	36	2,24
pétrole	81	44	2,06
importation	69	35	1,75
Etats-Unis	64	14	1,63
production	62	40	1,58
demande	53	36	1,35
politique	40	57	1,02
année	40	30	1,02

Geopo_15 : La réglementation des lobbies aux États-Unis et son impact sur les think tanks spécialisés dans les politiques publiques (2517 noms répertoriés)

NOM	Nb	%init	%R
lobbyiste	42	40	1,67
think	41	46	1,63
tank	41	46	1,63
organisation	41	32	1,63
intérêt	37	27	1,47
but	31	32	1,23
entreprise	27	30	1,07

Geopo_16 : Le lobbying dans l'Union européenne : les intérêts des entreprises concordent-ils avec la politique étrangère ? (3117 noms répertoriés)

NOM	Nb	%_init	%R
union	93	44	2,98
intérêt	65	18	2,09
commission	57	47	1,83
état	57	2	1,83
lobbying	51	51	1,64
membre	51	14	1,64
groupe	45	27	1,44
banane	45	24	1,44
conseil	41	37	1,32
décision	41	24	1,32
gouvernement	36	19	1,15
marché	35	3	1,12
question	34	32	1,09
organisation	33	27	1,06

Geopo_17 : De Doha à Cancun: les enjeux du cycle de négociations (1346 noms répertoriés)

NOM	Nb	%init	%R
cycle	45	31	3,34

négociation	36	42	2,67
sujet	26	38	1,93
Doha	26	35	1,93
pays	25	20	1,86
accord	20	30	1,49
Ped	20	0	1,49
question	17	35	1,26
Omc	16	0	1,19

Geopo_18 : Du bon usage de la mondialisation (1033 noms répertoriés)

NOM	Nb	%init	%R
pays	63	30	6,10
mondialisation	34	56	3,29
ouverture	28	39	2,71
marché	28	29	2,71
échange	19	21	1,84
année	16	69	1,55
évolution	14	64	1,36
développement	14	43	1,36
intégration	13	46	1,26
économie	13	31	1,26
inégalité	12	58	1,16
pauvreté	12	42	1,16
processus	11	55	1,06
gouvernement	11	36	1,06
écart	11	18	1,06
institution	11	9	1,06
Ped	11	0	1,06

Geopo_19 : Force, faiblesse, puissance ? (1344 noms répertoriés)

NOM	Nb	%init	%R
Etats-Unis	42	0	3,13
puissance	26	23	1,93
Europe	25	32	1,86
européen	23	48	1,71
moyen	21	14	1,56
Kagan	20	50	1,49
état	19	79	1,41
pays	16	12	1,19
américain	15	27	1,12
force	14	36	1,04
politique	14	29	1,04
partenaire	14	14	1,04

Geopo_20 : les pouvoirs de guerre en debat à Washington (5885 noms répertoriés)

NOM	Nb	%init	%R
congrès	176	28	2,99
président	139	39	2,36
guerre	92	39	1,56
administration	89	28	1,51
pouvoir	88	25	1,50
parlementaire	78	51	1,33
Etats-Unis	77	26	1,31
politique	64	36	1,09

Geopo_21 : La crise budgétaire des Etats fédérés américains (1488 noms répertoriés)

NOM	Nb	%init	%R
état	153	0	10,28
impôt	48	29	3,23
revenu	33	24	2,22
année	27	19	1,81
crise	26	50	1,75
milliard	22	5	1,48
dollar	21	5	1,41
situation	17	41	1,14
budget	17	12	1,14
problème	16	44	1,08
dépense	16	31	1,08
programme	16	6	1,08

Geopo_22 : Libertés civiles et lutte anti-terroriste aux Etats-Unis (1446 noms répertoriés)

NOM	Nb	%init	%R
cour	24	50	1,66
mesure	19	37	1,31
anti	19	32	1,31
lutte	18	28	1,24
état	17	0	1,18

Geopo_23 : Les chances et la signification d'une politique européenne de sécurité et de défense dans le nouveau contexte international (2737 noms répertoriés)

NOM	Nb	%init	%R
défense	60	27	2,19
Ue	57	0	2,08
Otan	42	0	1,53
Etats-Unis	36	17	1,32
Europe	33	24	1,21
européen	31	61	1,13
sécurité	30	20	1,10
capacité	29	17	1,06
état	29	0	1,06
alliance	28	11	1,02

Geopo_24 : Perceptions de la mondialisation en France et aux Etats-Unis (1338 noms répertoriés)			
NOM	**Nb**	**%init**	**%R**
mondialisation	59	42	4,41
%	51	27	3,81
sondage	30	70	2,24
opinion	28	43	2,09
personne	26	50	1,94
entreprise	23	17	1,72
Etats-unis	22	0	1,64
France	21	71	1,57
pays	21	14	1,57
emploi	19	11	1,42
sentiment	17	47	1,27
perception	16	44	1,20
conséquence	14	0	1,05

Geopo_25 : Mondialisation et démocratie (1238 noms répertoriés)			
NOM	**Nb**	**%init**	**%R**
démocratie	114	36	9,21
élection	33	21	2,67
mondialisation	30	50	2,42
culture	30	40	2,42
temps	25	52	2,02
procédure	22	50	1,78
pays	17	41	1,37
espace	16	56	1,29
droit	14	29	1,13

Geopo_26 : La guerre dans le siècle (1219 noms répertoriés)			
NOM	**Nb**	**%init**	**%R**
guerre	89	42	7,30
siècle	21	52	1,72
puissance	21	29	1,72
conflit	20	40	1,64
état	13	69	1,07
sécurité	13	38	1,07
arme	13	38	1,07

Geopo_27 : Les États-Unis à l épreuve de la vulnérabilité (1340 noms répertoriés)			
NOM	**Nb**	**%init**	**%R**
Etats-Unis	45	2	3,36
attentat	25	56	1,87
11 septembre	23	61	1,72
défense	17	29	1,27
politique	16	75	1,19
américain	15	80	1,12
conséquence	15	40	1,12
pays	15	33	1,12
relation	14	50	1,04
intérêt	14	7	1,04

Geopo_28 : 11 septembre : premières leçons stratégiques (900 noms répertoriés)			
NOM	**Nb**	**%init**	**%R**
défense	20	35	2,22
monde	17	59	1,89
sécurité	16	50	1,78
moyen	16	25	1,78
système	14	29	1,56
stratégie	12	42	1,33
forces	10	60	1,11
guerre	10	30	1,11

Geopo_29 : Géoéconomie du bassin caspien (2048 noms répertoriés)			
NOM	**Nb**	**%init**	**%R**
pays	90	43	4,39
région	54	43	2,64
économie	49	59	2,39
état	47	2	2,29
%	43	5	2,10
Kazakhstan	33	42	1,61
Russie	30	47	1,46
Asie	28	39	1,37
Turkménistan	25	52	1,22
système	25	52	1,22
Azerbaïdjan	24	58	1,17
bassin	22	36	1,07
Caspien	21	38	1,03
hydrocarbure	21	24	1,03
secteur	21	14	1,03

Geopo_30 : Les transformations de l'US Army (8419 noms répertoriés)			
NOM	**Nb**	**%init**	**%R**
Army	195	48	2,32
force	160	45	1,90
guerre	98	39	1,16
armée	90	42	1,07
unité	90	19	1,07

Geopo_31 : Après Enron. Wall Street et le gouvernement d'entreprise (2371 noms répertoriés)			
NOM	**Nb**	**%init**	**%R**
société	44	23	1,86
Enron	40	45	1,69
audit	38	11	1,60

entreprise	33	36	1,39
loi	32	62	1,35

Geopo_32 : Pour une relance du cycle du développement : refonder le consensus multilatéral après Cancun (2642 noms répertoriés)

NOM	Nb	%init	%R
pays	92	23	3,48
Omc	61	0	2,31
Cancun	55	36	2,08
négociation	43	16	1,63
développement	42	26	1,59
accord	29	28	1,10
libéralisation	28	18	1,06
sujet	28	14	1,06

Peopl_1 : Don Juan (1417 noms répertoriés)

NOM	Nb	%init	%R
Don	76	33	5,36
Juan	62	34	4,38

Peopl_2 : Joyce J.(1915 noms répertoriés)

NOM	Nb	%init	%R
Joyce	70	16	3,66
Ulysse	21	19	1,10

Peopl_3 : Dostoïevski F.M. (1269 noms répertoriés)

NOM	Nb	%init	%R
Dostoïevski	41	20	3,23
homme	24	17	1,89
vie	16	31	1,26
Christ	14	29	1,10
Russie	14	21	1,10

Peopl_4 : Pascal B. (1977 noms répertoriés)

NOM	Nb	%init	%R
Pascal	108	28	5,46

Peopl_5 : Baudelaire C. (1676 noms répertoriés)

NOM	Nb	%init	%R
Baudelaire	62	15	3,70
art	48	17	2,86
poésie	33	18	1,97
poète	29	21	1,73
forme	21	10	1,25
conscience	19	16	1,13
langage	17	6	1,01

Peopl_6 : César J. (2016 noms répertoriés)

NOM	Nb	%init	%R
César	95	23	4,71
Pompée	40	32	1,98
guerre	39	26	1,93

Peopl_7 : Balzac H. de (3791 noms répertoriés)

NOM	Nb	%init	%R
Balzac	143	32	3,77
oeuvre	62	31	1,64
roman	53	25	1,40
vie	44	18	1,16

Peopl_8 : Shakespeare W. (2158 noms répertoriés)

NOM	Nb	%init	%R
Shakespeare	45	27	2,09
pièce	25	36	1,16
Henry	24	12	1,11

Peopl_9 : Hugo V. (3480 noms répertoriés)

NOM	Nb	%init	%R
Hugo	96	29	2,76
oeuvre	37	24	1,06
histoire	35	34	1,01

Peopl_10 : Malraux A. (2145 noms répertoriés)

NOM	Nb	%init	%R
Malraux	92	38	4,29
art	44	23	2,05
homme	29	38	1,35
roman	28	18	1,31
livre	24	58	1,12

Peopl_11 : Léonard de Vinci (3272 noms répertoriés)

NOM	Nb	%init	%R
léonard	67	45	2,05
fo	38	8	1,16
dessin	33	30	1,01

Peopl_12 : Bosch J. (2338 noms répertoriés)

NOM	Nb	%init	%R
Bosch	68	38	2,91
monde	51	49	2,18
oeuvre	50	48	2,14
regard	25	68	1,07
image	24	50	1,03

Peopl_13 : Botticelli S. (1185 noms répertoriés)

NOM	Nb	%init	%R
Botticelli	47	34	3,97
composition	19	37	1,60
peintre	18	33	1,52
tableau	18	33	1,52
Médicis	14	36	1,18

oeuvre	13	38	1,10
Florence	12	33	1,01

Peopl_14 : Apollinaire G. (1033 noms répertoriés)

NOM	Nb	%init	%R
Apollinaire	17	18	1,65
oeuvre	13	54	1,26
vie	13	15	1,26
poème	12	33	1,16
poésie	12	17	1,16
revue	11	36	1,06
art	11	18	1,06
amour	11	18	1,06

Peopl_15 : Bartók B. (2002 noms répertoriés)

NOM	Nb	%init	%R
Bartók	76	29	3,80
musique	51	20	2,55
oeuvre	28	32	1,40
langage	25	40	1,25
piano	24	17	1,20
mélodie	23	4	1,15
année	21	67	1,05

Peopl_16 : Mozart W.A. (1082 noms répertoriés)

NOM	Nb	%init	%R
Mozart	32	25	2,96
K.	21	10	1,94
oeuvre	18	33	1,66
musique	13	46	1,20
langage	13	8	1,20
an	12	42	1,11
année	12	42	1,11
mort	12	33	1,11
maître	11	0	1,02

Peopl_17 : Beethoven L. van (889 noms répertoriés)

NOM	Nb	%init	%R
Beethoven	62	32	6,97
oeuvre	35	40	3,94
musique	26	27	2,92
vie	15	27	1,69
musicien	14	36	1,57
symphonie	11	18	1,24

Peopl_18 : Bach J.-S. (1457 noms répertoriés)

NOM	Nb	%init	%R
Bach	81	31	5,56
musique	46	30	3,16
art	20	55	1,37
musicien	17	53	1,17
oeuvre	15	40	1,03

Peopl_19 : Platon (2885 noms répertoriés)

NOM	Nb	%init	%R
Platon	83	36	2,88
âme	52	27	1,80
chose	39	26	1,35
forme	38	34	1,32
dialogue	36	31	1,25
Socrate	34	18	1,18
être	33	85	1,14
pensée	33	39	1,14
savoir	33	36	1,14

Peopl_20 : Diderot D. (1467 noms répertoriés)

NOM	Nb	%init	%R
Diderot	30	23	2,04
nature	22	27	1,50
homme	21	29	1,43
art	16	44	1,09
oeuvre	16	19	1,09
philosophe	15	13	1,02

Peopl_21 : Hegel G.W.F. (2079 noms répertoriés)

NOM	Nb	%init	%R
Hegel	65	38	3,13
logique	30	37	1,44
monde	30	23	1,44
langage	27	44	1,30
liberté	27	22	1,30
dieu	25	32	1,20
philosophie	24	25	1,15
Esprit	24	17	1,15
absolu	23	43	1,11
état	23	39	1,11
raison	22	45	1,06
histoire	22	32	1,06
principe	22	23	1,06
conscience	21	43	1,01

Peopl_22 : Leibniz G.W. (2967 noms répertoriés)

NOM	Nb	%init	%R
Leibniz	85	16	2,86
dieu	45	13	1,52
principe	38	8	1,28

monade	37	30	1,25
monde	34	12	1,15
chose	31	6	1,04
nombre	30	17	1,01

Peopl_23 : Nietzsche F. (2341 noms répertoriés)

NOM	Nb	%init	%R
Nietzsche	95	27	4,06
puissance	47	36	2,01
volonté	46	30	1,96
être	41	56	1,75
monde	32	12	1,37
vérité	27	48	1,15
vie	27	33	1,15

Peopl_24 : Husserl E. (1679 noms répertoriés)

NOM	Nb	%init	%R
Husserl	53	23	3,16
humanité	47	15	2,80
philosophie	34	21	2,03
phénoménologie	25	36	1,49
question	23	30	1,37
science	21	19	1,25
logique	20	35	1,19
sens	18	11	1,07
métaphysique	17	18	1,01

Peopl_25 : Sartre J.-P. (1921 noms répertoriés)

NOM	Nb	%init	%R
Sartre	72	38	3,75
liberté	29	41	1,51
conscience	28	18	1,46
néant	22	55	1,15

Peopl_26 : Napoléon 1er (999 noms répertoriés)

NOM	Nb	%init	%R
Napoléon	52	21	5,21
empereur	15	27	1,50
légende	14	71	1,40
France	13	31	1,30
homme	10	10	1,00

Peopl_27 : Alexandre Le Grand (1629 noms répertoriés)

NOM	Nb	%init	%R
Alexandre	99	16	6,08
roi	29	28	1,78
empire	22	45	1,35
pouvoir	17	18	1,04

Peopl_28 : Tchekhov A.P. (1245 noms répertoriés)

NOM	Nb	%init	%R
Tchekhov	63	22	5,06
vie	22	41	1,77
lettre	14	14	1,12
année	13	77	1,04
Moscou	13	15	1,04

Peopl_29 : Louis XIV (2772 noms répertoriés)

NOM	Nb	%init	%R
Louis	60	52	2,16
xiv	54	0	1,95
roi	53	40	1,91
guerre	46	41	1,66
France	44	52	1,59

Peopl_30 : Mallarmé S. (1341 noms répertoriés)

NOM	Nb	%init	%R
Mallarmé	44	14	3,28
poème	42	24	3,13
poète	24	29	1,79
mot	22	27	1,64
musique	17	41	1,27
sens	14	29	1,04

ANNEXE D. REPÉRAGE DES TITRES DE SECTION

D.1. Trois caractéristiques formelles pertinentes

Trois caractéristiques formelles ont été utilisées pour déterminer si ce que l'on considère comme un paragraphe (délimité par deux retours à la ligne) est ou non un titre de section. Pour définir ces caractéristiques, trois indices ont été repérés pour les 1299 titres de section des trois corpus : (i) le fait d'être numéroté, (ii) la longueur du titre et (iii) l'absence ou présence d'une ponctuation finale.

