Une pragmatique à fleur de texte :

approche en corpus de l'organisation textuelle

Mémoire présenté pour l'obtention d'une
Habilitation à Diriger des Recherches
Spécialité : linguistique

Marie-Paule Péry-Woodley

Université de Toulouse-Le Mirail

2000

Table des matières

Liste des tableaux

Liste des figures

Remerciements

C’est à travers les collaborations et les amitiés nouées dans les divers lieux où je me suis installée, pour quelques mois ou quelques années, que les travaux présentés ici ont pris forme. Pour la période antérieure à mon arrivée à Toulouse, je tiens particulièrement à remercier :

Teresa O’Brien, Tony James et Terry Lewis (University of Manchester) ;

Dick et Joan Allwright (University of Lancaster) ;

Knud Lambrecht (University of Texas at Austin) et Wallace Chafe (University of California at Santa Barbara), rencontrés lors d’un séjour à Berkeley ;

Marie Hayet, Kristina Jokkinen, Jock McNaught, Juan C. Sager (Centre for Computational Linguistics, University of Manchester Institute of Science and Technology) ;

Lydia Nicaud, Violaine Prince, Gérard Sabah, Anne Vilnat (Equipe Langage et Cognition, Laboratoire d’Informatique pour la Mécanique et les Sciences de l’Ingénieur – UPR A3251-CNRS) ;

A Toulouse, j'ai trouvé un très bel accueil au sein de l'Equipe de Recherche en Syntaxe et Sémantique, mon équipe de rattachement, et dans l'atelier Texte et Communication du Pôle de Recherche en Sciences Cognitives de Toulouse. Je remercie tous ceux qui y ont contribué, et tout particulièrement :

Elsa Pascual : travailler avec elle a été une joie totale. Je lui voue amitié et gratitude.

Andrée Borillo, qui m'a donné de son savoir et de son expérience sans compter. Je la remercie pour son soutien, ses encouragements, et pour sa conception de notre métier.

Anne Condamines pour son accueil dès mes premiers jours à Toulouse, et pour la réflexion partagée sur le rôle des corpus.

L'ERSS est un lieu propice aux échanges d'un domaine de recherche à un autre, d'une génération à l'autre : merci à Laure Sarda, à Josette Rebeyrolle, à Hélène Miguet, à Cécile Fabre, à Anne Le Draoulec, avec qui je partage bien plus qu'un bureau. Merci aussi à Francis Cornish et à Marc Plénat.

Au sein de l'atelier Texte et Communication, avec Jacques Virbel, Christophe Luc et Mustapha Mojahid, j'ai le plaisir de pouvoir continuer le travail commencé avec Elsa. Je les en remercie, ainsi que Jean-Luc Nespoulous et Claudine Garcia-Debanc.

Le dialogue avec Benoît Habert, entamé à Manchester en 1982, traverse les périodes.

Merci à Michel Charolles d'avoir accepté d'être rapporteur, à Christian Jacquemin et à Douglas Biber d'avoir bien voulu être au jury.

Préambule

Ce mémoire regroupe des travaux motivés et influencés par les divers environnements de recherche et d'enseignement qui ont jalonné mon itinéraire : c'est le goût des langues et de la description linguistique qui m'a d'abord conduite en Irlande, à l'Université de Galway, pour une étude des syntagmes prépositionnels calqués sur le gaélique dans l'anglais de la région. Après un détour passionnant du côté de la neuropsycholinguistique à l'Université de Keele (Grande-Bretagne), ma carrière de chercheur en linguistique a réellement débuté au Département d'Etudes Françaises de l'Université de Manchester, à travers des interrogations liées à l'enseignement du français langue étrangère. Ces questionnements ont débouché sur une thèse à l'Université de Lancaster, consacrée à la cohérence textuelle et aux marques de structuration du texte, et fortement infléchie par un séjour à l'Université de Californie à Berkeley. Mon orientation vers les linguistiques du discours, et mon intérêt pour la perspective cognitive, se confirment ensuite, mais avec un biais croissant vers la modélisation et les applications informatiques. Une réorientation grâce à une formation en Sciences Cognitives à l'Université de Manchester va me permettre de me transférer dans un environnement propice, au Centre de Linguistique Informatique de l'Institut de Science et de Technologie (UMIST) de cette même université. La mise en place de ces thématiques à l'Université de Toulouse-Le Mirail me donne l'occasion, en 1994, de rejoindre le Département de Sciences du Langage et de contribuer à la mise en place d'une filière Traitement Automatique des Langues (1999). L'Equipe de Recherche en Syntaxe