D.1.1. Présence d'un système de numérotation

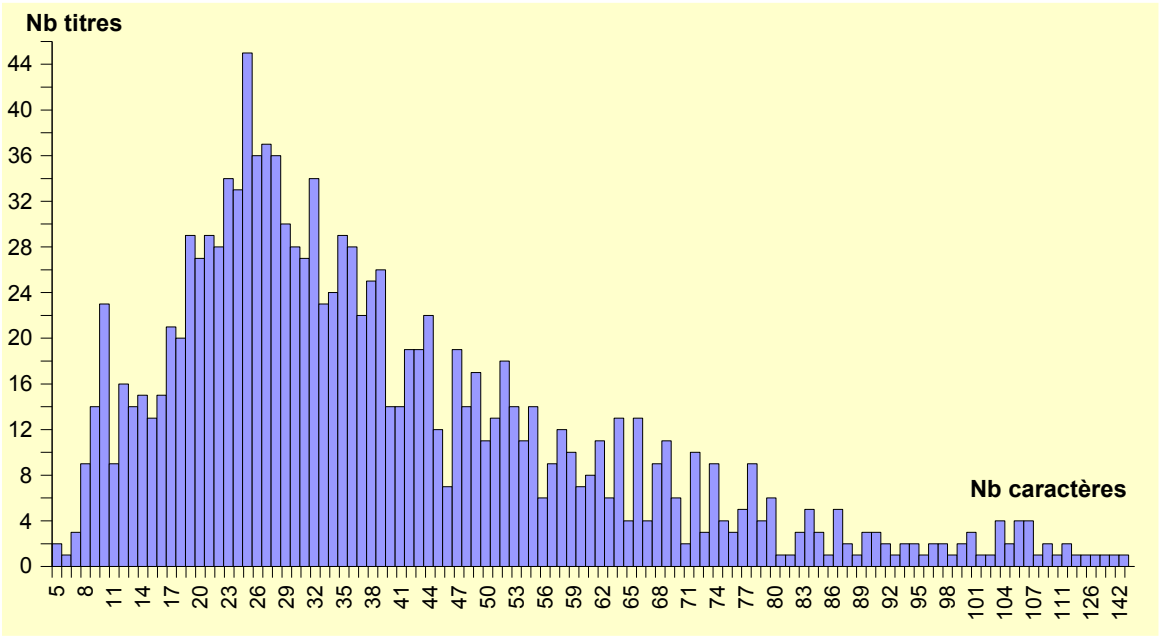
Dans notre corpus, 54 % des titres présentent une numérotation. L'indice de numérotation des titres est donc pertinent dans la moitié des cas. Tout paragraphe commençant par une numérotation peut-être potentiellement un titre. Mais il peut également faire office d'item d'énumération. Il faut donc procéder à une recherche contextuelle qui consiste à considérer les paragraphes précédent et/ou suivant. Si l'un de ceux-ci est numéroté selon le même système et au même niveau (chiffres arabes/latins ou lettres minuscules/majuscules qui se suivent logiquement), le paragraphe est un item. Si par contre, le système de numérotation change et/ou ne suit pas une suite logique avec le paragraphe précédent, le paragraphe est étiqueté « titre ».

Ce premier indice dans le repérage des titres est selon nous le plus fiable. Cependant, il ne recouvre que 50 % des cas et, de ce fait, doit être complété. Un titre, numéroté ou non, est un paragraphe qui présente plusieurs particularités : une complétude syntaxique facultative, une taille relativement courte et une ponctuation finale facultative. Bien que la première caractéristique reste très pertinente, nous nous sommes principalement intéressée aux deux derniers particularismes (l'optionnalité de complétude syntaxique peut d'ailleurs être mise en relation avec celle de la ponctuation finale).

D.1.2. Longueur des titres

Les graphiques 1 et 2 présentent la longueur en nombre de caractères et en nombre de mots des 1299 titres de section des trois corpus d'étude. Ils montrent clairement que la majorité des titres présentent une longueur de moins de 100 caractères, avec un pic autour des 20 caractères ; et de moins de 20 mots, avec un pic vers les 6-7 mots.

Dans les cas de « paragraphe » numéroté, l'indice de longueur maximale d'un titre de section est élargie à 250 caractères. Dans les autres cas, cette longueur est réduite à 150 caractères, voire plus selon qu'il y a ou non ponctuation finale.



Graphique 1: Longueur des titres de section en nombre de caractères

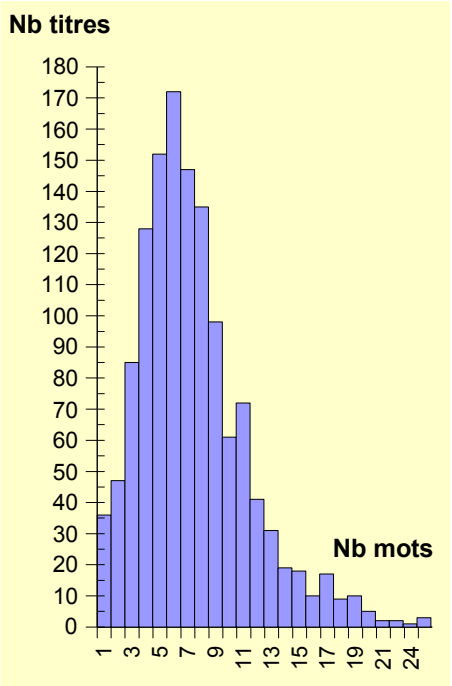
Pour l'instant, nos indices sont encore trop généraux pour que seuls les titres de sections soient repérés. En effet, de nombreux paragraphes non numérotés présentent de telles caractéristiques, comme l'illustrent les exemples qui montrent des paragraphes de type texte (et non titre) comportant moins de 150 caractères (ces paragraphes sont surlignés). Nous y avons également fait figurer les titres de sections environnants afin de donner une idée de différents types de titre de sections :

(1) **13. LE *TRAFIC MARITIME PASSAGER TRANSMANCHE EN 1996*** [titre niveau 1]
La structure du trafic passager transmanche ne se modifie pas. [...]
Les ports les plus touchés par la concurrence du Tunnel sous la Manche sont ceux de Boulogne et de Dieppe. [...]
Les résultats de 1997 devraient conforter cette structuration du trafic.[ATLAS_1]

(2) ***Formations et métiers*** [titre niveau 3]
Les élèves du second cycle professionnel se partagent à peu près par moitié entre les métiers de l'industrie et ceux du tertiaire.
Dans les métiers du secteur secondaire, un élève sur trois suit [...] [ATLAS_2]

(3) *La Maison Blanche rejoint Kevin O'Connell dans l'idée que [...]*
À un stade très peu abouti de la réflexion des responsables, quelques éléments concrets sont mentionnés pour une nouvelle approche du problème.
Préparer la réponse en cas de diffusion d'imagerie métrique [titre niveau 3] *[GEOPO_2]*

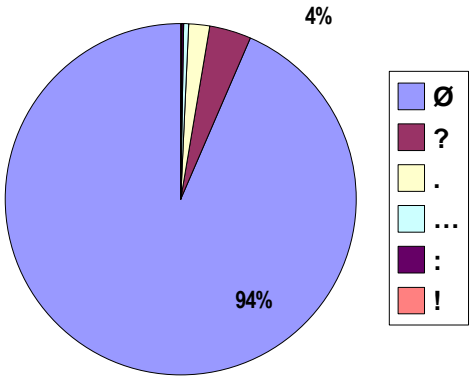
(4) Simples répliques *ou tirades construites, ils annoncent Le Théâtre en liberté [...]*
Enfin, il écrit son dernier drame, le plus puissant, le plus touffu, rassemblant l'essentiel de ses préoccupations : Torquemada (1869).
15) LE DESSINATEUR [titre niveau 1] *[PEOPL_9]*



Graphique 2: Longueur des titres de section en nombre de mots

D.1.3. Absence de ponctuation finale

Notre dernier indice relève de la présence ou de l'absence de ponctuatior en fin de titre (Ø). Le graphique 3 montre que très peu de titres de section prés une ponctuation finale, et dans les cas de ponctuation finale, c'est le {?} qui d Cette information nous permet de faire deux traitements pour le repérage des titr s'il n'y a pas de ponctuation finale, on peut considérer comme étant un tit « paragraphe » de moins de 150 caractères, comme défini précédemment ; (2) ponctuation finale, cette longueur est réduite à 80 caractères s'il y a un point fin à 117 caractères qu'il y a une autre ponctuation finale { ?:…}[199]. Ces de décisions ont été prises par rapport aux titres étudiés dans le corpus et sont à vé sans doute à adapter aux corpus.



Graphique 3 : Répartition des ponctuations finales pour les titres de section

D.2. Programme de repérage

Pour le programme de repérage, les unités auxquelles est attribué le statut de titre de section sont : des paragraphes de longueur inférieure à 250 caractères répondant à une des cinq conditions suivantes :

1. il y a mention explicite qu'il s'agit d'un titre (ex. TITRE :, <TITRE>, etc.)
2. le premier mot est un label appartenant à la liste {introduction, conclusion, chapitre, partie, section}
3. il y a un système de numérotation et l'unité n'a pas été étiquetée comme étant un item d'énumération. Ce système de numérotation peut être :
   - un label suivi d'une numérotation tel que {partie x, chapitre x, section x, paragraphe x, § x} où x appartient à un des systèmes de numérotation présentés ci-dessous, ou
   - une numérotation appartenant à la liste {1, 2, 3 ; I, II, III ; A, B, C ; i, ii, iii ; a, b, c}. Chaque numéro peut être suivi d'un des signes typographiques {.) - /} précédé ou non d'un espace. Chaque numéro peut-être suivi d'un ou de plusieurs autre(s) système(s) de numérotation, pouvant être suivi(s) d'un des signes typographiques {.) - /} précédée ou non d'un espace.
4. le paragraphe mesure moins de 177 caractères et
   - il n'y a pas de ponctuation finale, ou
   - il y a une ponctuation finale appartenant à la liste { ? … ! } pouvant être suivie de guillemets fermantes ou d'une astérisque
5. le paragraphe mesure moins de 80 caractères et présente une ponctuation finale appartenant à la liste {. : } pouvant être suivie de guillemets fermantes ou d'une astérisque

D.3. Validation du programme de repérage

Une vérification manuelle de cette détermination nous permet de déterminer le niveau des titres de sections en nous basant sur les documents d'origine pour lesquels la mise en forme matérielle a été conservée. Nous pouvons mesurer les taux de précision et de rappel de ce programme qui sont bien satisfaisants.

Précision	Rappel
99,8	96

Nous remarquons que ces résultats montrent des variations selon les corpus. Nous retrouvons là l'idée que, selon le type de texte, la mise en titre change.

Corpus	Nbre de titres	Précision	Rappel
atlas	534	99,8	94
geopo	411	100	99,75
peopl	419	99,5	95

199 Pour les titres présentant la ponctuation finale {:}, la recherche par rapport au contexte effectuée au niveau de l'indice de numérotation permet de valider ou invalider le fait que l'on a affaire à un titre ou à une amorce d'énumération.

Annexe E. SP ({*à, dans, en, sur, depuis, au cours de, lors de*} + SN) les plus fréquents et apparaissant souvent en position initiale

Les tableaux suivants indiquent pour chaque sous-corpus les SP de type Prep + SN où Prep = {à, dans, en, sur, depuis, au cours de, lors de} qui présentent plus de 10 occurrences (Occ) et dont la proportion à apparaître en position initiale (%_Init) dépasse les 25%. Les lignes coloriées correspondent aux SP types exprimant de façon certaine un temps ou un lieu.

ATLAS

Prep	Det	Tête	*%_Init*	Occ
dans	0	nombre	*0,67*	12
à	Det	inverse	*0,60*	30
dans	Det	situation	*0,56*	16
dans	Det	France	*0,55*	22
dans	Det	Finistère	*0,47*	19
depuis	Det	milieu	*0,46*	13
à	Det	métier	*0,45*	11
dans	Det	est	*0,45*	11
en	0	terme	*0,45*	29
à	Det	rentrée	*0,44*	34
dans	Det	enseignement	*0,43*	42
dans	Det	contexte	*0,42*	59
en	0	île-de-France	*0,42*	19
dans	Det	primaire	*0,42*	12
au cours de	Det	année	*0,41*	37
en	0	Basse-Normandie	*0,40*	52
à	Det	terme	*0,40*	20
en	0	Vendée	*0,38*	13
en	0	milieu	*0,38*	13
dans	Det	cas	*0,38*	21
dans	Det	pays	*0,37*	27
en	0	Lorraine	*0,36*	11
à	0	Paris	*0,36*	36
à	Det	échelle	*0,36*	25
à	Det	regard	*0,36*	14
en	0	Bretagne	*0,35*	100
dans	Det	région	*0,35*	132
à	Det	âge	*0,35*	23
dans	Det	nord	*0,33*	39
dans	0	département	*0,33*	24
dans	Det	mouvement	*0,33*	15
en	0	Normandie	*0,33*	15

Prep	Det	Tête	*%_Init*	Occ
dans	Det	orne	*0,32*	22
dans	Det	système	*0,31*	13
dans	Det	midi	*0,31*	13
en	0	pays	*0,31*	13
dans	Det	Manche	*0,30*	23
dans	Det	ensemble	*0,30*	60
depuis	Det	année	*0,29*	24
en	0	1900	*0,29*	750
dans	Det	circonscription	*0,28*	32
dans	Det	centre	*0,27*	11
en	0	Mayenne	*0,27*	11
depuis	Det	fin	*0,27*	11
au cours de	Det	décennie	*0,27*	11
à	Det	évolution	*0,27*	11
depuis	0	1900	*0,27*	66
à	Det	changement	*0,27*	11
à	Det	population	*0,27*	11
dans	Det	degré	*0,27*	11
à	Det	contraire	*0,27*	37
en	0	France	*0,27*	111
dans	Det	scrutin	*0,27*	15
à	Det	classe	*0,27*	15
dans	Det	moitié	*0,27*	15
sur	0	quatre	*0,27*	15
dans	Det	département	*0,26*	155
dans	Det	établissement	*0,26*	27
sur	0	trois	*0,25*	16
en	0	Europe	*0,25*	12
à	Det	lycée	*0,25*	28
en	0	croissance	*0,25*	12
depuis	Det	début	*0,25*	20
à	0	Guernesey	*0,25*	12

Prep	Det	Tête	%_Init	Occ
à	0	1900	*0,25*	28
en	0	Grande-Bretagne	*0,25*	12
Nb total de types :		66	Nb total d'Occ :	2580

GEOPO

Prep	Det	Tête	%_Init	Occ
dans	Det	discours	*0,75*	12
à	Det	inverse	*0,74*	27
dans	Det	condition	*0,71*	35
à	0	terme	*0,64*	11
à	Det	élection	*0,57*	14
dans	Det	cas	*0,56*	45
à	Det	terme	*0,55*	20
depuis	Det	11 septembre	*0,55*	11
sur	Det	point	*0,54*	26
dans	Det	rapport	*0,50*	12
dans	Det	perspective	*0,50*	20
au cours de	Det	année	*0,50*	16
à	Det	sud	*0,47*	15
dans	Det	contexte	*0,47*	60
à	0	distance	*0,46*	13
à	Det	négociation	*0,46*	13
dans	Det	ensemble	*0,45*	11
sur	0	lequel	*0,44*	16
à	Det	stade	*0,43*	14
à	Det	intervention	*0,42*	12
dans	Det	lutte	*0,38*	13
depuis	Det	fin	*0,38*	24
à	Det	plan	*0,38*	24
à	Det	majorité	*0,36*	11
dans	Det	société	*0,35*	17
à	Det	printemps	*0,35*	17
sur	Det	plan	*0,34*	38
à	Det	oeil	*0,33*	12
dans	Det	Balkans	*0,33*	18
en	0	transition	*0,33*	12
en	0	Asie	*0,32*	22
dans	Det	pays	*0,32*	79
dans	Det	conflit	*0,31*	13
dans	Det	situation	*0,30*	33
en	0	1900	*0,30*	358
à	Det	contraire	*0,30*	44
en	0	Bosnie	*0,29*	24
depuis	Det	année	*0,28*	25
à	Det	niveau	*0,28*	76
dans	Det	année	*0,27*	55
dans	Det	camp	*0,27*	11
à	0	Bruxelles	*0,27*	33
dans	Det	décision	*0,27*	11
à	Det	moment	*0,27*	22
à	Det	programme	*0,27*	15
en	0	France	*0,26*	31
en	0	Afghanistan	*0,26*	39
à	Det	combat	*0,25*	16
sur	Det	avenir	*0,25*	12
à	Det	Liban-sud	*0,25*	24
en	0	Syrie	*0,25*	16
depuis	0	1900	*0,25*	69
Nb total de types :		52	Nb total d'Occ :	1617

PEOPL

Prep	Det	Tête	%_Init	Occ
à	Det	inverse	*0,55*	11
dans	Det	domaine	*0,50*	18
depuis	0	1900	*0,46*	13
à	0	côté	*0,39*	18
sur	Det	point	*0,38*	13
sur	Det	art	*0,38*	13
en	0	prose	*0,37*	19
en	0	Espagne	*0,33*	12
à	Det	mort	*0,33*	21
à	Det	époque	*0,32*	19
en	0	1900	*0,31*	512
à	Det	bibliothèque	*0,31*	16
en	0	France	*0,31*	26
à	Det	contraire	*0,31*	59
dans	Det	année	*0,29*	17
à	0	lequel	*0,28*	18
à	Det	homme	*0,27*	11
dans	Det	unité	*0,27*	11
à	Det	titre	*0,27*	11
dans	0	lequel	*0,27*	22
à	0	Paris	*0,27*	49
à	Det	travail	*0,25*	12
dans	Det	art	*0,25*	12
Nb total de type :		23	Nb total d'Occ :	933

Annexe F. Collocations rares en initiale

Le tableau suivant présente les 62 collocations rares en position initiale *i.e.* qui représentent moins de 1% des phrases du corpus

INIT1	Tt	Ts	Nb Phrases
APPO	.	Sujet Inversé	1
ARGU	SNind	.	1
TEXT	.	autre	1
TOPI	.	clivée	1
autre	.	existentielle	1
autre	PRO demo	.	1
autre	SNdem_R		1
autre	SNind_R		1
autre	SNdem		1
APPO	.	clivée	2
TEXT	.	topicalisation	2
autre	.	impersonelle	2
autre	SNposs	.	2
autre	NP_R	.	2
autre	SNdem	.	2
APPO	.	existentielle	3
ARGU	SNind_R	.	3
TEXT	PRO demo	.	3
autre	.	clivée	3
autre	.	autre	3
autre	autre_R	.	3
autre	NP	.	3
APPO	PRO demo	.	4
MODA	.	Sujet Inversé	4
MODA	PRO demo	.	4
TEXT	.	Sujet Inversé	4
autre	.	Commentaire	4
autre	SNind		4
TEXT	NP_R	.	5
TEXT	.	existentielle	7
TEXT	SNposs	.	7
TEXT	SNind_R	.	7

INIT1	Tt	Ts	Nb Phrases
APPO	SNind_R	.	8
MODA	SNposs	.	9
TEXT	.	clivée	10
TEXT	SNdem	.	10
TEXT	NP	.	10
MODA	.	existentielle	11
MODA	.	autre	11
MODA	SNdem	.	11
TEXT	.	Impersonelle	11
TEXT	autre_R	.	11
TEXT	SNind	.	11
APPO	SNposs	.	12
TEXT	.	Commentaire	12
TEXT	SNdem_R	.	12
autre	PRO3	.	12
TEXT	autre	.	13
APPO	.	autre	14
MODA	autre_R	.	14
MODA	autre	.	14
autre	SNdef_R	.	14
CIRC	.	topicalisation	16
MODA	NP_R	.	17
APPO	.	Impersonelle	18
APPO	autre_R	.	18
APPO	SNind	.	18
MODA	NP	.	18
MODA	SNind_R	.	19
autre	SNdef	.	20
MODA	.	clivée	21
MODA	SNdem_R	.	21
autre	autre		23

Annexe G. Éiquettes générées par Syntex

La liste ci-dessous est une sorte de glossaire présentant les différentes catégories syntaxiques fournies par l'outil Syntex (catégories « syntexiques »). Ce glossaire est particulièrement utile à la bonne lecture de l'annexe H.

A

Adj = adjectif
Adv = adverbe
AdvGP = locution adverbiale

C

CCoord = conjonction de coordination
CSub = conjonction de subordination et dans certains cas, construction impersonnelle

N

NomFP = nom commun féminin pluriel
NomPr = nom propre
NomXXTitre = nom pour qualifier le statut d'une personne (Monsieur, Lieutenant, etc.)
NomXXDate = nom de localisation temporelle (1900, lundi, mars, siècle, etc.)
NomXXMes = nom de mesure (kilomètre, gramme, etc.)
NomXXNum = nom de chiffre

P

PPa = participe passé
PPr = participe présent
Prep = préposition
Pro = pronom
ProRel = pronom relatif

V

VCONJ = verbe conjugué
VINF = verbe infinitif

Annexe H. Descriptif simplifié du programme d'annotation automatique

Le programme décrit dans cette annexe s'applique à toutes les phrases d'un corpus, qu'elles soient ou non des phrases d'un point de vue syntaxique. De ce fait, les titres et les items d'énumération subissent le même programme que les phrases canoniques.

H.1. Remarques utiles à la compréhension de ce qui suit

- Nous référons aux éléments constitutifs de phrase en cours d'analyse par la catégorie syntaxique qui leur est attribuée selon Syntex et/ou leur lemme et/ou leur occurrence. Les différentes catégories 'syntexiques' sont donnés en annexe G.
- Le code utilisé est le suivant : <catégorie Syntex[lemme] « occurrence »>.
- Parfois, nous cherchons tous les éléments d'une telle catégorie exceptés certains lemmes ou occurrences spécifiques. Cela est noté : <catégorie Syntex![[lemme1],[lemme2]]>, ce qui indique que nous cherchons les éléments appartenant à la catégorie désignés sauf si le lemme de l'élément est égal à lemme1 ou lemme2
- Si les lemmes ou occurrences à ne pas prendre en compte sont nombreux, ils font l'objet d'une liste indiquée à la suite des listes de schémas. Cette liste est référencée de la façon suivante : <catégorie Syntex![ListeX]>
- Pour faire référence à une catégorie Syntex, un lemme ou une occurrence incomplet(e), nous utilisons l'astérisque : <catégo*>
- <E1> <E2> <E3> indique que l'élément E1 précède E2 qui précède E3.
- Pour faire référence à un élément facultatif; nous utilisons les parenthèses : <E1>(<E2>)<E3> indique que la suite <E1><E2><E3> convient autant que la suite <E1><E3>
- {<E1>,<E2>} <E3> indique que l'élément E1 ou E2 précède E3.
- <E1>* indique que l'élément doit commencer par E1
- *<E1> indique que l'élément doit se terminer par E1
- Pour référer à une suite d'occurrences sans prendre en compte la catégorisation de Syntex, c'est-à-dire telle que nous la rencontrons dans le texte, nous indiquons la suite entre guillemets « xxx »
- les annotations générées sont exprimées entre barre verticales : |annotation|
- le symbole # réfère à n'importe quel caractère numérique

H.2. Repérage des paragraphes et des phrases

- Le repérage des titres de section est présenté en annexe D. Pour les textes utilisés ici, le découpage en paragraphe original a pu être conservé. Chaque unité textuelle alors entourée de sauts de paragraphes (¶ ou) est considérée a priori comme étant un paragraphe |para|. Seule exception : les structures énumératives.
- Pour définir si l'on se situe dans une structure énumérative, voici la procédure effectuée :
- Si l'unité considérée à priori comme étant un paragraphe est précédée d'une puce (-, *, •) ou d'un élément appartenant à un système de numérotation ({1, 2, 3, …}, {A, B, C, …}, {a, b, c, …}, {I, II, III, …} ou {i, ii, iii, …})
- Et que les paragraphes l'environnant (les précédent et/ou le suivant) aussi sont précédés de le même puce ou du numéro précédent ou suivant selon le cas, alors il y a énumération
- Alors le paragraphe se trouve dans une structure énumérative.
- Et alors tous les paragraphes successifs présentant cette puce ou cette numérotation sont considérés comme des items d'énumération |item| et sont comptabilisés comme ne formant qu'un seul paragraphe. Si le paragraphe précédent n'est composé que d'une phrase et que cette phrase se finit par la

ponctuation {:}, inclure ce paragraphe amorce dans l'énumération. La phrase sera alors annoté |amorce|

- Le découpage en phrases précédant ce programme est celui effectué par le programme Syntex (Bourrigault & Fabre 2000) et correspond à une définition typographique de la phrase : une suite de mots situés entre deux ponctuations fortes (. ? ! alinéa). Par défaut, chaque phrase est annotée |texte|, sauf si celle-ci a été annotée |amorce] ou |item|.

H.3. Repérage des différents éléments

H.3.1. Squelette du programme de repérage

Pour chaque sous-corpus, en commençant par la première 'phrase' (titres compris) du premier texte :

1. Suppression des puces et numérotations présentes en début de phrase
2. Repérage des connecteurs selon les schémas donnés dans la liste Liste(Connect).
3. Repérage du verbe principal.
   - Dans certaines situations il faut chercher un autre verbe que celui repéré car il ne s'agit pas du verbe principal. Ces situations se rencontrent lorsque :
     - le verbe est précédé d'une <CSub> ou d'un <ProRel>,
     - le verbe est entre parenthèses ou entre tirets.
   - Si Syntex a repéré un sujet régi par ce verbe, mémoriser son occurrence.
4. Découpage la phrase en deux parties : le prédicat et la position préverbale. Trois situations se rencontrent :
   - Si un verbe principale a été repéré et qu'il ne correspond pas au premier élément (ou celui suivant un adverbe ou un |Connect|), le texte précédent le verbe principal correspond à la partie préverbale et le reste de la phrase correspond au prédicat
   - Si le verbe est le premier élément (ou celui suivant un adverbe ou un |Connect|) i.e. il n'y a pas de position préverbale, on annote la phrase comme étant une construction spéciale : un sujet inversé ou, s'il y a un point d'interrogation final, une interrogative. On passe ensuite à l'étape 6.
   - Si aucun verbe n'a été repéré, on étiquette la phrase comme étant sans verbe et la position préverbale équivaut à toute la phrase.
5. Repérage des éléments présents en position préverbale. Il s'agit de distinguer le premier élément détaché |INIT1|, les suivants |INIT2| et le sujet grammatical |ThTop|
   - Extraction des différents blocs syntaxiques selon Syntex[200]. Opération récursive qui analyse chaque bloc syntaxique jusqu'à arriver au verbe principal.
   - Identification de la nature de chaque bloc : vérifier si le bloc en cours d'analyse est sujet grammatical (Module(ThTop) ci-après).
     - Si le bloc est le sujet grammatical, le prendre dans son entier : jusqu'au verbe principal. Puis aller en 5.3
     - Si le bloc n'est pas le sujet grammatical, c'est un élément détaché en initiale. Il faut alors le repérer dans son ensemble : jusqu'à la première ponctuation faible <Typo[,-:]> rencontrée. Si un élément détaché a déjà été repéré, vérifier s'il s'agit d'un nouvel élément détaché |INIT2| ou d'un élément prolongeant l'élément détaché précédent (Liste(INITProlong))
6. Analyse du prédicat pour identifier les cas de constructions spéciales
   - Si deux parties ont été découpées en 4, procéder à une reconcaténation entre le bloc identifié |ThTop| (dans certains cas, le(s) bloc(s) |INIT|) et le prédicat.
   - Si cette chaîne correspond à un schéma caractéristique d'une certaine construction spéciale (voir H.3.3) le |ThTop| est vidé de son contenu et on recherche le focus de la construction spéciale (procédure non détaillée ici)

200 L'outil Syntex calcule les relations de dépendances entre les éléments d'une phrase. Ces relations qui peuvent de nature sujet, attribut, déterminant, etc. relient un recteur et un régi. Dans le cas de la relation sujet, le sujet est régi par le verbe recteur, dans le cas de la relation déterminant, le déterminant est régi par le nom recteur, dans le cas de la relation prépositionnelle, c'est la préposition qui est recteur d'un élément régi (un nom ou autre). Extraire un bloc Syntex correspond à suivre les relations définies par Syntex, pour extraire le syntagme le plus complet.

7 Enregistrement des différents éléments de la phrase, de leur étiquettes morpho-syntaxique et de leur tête syntaxique : le |Connect|, le(s) |INIT|, le |ThTop| ou le |type de ThSpe| et le |Pred|
8 Passer à la 'phrase' suivante et recommencer à l'étape 1.

H.3.2. Listes des patrons utilisés pour le repérage des différents éléments

Liste(Connect) : schémas types de Connecteurs (voir également l'Annexe J.)

- {<CCoord![ni]>,<Adv![liste(Adv-C)]>,<Prep>} <Typo[,]>
- {<CCoord![ni]>,<Adv>,<Prep>} <Adv![liste(Adv-C)]> <Typo[,]>
- <Prep> <Prep> <Typo[,]>
- « qui plus est »
- « d'où »

Liste(Adv-C) : adverbes (tels que étiquetés par Syntex) non connecteurs

- « ici »
- « aujourd'hui »
- « hier »
- locution adverbiale comportant le lemme <[temps, heure, fois]>
- adverbe ayant pour terminaison « -ment »

Module(ThTop) : procédure pour déterminer si le bloc Syntex correspond au Thème topical

- Si
  - le bloc repéré ne commence pas par un élément <NomXXDate>*, et
  - Si le bloc repéré ne commence pas par <{[fin], [milieu], [début], [côté]}>*, et
  - Si le bloc repéré :
    - a pour tête syntaxique l'élément repéré comme étant le sujet du verbe principal, ou
    - correspond aux expressions régulières listées dans liste(ThTop), ou
    - si aucun sujet n'a été repéré par Syntex et que le bloc correspond à la chaîne « d'autres », ou
    - si aucun sujet n'a été repéré par Syntex et que le bloc relève du patron : <Prep [de]> <Det [le]> <Nom*> *, ou
    - si aucun sujet n'a été repéré par Syntex et que le bloc relève du patron : <Prep [de]> <Adj [#*]> <Nom*> *, et
  - Si le bloc ne répond pas à une règle du module(nonThTop),
- Alors le bloc constitue le ThTop de la phrase

Liste(ThTop) : schémas types d'élément sujet

- <Adj!{[nul] , [rien]}> <Det> *
- <Adj!{[nul] , [rien]}> {<NomXX*>, <NomPr>, !<NomXXDate>} *
- <Adj [#*]> <Nom* !{[milieu], [début], [fin], [côté]}> *
- <Adv> <Det> *
- <Adv> {<NomXX*>, <NomPr>, !<NomXXDate>} *
- <Pro>
- <Pro> *
- <Det> *
- {<Nom* !{[milieu], [début], [fin], [côté]}>, !<NomXXDate>} *
- <VINF> *"
- <Adj !{[nul] , [rien]}> *

Liste(non-sujet) : règles de repérage des éléments non sujets

- Si
  - le bloc repéré n'est pas jouxte au verbe principal repéré et qu'il correspond au patron : {<Det [le]>, <Pro [le]>}* <NomXXDate {[#*], liste(jour de la semaine)>*" * , ou
  - Si un INIT1 a déjà été repéré, qu'il commence par un élément <Csub> et qu'il ne contient pas d'élément <VCONJ>, ou
  - Si le bloc repéré se situe en amont de l'élément repéré comme étant le sujet du verbe principal (et qu'un élément sujet a été repéré) et ne correspond pas aux patrons :

        - {<Pro>, <Det>} *
        - {<Pro>, <Det>} <NomXXDate> *
        - <NomPr> *
        - {<Adj>, <Adv>} {<Det>, <Nom*>} *, ou
    - Si le bloc se situe en amont de l'élément repéré comme étant le sujet du verbe principal (et qu'un élément sujet a été repéré et correspond au patron : <Det>, <Pro>} <Adj> {<NomXXDate*>} *, ou
    - Si le bloc correspond à un élément prolongeant un élément détaché précédent (Liste(INITProlong)), ou
    - Si le bloc correspond au patron : « une fois » * <Typo[,]>, ou
    - Si le bloc se termine par la suite de caractère : { « plus tard, », « plus tôt, », « plus loin, », « plus près, »},
- alors le bloc n'est pas un ThTop

Liste(InitProlong) : schémas types des éléments détachés prolongeant un précédent élément détaché

- « c'est-à-dire » * (et les deux variantes : forme abrégée « càd » et forme sans traits d'union « c'est à dire »)
- <[comme]> *
- <[tel]> *
- <ProRel> *
- {<Ppa>,<Adj>} <Typo> quand le précédent élément détaché correspond au patron {<Ppa>,<Adj>} <Typo>
- <[#*]> quand le précédent élément détaché se termine par * <[#*]> <[-]>

H.3.3. Caractérisation des constructions spéciales (ThSpe)

Type de ThSpe	*Expression régulière ou Indices de caractérisation*
Construction clivée	Phrase contenant un élément détaché de forme « ce qu... » qui précède un sujet <Pro[ce]> et un verbe principal <VCONJ [être]>
	Phrase (une fois les INIT et les Connect enlevés) commençant par les chaînes : « c'est qu »*
	Phrase contenant un sujet <Pro[ce]> et un verbe principal <VCONJ [être]> et qui ne répond pas au patron des dislocations
Construction présentationnelle	Phrase (une fois les INIT et les Connect enlevés) commençant par les chaînes : « il existe »*, « il y (en) a »*, « il s'agit »*, « voici »*, « voilà »*
Construction impersonnelle	Phrase (une fois les INIT et les Connect enlevés) étiquetée par Syntex comme étant une <CSub>
	Phrases avec un sujet <Pro[il]« il »> qui précède un prédicat de forme[201] :

201 Ces prédicats peuvent bien évidemment être à la forme négative

Type de ThSpe	*Expression régulière ou Indices de caractérisation*	
		<VCONJ{[devoir], [pouvoir]> (<Adv>) (<Pro>) <VINF>*
		<Pro[en]> <VCONJ{[aller], [demeurer]}>*
		<VCONJ[s'ensuivre]>*
		<Pro[en]> <VCONJ[être]> <de]> <[même]>*
		(<Pro{[le], [lui]}>) <VCONJ{[apparaître], [paraître], [sembler], [rester]}> (<Adv>) <CSub[que]>*
		(<Pro{[le], [lui]}>) <VCONJ{[apparaître], [paraître], [sembler], [rester]}> (<Adv>) <CSub[que]>*
		(<Pro{[le], [lui]}>) <VCONJ{[apparaître], [paraître], [sembler], [rester]}> (<Adv>) {<Adj>, Ppr>) <Prep[de]> <VINF>*
		<VCONJ{[apparaître], [paraître], [sembler], [rester]}> (<Adv>) {<Adj>, Ppr>) <CSub[que]>*
		(<Pro{[le], [lui]}>) <VCONJ{[importer], [suffire], [s'avérer], [se trouver]}> (<Adv>) <CSub[que]>*
		(<Pro{[le], [lui]}>) <VCONJ{[importer], [suffire], [s'avérer], [se trouver], [convenir]}> (<Adv>) {<Adj>, Ppr>) <Prep[de]> <VINF>*
		(<Pro{[le], [lui]}>) <VCONJ>{[falloir]>*
		<Pro[en]> <VCONJ{[résulter], [subsister], [découler]}>*
		<VCONJ{[résulter], [ressortir], [découler], [se dégager]}>* <Prep[de]> * <Csub[que]>*
		<VCONJ[prévoir][202]> <Csub[que]>*
		<VCONJ[prévoir][203]> <Prep[de]> <VINF>*
		<VCONJ[se pouvoir]> <CSub[que]>*
Construction en « on... »	Toute phrase dont le sujet est le pronom « on » ou « nous »	
Sujet inversé	Si le ThTop est vide ou ne comporte que des pronoms clitiques objets (« le », « la », « les », « lui », « en », « y ») et qu'il n'y a pas de ponctuation finale d'interrogation	
Dislocation	Phrase contenant un INIT catégorisé |SN| qui précède un sujet <Pro [il]>	
	Phrase contenant un INIT catégorisé |SN| qui précède un sujet <Pro [ce]> et un verbe principal <VCONJ[être]>	
Autres constructions spéciales	Phrase ayant pour sujet le pronom « rien » ou « tout »	
	Phrase (une fois les INIT et les Connect enlevés) commençant par les chaînes : « c'est pourquoi »*, « voilà pourquoi », « voici pourquoi »*, « autant dire qu »*, « c'est d'autant »*, « d'autant plus qu »*, «le fait est qu »*, « force est de constater qu »*, « ce d'autant plus qu »*, « c'est-à-dire »*, « ce qui »*,	
	Phrase ayant pour sujet <[tel]> et pour verbe principal <VCONJ[être]>	

H.4. Caractérisation des différents éléments

Après avoir repérer les différents type d'éléments en position initiale de chaque phrase (étape décrite en H.3), le programme effectue : une caractérisation des |ThTop| et des |INIT|. Cette partie détaille en partie notre identification des différents degrés d'accessibilité et des fonctions et rôles sémantiques d'INIT. Nous ne détaillons pas comment notre programme identifie la catégorie morpho-syntaxique des ThTop et des INIT, cette étape de la caractérisation étant relativement simple et sans nouveauté.

202 Ici, seuls les temps composés sont acceptés (« est prévu », « était prévu », etc.)

203 Ici, seuls les temps composés sont acceptés (« est prévu », « était prévu », etc.)

H.4.1. Caractérisation du degré d'accessibilité

Deux identifications nécessaires à l'identification des différents degré d'accessibilité font appel à des informations autres que la seule catégorie morpo-syntaxique du |ThTop| : les cas de reprise lexicale et les cas de description courte.

1. Assignation, sur la base du jeu d'étiquettes TreeTagger, d'une catégorie syntaxique (SNdef, SNdem, PRO3, SN sans déterminant, etc.) et le cas échéant d'un genre et d'un nombre.
2. Recherche d'une reprise lexicale dans les cas de ThTop non pronominaux : si la tête syntaxique du ThTop a déjà été rencontrée dans la section de plus bas niveau en cours, localiser la position de cette mention précédente : dans quel type d'élément elle apparaît (ThTop, INIT, Focus, Pred), combien de phrases avant, combien de paragraphes avant et si la phrase dans laquelle elle apparaît est une première phrase de paragraphe.
3. Les degrés d'accessibilité sont ensuite assignés de la façon suivante :
   0. le degré 0 est assigné à toutes les phrases ayant été catégorisées ThSpe ou les phrases dont le ThTop a été catégorisé SNindef (i.e. avec un déterminant indéfini ou numéral), infinitif, SN sans déterminant ou ne répondant d'aucune catégorie (ex : « Agrégés et enseignantes » étiqueté <PpaMP> <CCoord> <AdjFP>).
   1. le degré 1 est assigné aux ThTop catégorisés noms propres pour lesquels aucune reprise lexicale n'a été repérée.
   2. Le degré 2 est assigné aux ThTop catégorisés SNdef pour lesquels aucune reprise lexicale n'a pas été repérée et qui comportent plus de trois mots.
   3. Le degré 3 est assigné aux ThTop catégorisés SNdef pour lesquels une reprise lexicale a été repérée ou qui comportent moins de quatre mots.
   4. le degré 4 est assigné aux ThTop catégorisés noms propres pour lesquels une reprise lexicale été repérée.
   5. Le degré 5 est assigné aux ThTop catégorisés SNdem pour lesquels aucune reprise lexicale n'a pas été repérée et qui comportent plus de trois mots.
   6. Le degré 6 est assigné aux ThTop catégorisés SNdem pour lesquels une reprise lexicale a été repérée ou qui comportent moins de quatre mots.
   7. Le degré 7 est assigné aux ThTop de forme pronominale (pronoms personnels, démonstratifs ou 'autre' comme par ex. les autres) ou catégorisés SN possessifs.

RQ : le degré d'accessibilité de la phrase précédente est également noté afin d'observer les successions de phrases par leur degré d'accessibilité.

H.4.2. Caractérisation des INIT

Les INIT correspondent aux éléments détachés en initiale situés entre un connecteur 'pur' et le sujet grammatical (que ce soit un ThTop ou le sujet d'une construction spéciale). Le tableau ci-dessous présente les expressions régulières utilisées pour effectuer leur caractérisation automatique au niveau de la fonction et du rôle sémantique. La colonne « ordre » indique l'ordre d'application des expressions régulières.

Fonction (_rôle sémantique)	*ordre*	*Expressions régulières ou Indices de caractérisation*
APPO	6	<Prep[avec,sans]> <Nom*>*
	7	<Prep [de]> <Nom*> <Prep !{[à],[en]}> <Nom*>*
	8	{Ppa>, <Ppr !{« concernant », « suivant »}>}*
	9	|Sadj| <Typo>
	11	INIT catégorisé |SN| ou |REL| (sauf si une dislocation à gauche ou une construction clivée a été repérée et alors l'INIT est annoté |TOP|)
ARGU	1	INIT non détaché par une virgule précédent une phrase sans ThTop (et donc une construction spéciale)
	41	SP commençant par <Prep {[de][à]}> et qui ne correspond à aucun autre schéma

Fonction (_rôle sémantique)	*ordre*	*Expressions régulières ou Indices de caractérisation*
		d'INIT et qui précède un ThSpe catégorisé |sujet inversé|
CIRC_tps	2	<Nom*[début, milieu, fin]>*
	3	<Det> <NomXXDate>*
	4	<Det> < Adj> <NoXXDate>* (ex : *Tous les quatre ans,*)
	20	<Prep> (<Det>) <NomXXDate>*
	21	<[quand, lorsque, depuis, jusque, pendant, avant, après, dès, durant, lors
	22	{« une fois », « au cours d », « alors qu », « à l'occasion d », « au moment », « au début d », « à la fin d », « au milieu d », « tout au long d », « au terme d »}*
	24	{[en]} <{[1###], [2###], [###]}> *
	34	INIT catégorisé |SP| ou |SN| et comportant les lemmes : {[siècle], [décennie], [millénaire], [centenaire], [année], [an], [semestre], [trimestre], [mois], [quinzaine], [semaine], [jour], [nuit], [journée], [nuitée], [Xème], [Ième], [Vème], [printemps], [automne], [hiver], [été], [aujourd'hui], [rentrée], [date], [hier], [période], [durée], [temps], [moment], [époque], liste(jours de la semaine), liste(mois)
	37	{<Prep> , <AdvGP>} * <{[1###], [2###], [###]}> *
CIRC_spa	15	<Prep {[à], [en], [dans], [sur], [entre]}> (<Det>) <NomPr>*
	16	<Prep [de]> (<Det>) <NomPr> <Prep [à]> <NomPr>*
	25	<{[dans], [en], [vers], [sur], [pour]}> * <Nom*[pays], [ville], [région], [département], [état], [territoire], [académie], [canton], [continent], [littoral], [agglomération], [lieu], [île], [zone], [comté], [commune]}> *
	26	* <{[nord], [sud], [est], [ouest], [frontière]}>
	27	{« au sein d », « autour d », « de part et d'autre », « aux alentours », « aux environs », « à l'intérieur d », « à l'extérieur d »}*
	28	{« sur », « le long d »}* {« axe », « ligne »}*
	29	<{[dans], [à], [vers], [sur]}> * <{[nord], [sud], [est], [ouest], [frontière], [porte]}>
	38	<Prep {[dans], [sur], [à]}> <Det> <Nom* « dont l'occurrence commence par une majuscule »>
	39	<Prep {[à], en]}><Nom* « dont l'occurrence commence par une majuscule »>
	40	<Prep [de]> (<Det>) <Nom* « dont l'occurrence commence par une majuscule »> <Prep [à]> (<Det>) <Nom* « dont l'occurrence commence par une majuscule »>
CIRC_not	5	<Ppr {[suivant], [concernant]}>*
	14	<Prep {[chez], [pour]}> <NomPr>*
	30	<{[côté], [chez], [selon]}>*
	31	{« au sein d », « en cas d », « sur le plan », « sur ce plan », « d'après »}*
	23	{« parmi », « en ce qui concerne », « quant à », « concernant »}*
	32	<{[à],[de]}> * <{[niveau], [coté], [plan]}> *
	33	<{[dans], [de]}> <Det> (<Adj>) {« point(s) de vue », « optique(s) », « perspective(s) », « vision(s) », « cadre(s) », « cas(s) », « contexte(s) », « domaine(s) », « condition(s) »}*
	38	{« s'agissant », « au sujet », « à partir »} <[de]> !{<[1###]>, <[2###]>, <[###]>}*

Fonction (_rôle sémantique)	*ordre*	*Expressions régulières ou Indices de caractérisation*
	35	{« en termes d », « au vu »}*
	36	<[sur]> <[#*]>*
	41	<Csub {[pour*], [sur*]}>*
CIRC_ssi (condition)	18	{« si », « au cas où », « à condition »* « sous condition »}*
CIRC_but	19	« dans le but d »*
		<Prep{[pour], [afin]}> <[que]>*
		<Prep[pour]> <VINF>*
CIRC_0	17	{« comme », « tout comme »}*
	42	Tout INIT catégorisé |SP|, |FIN| ou |INF| et ne répondant pas à un patron précédent
MODA	12	<Adv><typo[,]>
	13	« selon toute »* !<Det>
	Pré-raitement : Expression présente dans la liste(MODA)	
TEXT	Pré-raitement : Expression présente dans la liste(TEXT)	
TOP	10	INIT catégorisé |SN| ou |REL| ou |INF| précédent une dislocation à gauche ou une construction clivée (repérée lors de la phrase décrite en H.3) Pour le |INF| il faut qu'elles ne commencent pas par une <Prep>
		<Pro>*

Liste(MODA) : expressions catégorisées MODA par le programme : « à dire vrai, », « à l' évidence, », « à priori, », « à titre d' »*, « à titre de comparaison, », « à vrai dire, », « apparemment, », « au final, », « au mieux, », « au total, », « bien entendu, », « autrement dit, », « bien évidemment, », « bien sûr, », « bref, », « d'une manière »*, « d'une façon »*, « dans ce sens, », « dans l'ensemble, », « de ce fait, », « de façon »*, « de facto, », « de fait, », « de la sorte, », « de manière générale, », « de manière plus générale, », « de même, », « de plus en plus, », « de plus, », « de surcroît, », « de toute évidence, », « de toute façon, », « de toute manière, », d'une certaine façon, », d'une certaine manière, », d'une façon »*, d'une manière »*, « en bref, », « en ce sens, », « en cela, », « en fin de compte, », « en général, », « en l'occurrence, », « en moyenne, », « en outre, », « en parallèle, », « en particulier, », « en pratique, », « en réalité, », « en règle générale, », « en théorie, », « en tout cas, », « en tout état de cause, », « en un mot, », « en un sens, », « enfin et surtout, », « finalement, », « fort heureusement, », « oui, », « par ailleurs, », « par conséquent, », « par contraste, », « par exemple, », « petit à petit, », « peu à peu, », « plus que tout, », « qui plus est, », « quoi qu' il en soit, », « sans doute, », « sans nul doute », « selon toute apparence, », « selon toute évidence, », « selon toute probabilité, », « selon toute vraisemblance, », « sur le fond, »

Liste(TEXT) : expressions catégorisées TEXT par le programme : « à ce propos, », « à contrario, », « à côté de cela, », « à deuxième vue, », « à la partie », « à la section », « à l'inverse, », « à l'opposé, », « à première vue, », « à seconde vue, », « au chapitre », « au contraire, », « au demeurant, », « au final, », « au paragraphe », « au premier abord, », « cinquièmement, », « conclusion », « d'abord, », « d'ailleurs, », « dans la {première, deuxième, etc.} {partie, section} », « dans la {partie, section} »*, « dans le {chapitre, paragraphe} »*, « dans le {premier, deuxième, etc.} {chapitre, paragraphe} », « dans un second temps, », « d'autre part, », « deuxièmement, », « du reste, », « d'un autre côté, », « d'un côté, », « d'une part, », « en clair, », « en conclusion, », « en conséquence, », « en contrepartie, », « en d'autres termes, », « en définitive, », « en deuxième »*, « en effet, », « en premier »*, « en revanche, », « en second »*, « en tout premier »*, « en troisième »*, « enfin, », « introduction », « par contre, », « par voie de conséquence, », « parallèlement, », « plus encore, », « premièrement, », « quatrièmement, », « septièmement, », « sixièmement, », « tout d'abord, », « troisièmement, »

Annexe I. Connecteurs repérés

Le tableau qui suit expose la liste complète des connecteurs repérés dans notre corpus – en dehors des hapax[204]. Les lignes de couleur correspondent aux connecteurs présents dans les trois sous-corpus. Une différence est faite ici entre une forme apparaissant avec une virgule d'une autre apparaissant sans virgule. Nous distinguons également les connecteurs « modifiés » vs. non « modifiés (ex. Mais vs. Mais pourtant,).

Connecteur	Atlas	Geopo	Peopl	Total
Mais	53	197	325	575
et	39	43	112	194
Ainsi,	26	112	29	167
Mais,	8	27	77	112
Enfin,	25	64	20	109
Or,	6	61	35	102
Cependant,	29	22	21	72
Or	6	29	29	64
Car	0	3	52	55
Certes,	7	29	17	53
Ainsi	16	5	32	53
Néanmoins,	14	27	3	44
Pourtant,	3	21	17	41
En fait,	5	18	14	37
De même,	6	24	5	35
D'où	0	13	19	32
Toutefois,	1	25	9	35
Et,	2	2	23	27
Aussi	9	2	15	26
Cependant	15	4	3	22
De plus,	11	6	4	21
Aussi,	5	11	1	17
Enfin	7	2	4	13
Pour autant,	9	3	0	12
Puis	4	1	7	12
D'une part,	2	8	1	11
Pourtant	2	4	4	10
Désormais,	0	3	7	10
Ensuite,	1	4	5	10
Surtout,	0	4	5	9
Puis,	1	0	8	9
Jamais	0	1	7	8
Là encore,	1	4	3	8
Car,	0	0	6	6
Ici,	1	0	5	6

Connecteur	Atlas	Geopo	Peopl	Total
Pourquoi	1	1	4	6
Voilà	0	1	5	6
Plus tard,	0	2	4	6
Ou	2	2	1	5
Et pourtant,	0	0	5	5
Dorénavant,	0	4	1	5
Depuis,	3	1	1	5
Entre-temps,	1	0	3	4
Alors	0	0	4	4
Ici	3	0	1	4
Ou encore,	0	2	2	4
Parfois,	3	0	1	4
Pas	3	1	0	4
Au-delà,	1	3	0	4
Et pourtant	1	0	3	4
Déjà,	0	1	3	4
Ailleurs,	1	1	1	3
Souvent	2	0	1	3
Mais pourtant,	0	3	0	3
Et puis,	0	0	3	3
Ou bien	1	0	2	3
Ou,	0	1	2	3
Autant	0	3	0	3
Là,	0	1	2	3
De là,	0	0	2	2
Non,	0	0	2	2
Mais alors,	0	0	2	2
Donc,	0	0	2	2
Certes	0	2	0	2
Et surtout,	0	0	2	2
Là-haut,	0	0	2	2
Peut-être,	0	0	2	2
Voici	0	0	2	2
Une fois encore,	1	1	0	2
Très vite,	0	1	1	2

204 Mot n'apparaissant qu'une fois dans l'ensemble du corpus.

Connecteur	Atlas	Geopo	Peopl	Total
Très souvent,	1	1	0	2
Surtout	1	0	1	2
Souvent,	0	2	0	2
Soit	0	2	0	2
Si,	0	1	1	2
Pour sa part,	0	2	0	2
Alors,	0	1	1	2
Plus encore,	0	2	0	2

Connecteur	Atlas	Geopo	Peopl	Total
Auparavant,	0	2	0	2
Parfois même,	1	0	1	2
Ben	0	2	0	2
Non	0	0	2	2
Néanmoins	2	0	0	2
Parfois	1	0	1	2
Total	343	825	997	2165

ANNEXE J. ÉVALUATION DU PROGRAMME D'EXTRACTION ET DE CARACTÉRISATION DES ÉLÉMENTS PRÉVERBAUX

Deux évaluations ont été effectuées sur les résultats du programme d'annotation de notre corpus. D'une part, l'annotation de 339 phrases prises aléatoirement dans les différents textes des trois sous-corpus a été évaluée par nous-mêmes. D'autre part, l'annotation de 230 INIT1 a été comparée à celle effectuée manuellement par 7 juges linguistes (doctorants et chercheurs).

Il est évident que cette évaluation ne permet pas d'asseoir la validité d'un outil. Ce n'est d'ailleurs pas là son ambition. Ce que nous voulons, c'est se donner une idée de la précision de notre programme d'annotation automatique. Dans une visée plus applicative qui dépasse le cadre de cette thèse, une validation multi-juge sur de plus grosses données aurait été essentielle.

J.1. Quelques définitions :

Précision : pertinence du programme, c'est-à-dire rapport entre le nombre de données correctement repérées et caractérisées sur le nombre total des données ramenées.

Rappel : couverture du programme, c'est-à-dire rapport entre le nombre de données repérées et caractérisées comme étant pertinentes et le nombre total de données réellement pertinentes.

Ainsi, si l'on note S l'ensemble des données à analyser et V l'ensemble des données à repérer et/ou annoter correctement, nous obtenons les formules suivantes (où ^ signifie l'intersection) :

$$précision = \frac{S \wedge V}{S} \qquad rappel = \frac{S \wedge V}{V}$$

J.2. Distinction entre éléments préverbaux et prédicat

Dans 98% des cas, la délimitation de la position préverbale s'est correctement effectuée. Dans les 7 cas d'erreur, (i) soit il n'y a pas eu de délimitation car aucun verbe conjugué n'a été trouvé et donc, aucun prédicat n'a été repéré,(ii) soit il y a eu un découpage mais erroné. Les principales causes d'erreurs sont dues à l'étiquetage du TreeTagger (dans 5 phrases) ou à une mauvaise gestion des pronoms relatifs ou des conjonctions de subordination[205]. Toutes les autres phrases ont été correctement découpées ; ce qui donne les taux suivants : **précision = 98, rappel = 100**.

J.3. repérage des éléments détachés (INIT1, INIT2) vs. éléments intégrés

Pour 3 phrases sur 339, des éléments sujets ont malheureusement été annotés comme étant des éléments détachés (INIT). Pour 4 phrases sur 339, des éléments sujets annotés comme tel étaient en fait des éléments détachés. On obtient alors une précision de 98,8 et un rappel de 99,1. Si l'on réduit le champ d'investigation aux

205 Les verbes conjugués repérés après un pronom relatif ou une conjonction de subordination ne peuvent pas constituer le verbe principal de la phrase et donc le début du prédicat. Dans de telles configurations, le programme 'saute' jusqu'au prochain verbe conjugué repéré.

phrases présentant au moins un INIT (ce qui équivaut à 100 phrases), on obtient les taux suivants : **précision = 96, rappel = 97**.

J.4. Caractérisation des éléments intégrés (ThTop vs. ThSpe)

Sur les 339 phrases test, 6 erreurs ont été recensées :

- 2 présentent un élément annoté ThSpe alors qu'il s'agit en réalité d'un ThTop; et
- 4 présentent un élément annoté ThTop alors qu'il s'agit en fait d'un ThSpe.

Ts		*Tt*	
Précision	*Rappel*	*Précision*	*Rappel*
98.8	99.4	99.4	98.8

Tableau 1: Précision et rappel pour la caractérisation des Tt/s

J.5. Distinction entre INIT1 et INIT2

Sur les 339 phrases évaluées, un seul cas d'erreur a été observé : un INIT2 a été distingué alors qu'il s'agissait de la continuation de l'INIT1 (précision = 100, rappel = 99,7). Si l'on réduit le champ d'investigation aux cent phrases présentant au moins un INIT, on obtient les taux suivants : **précision = 100, rappel = 99**.

J.6. Évaluation personnelle de l'annotation des INIT

Sur les 112 annotations d'INIT effectués sur les 339 phrases test (100 INIT1 et 12 INIT2), nous observons 9 erreurs.

3 de ces erreurs correspondent à une non attribution de rôle sémantique à un circonstant : un 0 (signe d'indéfinitude) a été attribuée au lieu du rôle sémantique existant (1 circonstant de condition, 1 de domaine et 1 de temps).

Six annotations erronées ont été relevées :

1. 3 cas de mauvaise attribution de l'étiquette du rôle sémantique d'un circonstant, en l'occurrence ici, une confusion entre CIRC_tps et CIRC_not ;
2. 3 erreurs d'attribution de la fonction : deux éléments annotés comme étant des éléments disloqués alors qu'en fait il n'y avait pas dislocation et un élément annoté circonstant alors qu'il jouait un rôle de modalisateur

Précision	*Rappel*
97,3	94,6

Tableau 2: Précision et rappel pour l'annotation des INIT - version personnelle

J.7. Évaluation multijuge de l'annotation des INIT_CIRC

230 INIT pour lesquels la fonction était majoritairement circonstancielle – CIRC et le rôle sémantique indéfini ont été soumis à la caractérisation manuelle de 7 juges linguistes (doctorants et chercheurs). Cette caractérisation est ici comparée à celle effectuée par le traitement automatique.

Avant de commencer ces comparaisons, il faut noter que la caractérisation manuelle n'est pas des plus évidente, comme le montrent les différents commentaires reçus de la part des juges, amis que je remercie encore pour leur dur labeur.

« C'est vraiment compliqué! Tu verras que souvent je n'ai pas classé les circonstants car j'avais du mal à interpréter la phrase. »

« Pas facile à faire pour plusieurs raisons : bien sûr des ambiguïtés, mais aussi le fait que je ne suis pas sûre que tout le monde donne la même définition/extension aux catégories proposées »

Cette difficulté est intrinsèque à la définition des circonstants, comme il l'a été exposé en partie V.6.2.3.

J.7.1. Caractérisation de la fonction des INIT

fonction	caractérisation manuelle	caractérisation automatique	recouvrement des deux caractérisations
CIRC (circonstant)	160	197	154
ARGU (argument inversé)	21	5	2
MODA (modalisateur)	24	7	5
APPO (apposition)	5	8	3
indéfinis(sables)	12	3	1
éléments de topicalisation	6	6	6

Tableau 3: Comparaison homme/programme des annotations sur la fonction des INIT

Le chapitre VII explique pourquoi la fonction CIRC a été attribuée dans les cas où la fonction reste incertaine entre circonstant et autre. Cette préférence ne se retrouve pas dans la caractérisation humaine. En effet, que ce soit entre les arguments inversés (ARGU) et les circonstants (CIRC) ou entre les adverbiaux modalisateurs (MODA) et les circonstants, les juges ont plus souvent penchés pour la première fonction. Il faut d'ailleurs remarquer que pour 6 arguments inversés et 6 modalisateurs, un rôle sémantique propre aux circonstants a été attribué (domaine, moyen, manière).

Cette différence de choix se retrouve dans la faiblesse des cas de recouvrement et est plus amplement détaillée dans la section suivante qui s'intéresse particulièrement au typage de la fonction CIRC. Il est montré dans cette section qu'un bon nombre de caractérisations ARGU et MODA est sujet à discussion, voire à contestation.

J.7.2. Annotation de la fonction CIRC

J.7.2.a) 6 éléments annotés CIRC par l'humain et différemment par le programme

CIRC_not(humain) <-> TEXT(programme) : 1

À l'opposé, nombre de départements méridionaux, ceux d'Alsace et quelques départements parisiens enregistrent une stabilisation, voire une baisse de la proportion de collégiens et lycéens dans la population scolaire.

CIRC_tps(humain) <-> MIL(programme) : 1

Les références aux personnages historiques sont nombreuses. En tête, et de très loin, vient Napoléon. En second, Catherine de Médicis, pour ses mots : "Eh bien, nous irons au prêche", quand elle croyait perdue la bataille de Dreux ; pour sa devise : Odiate e aspettate! (Haïssez et attendez!).

CIRC_manière(humaine) <-> MODA(programme) : 2

De façon assez classique, les Etats ont privilégié les hausses les moins visibles, c'est-à-dire les moins coûteuses électoralement.

De façon générale, c'est la doctrine Bush concernant la croisade contre le terrorisme et l'utilisation de la force pour dissuader les adversaires qui fit l'objet de critiques de plus en plus franches.

CIRC_not(humain) <-> APPO(programme) : 1

Sans crise grave ou de façon différée, des points de retournement marqués par des élections dans les pays à suffrage universel, jalonnent ces changements.

CIRC_not(humain) <-> indéfinis(sable) : 1

Ailleurs (dans Modeste Mignon), la simple notation d'un geste module les paroles d'un dialogue, leur donne relief, nuances, et finalement signification véritable : Balzac aurait pu, a priori, applaudir des deux mains à la déclaration du baron de Canalis, le "grand poète" : […]

Parmi ces 6 cas, seuls les deux derniers sont à considérer comme des erreurs (des silences) dans le traitement. En effet, les quatre premiers éléments correspondent bien à notre définition des fonctions caractérisées de façon automatique (voir VII.1.2) et leur valeur circonstancielle reste pour nous très contestable.

Pour le repérage des éléments circonstants, la précision est donc de 158/160 => 98,75 %

J.7.2.b) 43 éléments caractérisés CIRC par le programme et différemment par l'humain

Ces 43 éléments ont été caractérisés par l'humain comme étant des MODA (16), des ARGU (17), des APPO(2) ou des éléments indéfinis(sables) (9).

On remarque sur cet ensemble que certains éléments à fonction non circonstancielle se sont vus cependant attribuer un rôle sémantique (4 MODA et 5 ARGU). De plus, des éléments tels que les SP {par + SN} ont obtenu la fonction de MODA, ARGU, APPO et élément indéfini(ssable), ce qui laisse douter de la précision d'une telle caractérisation, qui pourrait alors être de l'ordre des circonstants de moyen ou de manière. Enfin, à voir la liste ci-dessous relatant de l'ensemble de ces 43 éléments, il apparaît clairement que peu d'entre eux ont de façon certaine une fonction non circonstancielle.

En prenant tous ces éléments en considération, nous obtenons un rappel de 154/197 => 78,2 %. En ne prenant que les 19 pour lesquels il est assez certain que la fonction n'est pas celle d'un circonstant (lignes indexées par une x), ce rappel devient 178/197 => 90,3 %

CIRC (programme) <-> MODA (humain) : 16

Par un trait de génie, *il a évité d'étaler son ego en usant de la première personne. (moyen)*
Par recherche du classicisme, *il a évité les termes trop précis, appelant un navire de guerre nauis longa sans préciser s'il est une birème ou une trière, appliquant cette interpretatio Romana aux dieux gaulois dont il fait Mercure, Jupiter, ou Minerve. (moyen)*
x Dans une large mesure, *l'actuelle crise budgétaire des Etats fédérés est le produit du succès de la stratégie conservatrice de "Nouveau Fédéralisme". (domaine)*
Dans la combinaison complexe qui produit au final le comportement politique, *il n'existe aucun déterminisme simple qui conduirait de la condition économique à l'orientation politique. (domaine)*
x Tout au plus, *lorsqu'il accepte de sortir du pseudo-tragique bourgeois, écrit-il une aimable pièce : Est-il bon ?*
x Par malheur, *trop de conventions nées de l'ignorance ont perverti les règles naturelles de la société.*
Même si sa connaissance est plutôt de seconde main, l'intérêt de Balzac pour Leibniz apparaît très tôt et ne se dément jamais.
En termes religieux, *Dieu ne se laisse pas trouver sans nous réprouver ou nous dissoudre en lui.*
En deux traits : *l'émotion vraie est celle que provoque un événement vrai ; elle se caractérise par la surprise, l'immédiateté de la réponse, le désordre, la croyance en la réalité de l'objet émouvant : or, l'émotion du comédien est répétée, perfectionnée, ordonnée,*
x De toute façon, *il appartient à la raison de critiquer les produits de l'enthousiasme.*
Comme l'écrit excellemment Félix Davin, *ce qui doit être recherché dans l'oeuvre de Balzac, c'est "la science inconnue dont la pensée conduit l'auteur malgré lui".*
A la différence des positions anciennes, *l'Ouest n'apparaît pas comme une zone de faible abstention.*
À la condition de ne plus faire du savoir et du non- savoir deux états, mais deux mouvements (apprendre et oublier), *on peut rendre compte des erreurs empiriques : elles résultent du mauvais ajustement d'un souvenir à une impression présente.*

x À la vérité, *cette question avait été déjà abordée en 1647 par le jésuite Grégoire de Saint-Vincent qui avait établi des relations générales entre intégrales.*
x au fond, *pour lui, tout accès à la "spiritualité" passe par un "maquillage" et rien n'est plus contradictoire avec l'homme spirituel que l'homme naturel.*
x En très bref, *le problème des partis peut s'énoncer ainsi : deux joueurs décidant d'arrêter une partie avant son achèvement, il s'agit de déterminer comment l'enjeu restant doit être réparti entre eux.*
Même si sa connaissance est plutôt de seconde main, *l'intérêt de Balzac pour Leibniz apparaît très tôt et ne se dément jamais.*

CIRC(programme) <-> ARGU(humain) : 17

x Ou alors les quelques 500 entretiens menés par le FBI avec des personnes d'origine arabe*, suite à un autre décret du Ministère de la Justice pris le 15 novembre 2001.*
Outre ses conséquences internationales, *la guerre contre le terrorisme a des effets tout aussi évidents sur la scène politique américaine.*
Hormis les spécialistes de la question, *cette crise ne semble pourtant pas intéresser l'opinion publique américaine, et encore moins les observateurs étrangers.*
Plus que des divergences opposant Républicains et Démocrates, *les futurs débats sur les questions internationales mettront en évidence les désaccords existant au sein de le GOP entre des parlementaires conservateurs et d'autres plus "centristes", à l'instar de ceux opposant les "faucons" et les "colombes" au [...]*
Chez les naturalistes, *Balzac allait trouver ce qu'il dit être l'idée première de La Comédie humaine, qui lui "vint d'une comparaison entre l'Humanité et l'Animalité".*
Par l'ordonnance, *il est créé dans chaque région, une Agence Régionale de l'Hospitalisation constituée sous la forme d'un Groupement d'Intérêt Public entre l'Etat et les caisses d'assurances maladie.*
x À ces trois catégories, *il faut ajouter les enseignants et élèves-enseignants des établissements de formation, regroupés dans les Instituts universitaires de formation des maîtres (IUFM) depuis la rentrée 1991 [...]*
x Contre cette tendance, *Théophile Gautier avait réagi dès ses années romantiques et continue d'être, à l'époque de Baudelaire, le grand porte-parole de la gratuité de l'art ; mais ici surgit une autre équivoque qui va motiver le second grand refus de Baudelaire.*
x À l'un deux, *l'épreuve coûta la raison ; Dostoïevski en fut marqué pour la vie.*
x Entre ces deux France, *les oppositions ne sont pas seulement de nombre, mais de nature : ici, le problème scolaire, même s'il s'est déplacé de l'école primaire au lycée, reste une question de locaux et de postes en retard sur les besoins ; là, au contraire, le problème est [...]*
Entre ces deux termes de l'action humaine, *il est une autre formule dont s'emparent les sages, et je lui dois le bonheur de ma longévité.*
x Contre l'empirisme, *toujours renaissant, il fait le procès de la certitude sensible.*
Contre l'intuitionnisme, *il déjoue l'attrait de l'ineffable, tient le droit du raisonnement, de l'articulation, de la détermination.*
x au romancier comme à l'historien, *il faut "mille détails" (Préface d'Une fille d'Ève).*
x De sa mère, *il ne parlera que par occasions.*
Grâce à un habile travail sur l'opinion, *Montcornet est obligé de quitter Les Aigues.*
À défaut d'épopée, *il a rêvé de s'élever au second rang de la hiérarchie littéraire avec le théâtre.*

CIRC(programme) <-> APPO(humain) : 2

Par ses silences et ses ellipses, *le roman fait que le lecteur sonde les âmes et découvre des "souffrances inconnues".*
Comme l'écrit Madeleine Ambrière-Fargeaud (Balzac et "La Recherche de l'Absolu"), *Balzac a manifesté une prédilection, parmi les naturalistes, pour ceux qui étaient en même temps des philosophes et même des voyants, et parmi les mystiques, pour ceux qui "voyaient" le monde et l'expliquaient de manière rationnelle et scientifique.*

CIRC(programme) <-> indéfinis(sables)(humain) : 9

Par leur formation rhétorique autant que par leurs informations plus complètes, *ces anciens avaient été sensibles aux procédés insinuants des Commentaires.*
Comme le dit Victor Hugo, *dans son oraison funèbre : "Tous ses livres ne forment qu'un livre" ; ajoutons, un livre dont le mouvement ne semble pas s'être encore arrêté, une oeuvre infinie.*
x En particulier les souffrances de l'abandon, *les humiliations, les faiblesses secrètes.*
De tous, un peu.
Comme lui, *il voudra "instituteur des hommes", et il estime nécessaire de contribuer à l'amélioration de l'homme dans le cadre de la société.*

x Dans son oeuvre tant scientifique que littéraire paraît toujours le souci de concilier les grandes vérités de la raison et de la foi avec une expression accessible à tous : *le succès sans précédent de les Provinciales tient essentiellement à son art d'y mettre les problèmes de théologie à la portée des gens du monde, y compris, disait -on, des femmes dont l'éducation ne comportait pas l'étude du latin.*
x Comment intégrer deux espèces contraires en un même genre, *les rendre amies*
x Des mobilités de travail, *amenant à se déplacer dans un département voisin, parfois pour quelques années, parfois définitivement, des mobilités de travail amenant à aller régulièrement à plusieurs centaines de kilomètres, sans pour autant déménager, des mobilités quotidiennes […]*
Dans une sorte de continuum scandé de bonds et de changements qualitatifs, *il y a encore ses racines et n'a pas rompu les liens.*

J.7.3. Caractérisation du rôle sémantique des CIRC

Cette comparaison a pour but premier d'apprécier le rôle sémantique des CIRC dont le rôle n'est pas défini de façon automatique. Cette appréciation se base sur les 154 éléments dont la fonction circonstancielle est attribuée tant de façon manuelle qu'automatique.

rôle sémantique		caractérisation manuelle	caractérisation automatique	recouvrement des deux caractérisations
temps		19	6	6
lieu		44	11	8
domaine		46	19	12
0	concession	9	117	
	cause	4		
	manière	7		
	condition	10		
	autres	6		
	indéfini	9		
	total	45		

Tableau 4 : Recouvrement des attributions manuelle et automatique des rôles sémantiques

Dans ce tableau, nous voyons que les éléments au rôle sémantique définis de façon automatique sont assez faibles (36/154), ce qui est normal vu le critère de sélection des éléments[206]. Le fait qu'il y ait certains circonstants dont le rôle est automatiquement défini provient du fait que le programme établi pour la caractérisation automatique a été amélioré depuis l'envoi aux 7 juges des éléments à caractériser. Avant de voir précisément quels rôles ont été attribués aux éléments laissés indéfinis par le programme, voyons les cas pour lesquels le rôle défini automatiquement diffère de celui défini humainement.

J.7.3.a) divergence d'attribution entre humain et programme

Tous les CIRC_tps caractérisés comme tels de façon automatique le sont également de façon manuelle. Seuls les CIRC_spa et les CIRC_not montrent des cas de non recouvrement. Pour les CIRC_spa, les 3 éléments non recouverts se sont tous vus attribuer le rôle de CIRC_not (le premier cas affichant selon notre jugement un circonstant plus de lieu que de domaine) :

> ***Sur ces axes,*** *les académies de Poitiers et de Lyon occupent des positions relais : ce sont des académies de milieu de carrière.*

206 Nous rappelons que les éléments servant de test ici ont été sélectionnés pour le caractère indéfini de leur rôle sémantique.

x Des Bijoux indiscrets au Neveu de Rameau, *des pamphlets sur la querelle des Bouffons (1753) aux Leçons de clavecin et principes d'harmonie par M. Bemetzrieder (1771), il prend part aux polémiques — Rameau contre Lulli, Pergolèse contre Rameau, Gluck contre Pergolèse — et conseille Grétry.*
x Dans Swedenborg, *Richer a découvert une idée qui deviendra une idée balzacienne : "L'univers visible est lié par une union indissoluble à l'univers immatériel, le tout est un par essence et varié par nature" (La Nouvelle Jérusalem).*

Pour les CIRC_not, nous avons 5 circonstants de condition, 1 CIRC_spa et 1 circonstant au rôle sémantique indéfini (CIRC_0).

Les circonstants de condition correspondent ici à ce que nous considérons comme des circonstants de domaine, c'est-à-dire des éléments qui posent un critère réduisant le monde d'interprétation à un univers de discours autre que spatio-temporel : un domaine d'activité ou de connaissance, une thématique, un point de vue, une ou des condition(s) particulière(s).

En cas de récession économique, *la chute de la consommation influe directement sur le niveau des prélèvements étatiques.*
Dans ces conditions, *les résultats des élections de novembre 2002 devraient avoir pour effet de laisser le champ libre aux Républicains sur les questions internationales dans les deux prochaines années.*
Dans ces conditions, *les Etats doivent prendre le risque d'augmenter les impôts ou d'en créer de nouveaux.*
Dans ces conditions, *l'élément fondamental pour l'équilibre budgétaire de les Etats est bien la décision fédérale!*
Dans ces conditions de forte "volatilité", *la seule option qui reste ouverte à les Etats consiste à augmenter leur impôt sur le revenu ou à le rendre plus progressif.*

L'élément caractérisé CIRC_spa a un rôle sémantique plus délicat à définir. Cependant, il rentre plus dans notre définition des CIRC_not que dans celle des CIRC_spa. Les circonstants de lieu que nous cherchons à caractériser comme tels sont ceux qui établissent un repère spatial, voir géographique.

Sur fond d'inconscient, *quelque conscience s'éclaire.*

Enfin, le circonstant laissé indéfini par l'humain correspond tout à fait à notre définition des CIRC_not, étant paraphrasable par « à propos d'un structure sociale… »

Sur une structure sociale et une histoire à l'époque de la Révolution puis au XIXe qui conduit au mouvement et à la contestation, *l'événement unificateur récent de la Résistance, déclenche un comportement politique qui va durer une génération".*

Pour récapituler, les 36 éléments dont le rôle sémantique a été automatiquement défini l'ont été de façon plutôt satisfaisante par l'humain, en gardant toutefois à l'esprit que la caractérisation automatique est inspiré de notre opinion personnelle et influencée par nos objectifs d'étude qui ne sont pas de caractériser les différentes circonstances possibles, mais de voir comment se répartissent dans les textes certains circonstants particuliers : ceux de temps, de lieu et, accessoirement, de domaine (des réserves se faisant autour du caractère un peu fourre-tout de cette catégorie).

Taux de rappel 1[207] : 26/36 => 72,2 %, Taux de rappel 2[208] : 34/36 =>94,4 %

Cependant, il nous semble assez délicat de se fier à ce taux de rappel. Dans le cadre de l'élaboration d'un outil de TAL, d'autres investigations seraient à effectuer. Mais cela sort du cadre de cette thèse. Cette comparaison entre caractérisation humaine et annotation automatique est principalement utile pour « se faire une idée » de l'adéquation entre le traitement mis en place et les divers jugements humains répertoriés.

207 Prend en compte tous les éléments dont la caractérisation manuelle diffère de celle automatique

208 Ne prend en compte que les éléments dont la caractérisation manuelle diffère de celle automatique après vérification personnelle (marqués d'une **x**).

J.7.3.b) caractérisation manuelle des CIRC automatiquement indéfinis (CIRC_0)

Pour les 117 CIRC_0 dont le rôle sémantique a été laissé indéfini par le programme, la caractérisation manuelle de leur rôle sémantique donne les résultats suivants :

rôles sémantiques	*Nb de cas*	
temps	13	48
lieu	35	
domaine	38	45
condition	5	
but (*À cet effet, ils…*)	1	
perspective (*Au dire de Nietzsche, Dostoïevski…*)	1	
indéfini	9	31
concession (*Malgré cet héritage historique, il reste que …*)	8	
cause (*Par la faute d'un contrat léonin, Dostoïevski…*)	4	
manière (*Par une miséricorde gratuite, Dieu…*)	6	
moyen (*Par l'ingéniosité de sa conception, cette machine …*)	1	
quantité (*Au nombre de 750 000, les enseignants …*)	1	
restriction (*Malgré leur grand nombre, les petites écoles…*)	1	
opposition (*Contrairement à une légende malveillante, il…*)	1	

Tableau 5 : Caractérisation manuelle des éléments dont le rôle sémantique reste automatiquement indéfini

Les éléments qui nous intéressent particulièrement ici sont ceux qui sont laissés indéfinis par le traitement automatique et qui semblent correspondre aux rôles pertinents pour l'étude, c'est-à-dire les rôles de temps et de lieu ; ainsi que plus accessoirement ceux de domaine auxquels nous ajoutons les circonstants dits de condition, de but et de perspective qui correspondent à notre définition des CIRC_not.

Si l'on s'intéresse à la caractérisation des CIRC_tps, seuls 13/117 des éléments laissés indéfinis semblent appartenir à cette catégorie, ce qui reste relativement correct (taux de précision de 88,9%).

Pour la catégorie des CIRC_spa, le manque de précision est beaucoup plus important (35/117 => 70,1% de précision), et un regard plus précis sur ce manque est à apporter. Si l'on regarde de plus près ces 35 éléments, on remarque rapidement qu'il y a une certaine confusion entre les CIRC_spa et les CIRC_not, comme le montrent les quelques exemples suivants.

> ***Dans l'enseignement public,*** *un instituteur sur trois a plus de 45 ans, un autre moins de 30 ans.*
> ***En lycée professionnel,*** *les enseignants des disciplines générales sont mieux lotis que ceux des disciplines techniques, plus diversifiées.*
> ***De l'école maternelle à l'université,*** *le système éducatif français accueille environ quatorze millions d'élèves et étudiants.*
>
> ---
>
> ***De la pierre à l'homme,*** *du ver à l'étoile, l'univers reste un parce qu'il est un tout.*
> ***Dans Vremja,*** *il publie Humiliés et offensés (Unizennye i oskorblënnye, 1866), roman à la Dickens, très autobiographique, où pourtant apparaît un type qui, pour E. M. de Vogüé, n'est qu'un "traître de mélodrame", mais qui n'en est pas moins le premier de ces hom*
> ***Dans l'entourage du président,*** *les arguments en faveur du renforcement des prérogatives de l'Exécutif n'étaient pas jugées essentielles dans la seule perspective du règlement d'une crise ; elles s'imposaient également dans le bon fonctionnement d'une politique étrangère.*

Dans un éditorial récent du Washington Post, *un Sénateur démocrate de Caroline de Sud expliquait que l'administration Bush avait réussi à masquer la gravité de la situation budgétaire de l'Etat fédéral.*
Autour de ce standard historique national, *en même temps qu'il se construisait , de nouvelles spécificités régionales sont apparues, qui ne constituaient pas la négation d'un mouvement d'opinion nationale, mais son mode de construction dans les différentes régions et l'affirmation répétée au [...]*
Dans ce colombier auquel l'âme est comparée, *il n'y a que des colombes — des opinions vraies, des connaissances — , il n'y a pas de non — colombes que l'âme puisse attraper.*
Dans la littérature concernant la fiscalité étatique, *les fragilités du système budgétaire sont systématiquement mises en avant.*
Entre lui et Saltykov-Chtchédrine, *le satirique né, c'est une joute.*

On peut également remarquer que 42,8 % des cas (15/35) proviennent du corpus ATLAS (plus précisément du texte sur la France scolaire) et sont du même type que les trois premiers exemples cités ci-dessus. Il y a là une confusion entre la localisation dans le système scolaire (localisation interne au phénomène abordé) et la localisation dans un lieu spécifique (externe au phénomène). La même confusion peut se faire avec la localisation temporelle, comme pour le 3e exemple, où l'on peut se demander si ***De l'école maternelle à l'université,*** n'aurait pas pu aussi correspondre à une localisation temporelle. En ne considérant pas comme des CIRC_spa ces 15 cas, ainsi que les 8 autres exemples cités ci-dessus, le taux de précision pour les CIRC_spa passe à 89,7% (12/117)

Concernant les CIRC_not, le calcul est plus sujet à discussion, vu la faiblesse définitoire très vague de cette catégorie. Si l'on s'en tient aux chiffres du J.7.3.b, le taux de précision est de : 45/117 => 61,5 %. Cependant, comme nous venons de le voir, nous pourrions changer les chiffres de ce tableau en considérant que, lors de la caractérisation manuelle, seuls 12 éléments laissés indéfinis sont des CIRC_spa, les 23 restants étant des CIRC_not, ce qui donne les résultats suivants :

rôles sémantiques	***Nb de cas***
domaine	38
domaine mis en lieu	23
condition	5
but (*À cet effet, ils…*)	1
perspective (*Au dire de Nietzsche, Dostoïevski…*)	1
	68

Tableau 6 : CIRC_not laissés indéfinis par la caractérisation automatique

Ce changement descend le taux de précision pour cette catégorie (68/117 => 41,8 %). Il reste cependant indéniable que ces éléments sont des circonstants. Les analyses quantitatives portant plutôt sur ce point que sur la catégorie sémantique de « domaine ».

Annexe K. Extrait du sous-corpus GEOPO avec balises XML, après annotation automatique

Cette annexe présente un extrait du corpus GEOPO annoté et formaté selon la norme XML. La totalité du corpus GEOPO est disponible dans différentes versions à l'URL :

http://w3.univ-tlse2.fr/erss/textes/pagespersos/hodac/Corpus/.

On peut y trouver les fichiers .pdf d'origine, le fichier .txt, .xml, et une version html pour 'surfer' entre les éléments en position initiale. Ce corpus est libre de droit et peut être utilisé sans contraintes. Nous remercions chaleureusement l'IFRI pour leur confiance (http://www.ifri.org).

Ci dessous sont présenté tout d'abord l'extrait au format .txt et au format xml tel que annoté par notre programme. Cet extrait est issu d'un texte écrit par François Vergniolle de Chantal comme indiqué dans la version xml. Ce texte est également disponible en ligne dans sa version originale telle que diffusée par l'IFRI à l'URL :
www.ifri.org/.../publications/les_cahiers_du_cfe_1032456434976/publi_P_publi_cfe_fvcterrorisme_1037634219682

K.1. Version texte brut

LA LUTTE CONTRE LE TERRORISME : ESSAI DE BILAN INSTITUTIONNEL

AUTEUR : François Vergniolle de Chantal, Docteur en Sciences Politiques de l'IEP de Paris, est Maître de Conférences en civilisation américaine à l'Université de Bourgogne.

Depuis une quarantaine d'années, les républicains se sont fait fort de réduire le poids de l'Etat fédéral. Or l'actuelle lutte contre le terrorisme, menée par une équipe républicaine qui, pourtant, adhère totalement aux critiques contre le Big Government, remettrait en cause l'engagement conservateur en faveur de la décentralisation. Les différentes mesures annoncées depuis septembre 2001 vont toutes dans le même sens, un considérable renforcement de la présence de l'Etat fédéral. Comme toutes les guerres menées par les Etats-Unis, celle entamée contre le terrorisme risquerait, elle aussi, de renforcer la centralisation. Quels sont les aspects de ce retour de l'Etat central ? Comment s'opère la recentralisation, et avec quelles conséquences dans l'équilibre fédéral ? Finalement, quelles sont les conclusions à tirer de cette évolution ? En particulier, comment s'articule la lutte contre le terrorisme avec 'engagement conservateur en faveur des Etats fédérés ? Selon nous, la lutte contre le terrorisme ne serait pas similaire aux évolutions entraînées par les autres conflits. Elle débouche en fait sur un activisme tous-azimut, qui concerne aussi bien l'Etat fédéral que les Etats fédérés et les autorités locales (villes, comtés). Plutôt que de parler de centralisation, il faudrait évoquer un renforcement des fonctions légitimes de chacun des niveaux du gouvernement : la défense et la protections des citoyens pour le niveau fédéral ; les autorités locales, elles, gèrent les moyens de réponse immédiats aux agressions terroristes (police, pompier, santé). L'essentiel des problèmes suscités par la protection du territoire contre le terrorisme réside dans la coordination entre les différents organes. L'administration actuelle s'engage résolument dans cette voie, et entame une réorganisation massive des administrations nationales.

Face à l'urgence : les premières décisions de l'administration Bush

Dans le mois qui a suivi l'attentat du 11 septembre, l'administration a procédé à un certain nombre d'initiatives spectaculaires à plus d'un titre, notamment par l'intrusion massive des autorités fédérales dans différents domaines où, jusqu'alors, l'interventionnisme fédéral n'était pas de mise. A commencer par la sécurité aérienne, au vu, bien sûr, du déroulement des attentats : les attaques contre des objectifs civils semblaient alors être l 'objectif de prédilection des groupes islamistes. C'est pourquoi, sous la responsabilité du Secrétaire aux Transports, Norman Y. Mineta, un nouveau texte a été adopté par le Congrès dès le 19 novembre, le Aviation and Transportation Security Act (ATSA, Public Law 107-71). Ainsi est instituée la Transportation Security Administration (TSA), qui prend en charge la sécurité de l'aviation civile, auparavant de la responsabilité de la Federal Aviation Administration (FAA). A partir de février 2002, la nouvelle instance a " fédéralisé " les points de contrôle des 429 aéroports commerciaux des Etats-Unis, processus, qui, en fin de compte, devrait encore prendre quelques mois. Dorénavant, les compagnies privées de sécurité - jusqu'ici sous-traitantes des compagnies aériennes - ne sont donc plus responsables du contrôle des passagers ; près de 28000 fonctionnaires fédéraux doivent maintenant prendre le relais, et leur recrutement devrait se faire avec des critères plus exigeants que

ceux requis jusqu'alors. Pendant ce temps, les craintes d'attentats contre d'autres types de cibles civiles se multipliaient. Ainsi, un certain nombre d'élus démocrates (dont le Sénateur de New York Hillary R. Clinton) ont appelé en novembre à une prise en charge fédérale de la sécurité des 103 centrales nucléaires du pays par la Nuclear Regulatory Commission. Mais l'initiative est, pour le moment, restée lettre morte au Congrès : en l'état actuel de la situation, la protection des sites nucléaires est toujours assurée par les quelques 57000 réservistes et membres de la Garde Nationale qui ont été mobilisés suite aux attentats. Initialement chargés aussi de la sécurité dans les aéroports, ils en ont été rapidement relevés lors de la création de la TSA ; ils assurent maintenant exclusivement la défense des centrales nucléaires, et l'administration Bush semble s'en satisfaire.

Ces actions immédiates ont été renforcées par d'autres mesures, budgétaires, qui vont directement à l'encontre du libéralisme économique prôné par les républicains. Ainsi, le Président a immédiatement décidé des aides d'urgence : 40 milliards de dollars répartis entre l 'Etat de New York et le FBI, les agences de renseignement et l'armée ; à ce montant s'ajoute 15 milliards de dollars pour aider les compagnies aériennes. Autant dire que le non-interventionnisme économique de l'Etat fédéral a été immédiatement relegué au second rang devant l'urgence de la situation. La restriction budgétaire a tout de suite cédé la place à la nécessité de lutter contre le terrorisme.

Après quatre années d'excédents fédéraux, le budget de 2003 - qui débute en octobre 2002 - renoue avec les déficits. Sous l'effet conjugué du ralentissement économique et de la lutte contre le terrorisme (les démocrates rajouteraient aussi les baisses d'impôts parmi les facteurs explicatifs), le budget devrait afficher un déficit de l'ordre de 43 milliards de dollars. Les principaux postes budgétaires sont dorénavant la sécurité du territoire (homeland security) et la défense. Dans le premier cas, le budget passe de 15 milliards de dollars à 38 milliards, une part non-négligeable (un peu moins de trois milliards) étant consacrée à la lutte contre le bioterrorisme. A un niveau institutionnel, et plus seulement fonctionnel, l'administration Bush a décidé de renforcer considérablement les polices locales, pompiers, et services d'urgence, qui, tous, constituent la première ligne de défense vis-à-vis des attaques terroristes. Environ 3,5 milliards de dollars - soit une multiplication par dix des financements antérieurs - sont ainsi destinés aux autorités locales, municipales et étatiques, c'est-à-dire aux échelons politiques responsables de ces différents corps. En ce qui concerne la défense, le Secrétaire, Donald Rumsfeld, se trouve maintenant à la tête du second poste dans le budget fédéral. Le Président a obtenu une rallonge budgétaire de 48 milliards de dollars, soit une enveloppe qui dépasse le montant du budget militaire annuel de n'importe quel autre pays dans le monde. L'effort ainsi consenti est comparable à celui engagé par Truman lors de la Guerre de Corée. Comme il y a cinquante ans, les Etats-Unis sont véritablement entrés dans un budget de guerre : celui-ci devrait atteindre 396 milliards de dollars en 2003, et, si les prévisions se concrétisent, se chiffrer à 470 milliards en 2007.

A priori, l'administration Bush a adopté des dispositions budgétaires qui la placent en décalage par rapport aux discours républicains en faveur de la modestie budgétaire et de la nécessaire rigueur dans les dépenses. Dans ce domaine, l'Etat fédéral a bénéficié d'une nouvelle marge de manœuvre, inespérée au vu de l'orientation politique de l'équipe dirigeante. C'est d'autant plus vrai que ces mesures ne sont pas précisément des décisions sur lesquelles l'administration se serait engagée à revenir. Au contraire, la Présidence a, dans un second temps de sa lutte contre le terrorisme, élaboré un cadre plus général qui cherche à pérenniser les décisions prises à l'automne. L'accroissement des pouvoirs de l'Etat fédéral ne tient pas de l'accident de parcours. Il s'agit au contraire d'une priorité des pouvoirs publics.

L'élaboration d'un cadre de lutte contre le terrorisme

[...]

K.2. Version XML annotée

K.2.1. Descriptif des annotations indiquées

<CORPUS>	le corpus	
<DOCUMENT>	les différents documents (textes) constitutifs du corpus	
	num=""	identifiant du document dans le corpus
<AUTEUR>	information sur l'auteur du document **</AUTEUR>**	
<S>	section (le titre y est inclus)	
	niveau=""	niveau de la section
	num=""	numéro de la section dans le corpus

<TITRE>	titres du document **</TITRE>**	
	niveau=""	niveau du titre. les titres de niveau="0" sont les titres du docments, les autres sont les titres de section
</TITRE>		
	Titre_repris=""	nombre de fois qu'un élément du titre a été repris dans la section
<ENUM>	structure énumérative	
<P>	paragraphe	
	num=""	numéro du paragraphe dans le corpus
	nat= "(text,item)"	si le paragraphe est un item de structure énumérative, nat="item". Sinon, nat="txt"
<PHR>	phrase (les titres sont considérés comme des phrases)	
	num=""	numéro de la phrase dans le paragraphe
	nat= "(TopTh,Cliv,ILmp...)"	type de construction. TopTh = phrase avec sujet grammatical plein et à valeur référentielle. Cliv
	amorce="(0,1)"	si la phrase est une amorce, amorce = "1"
<CONNECT>	connecteur en initiale de phrase **</CONNECT>**	
<INIT>	élément détaché en initiale de phrase	
	pos=""	position de l'élément détaché : premier élément détaché après le connecteur s'il y a lieu (1), deuxième élément détaché (2), etc.
	morph= "(SP_,SN_,REL,...)"	caractère morpho-syntaxique de l'élément : syntagme prépositionnel (SP_), syntagme nominal (SN_), adverbe (ADV), proposition relative (REL), proposition subordonnée (FIN), proposition infinitive (INF), syntagme adjectival (ADJ), proposition participe présent (PPR), proposition participe passé (PPA), nature morphosyntaxique indéfinie (00_).
	fct= "(MODA,CIRC,...)"	fonction de l'élément détaché : apposition (APPO), circonstant (CIRC), modalisateur d'énonciation (MODA), organisateur textuel (TEXT), argument inversé (ARGU), construction détachée de topicalisation (TOPI)
	sem= "(tps,spa,...)"	rôle sémantique de lélément dtécahé s'il y a lieu : localisation temporelle (tps), spatiale (spa), notionnelle (not), etc.
	redeno= "(1,0)"	si la tête nominale de l'élément détaché reprend un élément déjà exprimé dans la section en cours, redeno ="1"
	posredeno=""	si redeno="1", localisation de l'occurrence précédente : élément__position du paragraphe_positon de la phrase. ex : "Pred__Para-4_Phr-25" signifie que l'occurrence précédente se situe dans la partie prédicat d'une phrase située 4 paragraphes avant celui-ci et 25 phrases avant celle-ci. Si cette phrase avait été la première phrase du paragraphe en question, l'attribu aurait eu pour valeur "Pred_§Para-4_Phr-25"
	RepriseduTitre= "(1,0)"	si un élément de l'élément détaché reprend un élément exprimé dans le titre de section directement supérieur, RepriseduTitre="1"
	MotFreq=""	nom récurrent du texte présent dans cet élément. Un nom récurrent recouvre 1% des occurrences nominales relevées dans le texte.
</INIT>		
<TOPTH>	Thème topical, ou en d'autres termes, sujet grammatical plein à valeur référentielle	
	morph= "(PRO3_MS, SNdem_FS,etc.)"	caractère morpho-syntaxique de l'élément accompagnée de son genre et nombre : pronom personnel de troisième personne fémini plueirl ("PRO3_FP"), syntagme nominal défini masculin pluriel ("SNdef_MP"),

		syntagme nominal san déterminant ("SNssDET"), etc.
	redeno= "(1,0)"	si la tête nominale du TopTh reprend un élément déjà exprimé dans la section en cours, redeno ="1"
	posredeno=""	si redeno="1", localisation de l'occurrence précédente : élément__position du paragraphe_positon de la phrase. ex : "Pred__Para-4_Phr-25" signifie que l'occurrence précédente se situe dans la partie prédicat d'une phrase située 4 paragraphes avant celui-ci et 25 phrases avant celle-ci. Si cette phrase avait été la première phrase du paragraphe en question, l'attribu aurait eu pour valeur "Pred_§Para-4_Phr-25"
	RepriseduTitre= "(1,0)"	si un élément du TopTh reprend un élément exprimé dans le titre de section directement supérieur, RepriseduTitre="1"
	MotFreq=""	nom récurrent du texte présent dans cet élément. Un nom récurrent recouvre 1% des occurrences nominales relevées dans le texte.
	maillon= "(1,0)"	calcul la position de ce TopTh dans une chaine de référence de type progression thématique à thème constant. Ce calcul brut se base sur la présence d'un pronom personnel de troisième personne (PRO3). Si le TOPTH est de forme PRO3 et que le TOPTH de la phrase précédente est un SN de même nombre et de même genre, alors ce TOPTH constitue le 2e maillon d'une TP constante... et ainsi de suite. Les SN définis et démonstratifs présentant la même tête que l'antécédent d'origine font également partie du calcul.
</TOPTH>		
<PRED>	Prédicat	
	RepriseduTitre= "(1,0)"	si un élément du prédicat reprend un élément exprimé dans le titre de section directement supérieur, RepriseduTitre="1"
	MotFreq=""	nom récurrent du texte présent dans cet élément. Un nom récurrent recouvre 1% des occurrences nominales relevées dans le texte.
<FOCUS>	Focus de la construction spéciale	
	morph= "(PRO3_MS, SNdem_FS,etc.)"	caractère morpho-syntaxique de l'élément accompagnée de son genre et nombre : pronom personnel de troisième personne fémini plueirl ("PRO3_FP"), syntagme nominal défini masculin pluriel ("SNdef_MP"), syntagme nominal san déterminant ("SNssDET"), etc.
	redeno= "(1,0)"	si la tête nominale du Focus reprend un élément déjà exprimé dans la section en cours, redeno ="1"
	posredeno=""	si redeno="1", localisation de l'occurrence précédente : élément__position du paragraphe_positon de la phrase. ex : "Pred__Para-4_Phr-25" signifie que l'occurrence précédente se situe dans la partie prédicat d'une phrase située 4 paragraphes avant celui-ci et 25 phrases avant celle-ci. Si cette phrase avait été la première phrase du paragraphe en question, l'attribu aurait eu pour valeur "Pred_§Para-4_Phr-25"
	RepriseduTitre= "(1,0)"	si un élément du Focus reprend un élément exprimé dans le titre de section directement supérieur, RepriseduTitre="1"
	MotFreq=""	nom récurrent du texte présent dans cet élément. Un nom récurrent recouvre 1% des occurrences nominales relevées dans le texte.
</FOCUS> **</PRED>** **</PHR>** **</P>** **</ENUM>**		

</S>
</DOCUMENT>
</COPUS>

K.2.2. Extrait du corpus annoté

<?xml version="1.0" encoding="UTF-8"?>

<DOCUMENT nom="GEOPO">

<TEXTE num="1">

<TITRE niveau="0">TITRE : La lutte contre le terrorisme : essai de bilan institutionnel</TITRE>

<AUTEUR><P num="2">AUTEUR : François Vergniolle de Chantal, Docteur en Sciences Politiques de l'IEP de Paris, est Maître de Conférences en civilisation américaine à l'Université de Bourgogne.</P></AUTEUR>

<P num="3" nat="text"><PHR num="1" nat="TopTh" amorce="0"><INIT pos="1" morph="SP_" fct="CIRC" sem="tps" ana="0" redeno="0" RepriseDuTitre="0">Depuis une quarantaine d' années, </INIT><TOPTH morph="SNdef_MP" redeno="0" RepriseDuTitre="0" maillon="0">les républicains </TOPTH><PRED RepriseduTitre="0">se sont fait fort de réduire le poids de l' Etat fédéral .</PRED></PHR><PHR num="2" nat="TopTh" amorce="0"><CONNECT>Or </CONNECT><TOPTH morph="SNdef_MS" redeno="0" RepriseDuTitre="0" maillon="0">l' actuelle lutte contre le terrorisme, menée par une équipe républicaine qui , pourtant , adhère totalement à les critiques contre le Big Government , </TOPTH><PRED RepriseduTitre="0">remettrait en cause l' engagement conservateur en faveur de la décentralisation .</PRED></PHR><PHR num="3" nat="TopTh" amorce="0"><TOPTH morph="SNdef_FP" redeno="0" RepriseDuTitre="0" maillon="0">Les différentes mesures annoncées depuis septembre 2001 </TOPTH><PRED RepriseduTitre="0">vont toutes dans le même sens, un considérable renforcement de la présence de l' Etat fédéral .</PRED></PHR><PHR num="4" nat="TopTh" amorce="0"><INIT pos="1" morph="SP_" fct="CIRC" sem="0" ana="0" redeno="0" RepriseDuTitre="0">Comme toutes les guerres menées par les Etats-Unis, </INIT><TOPTH morph="PROdem" redeno="0" RepriseDuTitre="0" maillon="0">celle entamée contre le terrorisme </TOPTH><PRED RepriseduTitre="0">risquerait, elle aussi, de renforcer la centralisation .</PRED></PHR><PHR num="5" nat="interro" amorce="0"><PRED RepriseduTitre="0">Quels sont les aspects de ce retour de l' Etat central ?</PRED></PHR><PHR num="6" nat="interro" amorce="0"><PRED RepriseduTitre="0">Comment s' opère la recentralisation, et avec quelles conséquences dans l' équilibre fédéral ?</PRED></PHR><PHR num="7" nat="TopTh" amorce="0"><INIT pos="1" morph="ADV" fct="MODA" sem="0" ana="0" redeno="0" RepriseDuTitre="0">Finalement, </INIT><TOPTH morph="NoSe" redeno="0" RepriseDuTitre="0" maillon="0">quelles sont les conclusions à tirer de cette évolution ? </TOPTH><PRED RepriseduTitre="0"></PRED></PHR><PHR num="8" nat="Interro" amorce="0"><INIT pos="1" morph="SP_" fct="MODA" sem="0" ana="0" redeno="0" RepriseDuTitre="0">En particulier, </INIT><PRED MotFreq="terrorisme" RepriseduTitre="0">comment s' articule la lutte contre le terrorisme avec ' engagement conservateur en faveur de les Etats fédérés ?</PRED></PHR><PHR num="9" nat="TopTh" amorce="0"><INIT pos="1" morph="SP_" fct="CIRC" sem="not" ana="0" redeno="0" RepriseDuTitre="0">Selon nous, </INIT><TOPTH morph="SNdef_FS" redeno="1" posredeno="Pred__Para-0_Phr-1" RepriseDuTitre="0" maillon="1">la lutte contre le terrorisme </TOPTH><PRED RepriseduTitre="0">ne serait pas similaire aux évolutions entraînées par les autres conflits .</PRED></PHR><PHR num="10" nat="TopTh" amorce="0"><TOPTH morph="PRO3_FS" redeno="0" RepriseDuTitre="0" maillon="2">Elle </TOPTH><PRED MotFreq="autorité" RepriseduTitre="0">débouche en fait sur un activisme tous-azimut, qui concerne aussi bien l' Etat fédéral que les Etats fédérés et les autorités locales (villes, comtés) .</PRED></PHR><PHR num="11" nat="II..._SN" amorce="0"><INIT pos="1" morph="INF" fct="CIRC" sem="0" ana="0" redeno="1" posredeno="Pred__Para-0_Phr-7" RepriseDuTitre="0">Plutôt que de parler de centralisation, </INIT> il faudrait évoquer <FOCUS morph="SNindef_MS" redeno="1" posredeno="Pred__Para-0_Phr-8">un renforcement des fonctions légitimes</FOCUS> de chacun des niveaux du gouvernement : la défense et la protections des citoyens pour le niveau fédéral ; les autorités locales, elles, gèrent les moyens de réponse immédiats aux agressions terroristes (police, pompier, santé) .</PHR><PHR num="12" nat="TopTh" amorce="0"><TOPTH morph="SNdef_MS" redeno="0" RepriseDuTitre="0" maillon="0">L' essentiel des problèmes suscités par la protection du territoire contre le terrorisme </TOPTH><PRED RepriseduTitre="0">réside dans la coordination entre les différents organes .</PRED></PHR><PHR num="13" nat="TopTh" amorce="0"><TOPTH morph="SNdef_MS" redeno="0" RepriseDuTitre="0" maillon="0">L' administration actuelle </TOPTH><PRED MotFreq="administration" RepriseduTitre="0">s' engage résolument dans cette voie, et entame une réorganisation massive des administrations nationales .</PRED></PHR></P>

<S niveau="1" num="2">

<TITRE niveau="1" TitreRepris="13"><PHR num="1" nat="TopTh" amorce="0"><INIT pos="1" morph="SP_" fct="CIRC" sem="0" ana="0" redeno="0" RepriseDuTitre="0">Face à l' urgence : </INIT><TOPTH morph="SNdef_FP" redeno="0"

RepriseDuTitre="0" maillon="0">les premières décisions de l' administration Bush </TOPTH><PRED RepriseduTitre="0"></PRED></PHR>
</TITRE>

<P num="5" nat="text"><PHR num="1" nat="TopTh" amorce="0"><INIT pos="1" morph="SP_" fct="CIRC" sem="tps" ana="0" redeno="0" RepriseDuTitre="0">Dans le mois qui a suivi l' attentat du 11 septembre, </INIT><TOPTH morph="SNdef_MS" redeno="0" RepriseDuTitre="1" maillon="0">l' administration </TOPTH><PRED MotFreq="autorité" RepriseduTitre="0">a procédé à un certain nombre d' initiatives spectaculaires à plus d'un titre, notamment par l' intrusion massive des autorités fédérales dans différents domaines où, jusqu'alors, l' interventionnisme fédéral n' était pas de mise .</PRED></PHR><PHR num="2" nat="TopTh" amorce="0"><INIT pos="1" morph="INF" fct="CIRC" sem="0" ana="0" redeno="0" MotFreq="sécurité" RepriseDuTitre="0">A commencer par la sécurité aérienne, </INIT><INIT pos="2" morph="0__" fct="CIRC" sem="not" ana="0" RepriseDuTitre="0">au vu, bien sûr, du déroulement des attentats : </INIT><TOPTH morph="SNdef_FP" redeno="0" RepriseDuTitre="0" maillon="0">les attaques contre des objectifs civils </TOPTH><PRED RepriseduTitre="0">semblaient alors être l' objectif de prédilection des groupes islamistes .</PRED></PHR><PHR num="3" nat="Autre_SP" amorce="0"> C' est pourquoi, <FOCUS morph="SP" redeno="0">sous la responsabilité du Secrétaire aux Transports,</FOCUS> Norman Y. Mineta, un nouveau texte a été adopté par le Congrès dès le 19 novembre, le Aviation and Transportation Security Act (ATSA, Public Law 107 - 71) .</PHR><PHR num="4" nat="SujetInv_SN" amorce="0"> Ainsi est instituée <FOCUS morph="SNdef_FS" MotFreq="administration" redeno="1" posredeno="Pred__Para-0_Phr-1">la Transportation Security Administration (</FOCUS>administration</PHR><PHR num="5" nat="TopTh" amorce="0"><INIT pos="1" morph="SP_" fct="CIRC" sem="tps" ana="0" redeno="0" RepriseDuTitre="0">A partir de février 2002, </INIT><TOPTH morph="SNdef_FS" redeno="0" RepriseDuTitre="0" maillon="0">la nouvelle instance </TOPTH><PRED RepriseduTitre="0">a " fédéralisé " les points de contrôle des 429 aéroports commerciaux des Etats-Unis, processus, qui, en fin de compte, devrait encore prendre quelques mois .</PRED></PHR><PHR num="6" nat="TopTh" amorce="0"><CONNECT>Dorénavant, </CONNECT><TOPTH morph="SNdef_FP" redeno="0" RepriseDuTitre="0" maillon="0">les compagnies privées de sécurité - jusqu'ici sous-traitantes de les compagnies aériennes - </TOPTH><PRED RepriseduTitre="0">ne sont donc plus responsables du contrôle des passagers ; près de 28000 fonctionnaires fédéraux doivent maintenant prendre le relais, et leur recrutement devrait se faire avec des critères plus exigeants que ceux requis jusqu'alors .</PRED></PHR><PHR num="7" nat="TopTh" amorce="0"><INIT pos="1" morph="SP_" fct="CIRC" sem="tps" ana="1" redeno="0" RepriseDuTitre="0">Pendant ce temps, </INIT><TOPTH morph="SNdef_FP" redeno="0" RepriseDuTitre="0" maillon="0">les craintes d' attentats contre d'autres types de cibles civiles </TOPTH><PRED RepriseduTitre="0">se multipliaient .</PRED></PHR><PHR num="8" nat="TopTh" amorce="0"><CONNECT>Ainsi, </CONNECT><TOPTH morph="SNindef_MS" redeno="1" posredeno="Pred__Para-0_Phr-7" RepriseDuTitre="0" maillon="0">un certain nombre d' élus démocrates (dont le Sénateur de New York Hillary R. Clinton) </TOPTH><PRED MotFreq="sécurité" RepriseduTitre="0">ont appelé en novembre à une prise en charge fédérale de la sécurité des 103 centrales nucléaires du pays par la Nuclear Regulatory Commission .</PRED></PHR><PHR num="9" nat="TopTh" amorce="0"><CONNECT>Mais </CONNECT><TOPTH morph="SNdef_MS" redeno="0" RepriseDuTitre="0" maillon="0">l' initiative </TOPTH><PRED RepriseduTitre="0">est, pour le moment, restée lettre morte au Congrès : en l'état actuel de la situation, la protection des sites nucléaires est toujours assurée par les quelques 57000 réservistes et membres de la Garde Nationale qui ont été mobilisés suite à les attentats .</PRED></PHR><PHR num="10" nat="TopTh" amorce="0"><INIT pos="1" morph="PPA" fct="APPO" sem="0" ana="0" redeno="0" MotFreq="sécurité" RepriseDuTitre="0">Initialement chargés aussi de la sécurité dans les aéroports, </INIT><TOPTH morph="PRO3_MP" redeno="0" RepriseDuTitre="0" maillon="0">ils en </TOPTH><PRED MotFreq="administration" RepriseduTitre="0">ont été rapidement relevés lors de la création de la TSA ; ils assurent maintenant exclusivement la défense des centrales nucléaires, et l' administration Bush semble s' en satisfaire .</PRED></PHR></P>

<P num="6" nat="text"><PHR num="1" nat="TopTh" amorce="0"><TOPTH morph="SNdemo_FP" redeno="0" RepriseDuTitre="0" maillon="0">Ces actions immédiates </TOPTH><PRED RepriseduTitre="0">ont été renforcées par d'autres mesures, budgétaires, qui vont directement à l'encontre du libéralisme économique prôné par les républicains .</PRED></PHR><PHR num="2" nat="TopTh" amorce="0"><CONNECT>Ainsi, </CONNECT><TOPTH morph="SNdef_MS" redeno="0" RepriseDuTitre="0" maillon="0">le Président </TOPTH><PRED MotFreq="milliard" RepriseduTitre="0">a immédiatement décidé des aides d'urgence : 40 milliards de dollars répartis entre l' Etat de New York et le FBI, les agences de renseignement et l' armée ; à ce montant s' ajoute 15 milliards de dollars pour aider les compagnies aériennes .</PRED></PHR><PHR num="3" nat="TopTh" amorce="0"><CONNECT>Autant </CONNECT><TOPTH morph="INF" redeno="0" RepriseDuTitre="1" maillon="0">dire que le non interventionnisme économique de l' Etat fédéral a été immédiatement relegué au second rang devant l' urgence de la situation . </TOPTH><PRED RepriseduTitre="0"></PRED></PHR><PHR num="4" nat="TopTh" amorce="0"><TOPTH morph="SNdef_FS" redeno="0" RepriseDuTitre="0" maillon="0">La restriction budgétaire </TOPTH><PRED MotFreq="terrorisme" RepriseduTitre="0">a tout de suite cédé la place à la nécessité de lutter contre le terrorisme .</PRED></PHR></P>

<P num="7" nat="text"><PHR num="1" nat="TopTh" amorce="0"><INIT pos="1" morph="SP_" fct="CIRC" sem="tps" ana="0" redeno="0" RepriseDuTitre="0">Après quatre années d' excédents fédéraux, </INIT><TOPTH morph="SNdef_MS" redeno="0" RepriseDuTitre="0" maillon="0">le budget de 2003 - qui débute en octobre 2002 - </TOPTH><PRED RepriseduTitre="0">renoue avec les déficits .</PRED></PHR><PHR num="2" nat="TopTh" amorce="0"><INIT pos="1" morph="SP_" fct="CIRC" sem="0" ana="0" redeno="0" MotFreq="terrorisme" RepriseDuTitre="0">Sous l' effet conjugué du ralentissement économique et de la lutte contre le terrorisme (les démocrates rajouteraient aussi les baisses d' impôts parmi les facteurs explicatifs), </INIT><TOPTH morph="SNdef_MS" redeno="1" posredeno="_tTopTh_§_Para-0_Phr-1" RepriseDuTitre="0" maillon="0">le budget </TOPTH><PRED MotFreq="lutte" RepriseduTitre="0">devrait afficher un déficit de l'ordre de 43 milliards de dollars .</PRED></PHR><PHR num="3" nat="TopTh" amorce="0"><TOPTH morph="SNdef_MP" redeno="0" RepriseDuTitre="0" maillon="0">Les principaux postes budgétaires </TOPTH><PRED MotFreq="sécurité" RepriseduTitre="0">sont dorénavant la sécurité du territoire (homeland security) et la défense .</PRED></PHR><PHR num="4" nat="TopTh" amorce="0"><INIT pos="1" morph="SP_" fct="CIRC" sem="not" ana="0" redeno="0" RepriseDuTitre="0">Dans le premier cas, </INIT><TOPTH morph="SNdef_MS" redeno="1" posredeno="_tTopTh__Para-0_Phr-2" RepriseDuTitre="0" maillon="0">le budget </TOPTH><PRED MotFreq="lutte" RepriseduTitre="0">passe de 15 milliards de dollars à 38 milliards, une part non négligeable (un peu moins de trois milliards) étant consacrée à la lutte contre le bioterrorisme .</PRED></PHR><PHR num="5" nat="TopTh" amorce="0"><INIT pos="1" morph="SP_" fct="CIRC" sem="not" ana="0" redeno="0" RepriseDuTitre="0">A un niveau institutionnel, et plus seulement fonctionnel, </INIT><TOPTH morph="SNdef_MS" redeno="1" posredeno="Pred__Para-2_Phr-9" RepriseDuTitre="1" maillon="0">l' administration Bush </TOPTH><PRED RepriseduTitre="0">a décidé de renforcer considérablement les polices locales, pompiers, et services d'urgence, qui, tous, constituent la première ligne de défense vis-à-vis des attaques terroristes .</PRED></PHR><PHR num="6" nat="TopTh" amorce="0"><TOPTH morph="NoSe" redeno="0" RepriseDuTitre="0" maillon="0">Environ 3,5 milliards de dollars - soit une multiplication par dix des financements antérieurs - </TOPTH><PRED MotFreq="autorité" RepriseduTitre="0">sont ainsi destinés aux autorités locales, municipales et étatiques, c' est-à-dire aux échelons politiques responsables de ces différents corps .</PRED></PHR><PHR num="7" nat="TopTh" amorce="0"><INIT pos="1" morph="SP_" fct="CIRC" sem="not" ana="0" redeno="1" posredeno="Pred__Para-0_Phr-2" RepriseDuTitre="0">En ce qui concerne la défense, </INIT><TOPTH morph="SNdef_MS" redeno="1" posredeno="Pred__Para-2_Phr-18" RepriseDuTitre="0" maillon="0">le Secrétaire, Donald Rumsfeld , </TOPTH><PRED RepriseduTitre="0">se trouve maintenant à la tête du second poste dans le budget fédéral .</PRED></PHR><PHR num="8" nat="TopTh" amorce="0"><TOPTH morph="SNdef_MS" redeno="1" posredeno="_tTopTh__Para-1_Phr-10" RepriseDuTitre="0" maillon="0">Le Président </TOPTH><PRED MotFreq="milliard" RepriseduTitre="0">a obtenu une rallonge budgétaire de 48 milliards de dollars, soit une enveloppe qui dépasse le montant du budget militaire annuel de n' importe quel autre pays dans le monde .</PRED></PHR><PHR num="9" nat="TopTh" amorce="0"><TOPTH morph="SNdef_MS" redeno="0" RepriseDuTitre="0" maillon="0">L' effort ainsi consenti </TOPTH><PRED RepriseduTitre="0">est comparable à celui engagé par Truman lors de la Guerre de Corée .</PRED></PHR><PHR num="10" nat="TopTh" amorce="0"><INIT pos="1" morph="SP_" fct="CIRC" sem="0_p" ana="0" redeno="0" RepriseDuTitre="0">Comme il y a cinquante ans, </INIT><TOPTH morph="SNdef_MP" redeno="0" RepriseDuTitre="0" maillon="0">les Etats-Unis </TOPTH><PRED MotFreq="milliard" RepriseduTitre="0">sont véritablement entrés dans un budget de guerre : celui -ci devrait atteindre 396 milliards de dollars en 2003, et, si les prévisions se concrétisent, se chiffrer à 470 milliards en 2007 .</PRED></PHR></P>

<P num="8" nat="text"><PHR num="1" nat="TopTh" amorce="0"><INIT pos="1" morph="SP_" fct="MODA" sem="0" ana="0" redeno="0" RepriseDuTitre="0">A priori, </INIT><TOPTH morph="SNdef_MS" redeno="1" posredeno="_tTopTh__Para-1_Phr-6" RepriseDuTitre="1" maillon="0">l' administration Bush </TOPTH><PRED RepriseduTitre="0">a adopté des dispositions budgétaires qui la placent en décalage par rapport à les discours républicains en faveur de la modestie budgétaire et de la nécessaire rigueur dans les dépenses .</PRED></PHR><PHR num="2" nat="TopTh" amorce="0"><INIT pos="1" morph="SP_" fct="CIRC" sem="not" ana="1" redeno="0" RepriseDuTitre="0">Dans ce domaine, </INIT><TOPTH morph="SNdef_MS" redeno="0" RepriseDuTitre="0" maillon="0">l' Etat fédéral </TOPTH><PRED RepriseduTitre="0">a bénéficié d' une nouvelle marge de manoeuvre, inespérée au vu de l' orientation politique de l' équipe dirigeante .</PRED></PHR><PHR num="3" nat="Autre" amorce="0"> C' est d' autant <FOCUS morph="NoSe" redeno="0">plus vrai que ces mesures</FOCUS> ne sont pas précisément des décisions sur lesquelles l' administration se serait engagée à revenir .</PHR><PHR num="4" nat="TopTh" amorce="0"><INIT pos="1" morph="SP_" fct="ORGA" sem="0" ana="0" redeno="0" RepriseDuTitre="0">au contraire, </INIT><TOPTH morph="SNdef_FS" redeno="0" RepriseDuTitre="0" maillon="0">la Présidence </TOPTH><PRED MotFreq="terrorisme" RepriseduTitre="0">a, dans un second temps de sa lutte contre le terrorisme, élaboré un cadre plus général qui cherche à pérenniser les décisions prises à l' automne .</PRED></PHR><PHR num="5" nat="TopTh" amorce="0"><TOPTH morph="SNdef_MS" redeno="0" RepriseDuTitre="0" maillon="0">L' accroissement des pouvoirs de l' Etat fédéral </TOPTH><PRED RepriseduTitre="0">ne tient pas de l' accident de parcours .</PRED></PHR><PHR num="6" nat="Present_SP" amorce="0"> Il s' agit <FOCUS morph="SP" MotFreq="pouvoir" redeno="1"

posredeno="_tInit1__Para-0_Phr-2">au contraire d' une priorité des pouvoirs publics .</FOCUS>pouvoir</PHR></P>
</S>
<S niveau="1" num="3">
<TITRE niveau="1" TitreRepris="3"><PHR num="1" nat="TopTh" amorce="0"><TOPTH morph="SNdef_MS" redeno="0" RepriseDuTitre="0" maillon="0">L' élaboration d' un cadre de lutte contre le terrorisme </TOPTH><PRED RepriseduTitre="0"></PRED></PHR></TITRE>
[...]
</S>
</TEXTE>
</DOCUMENT>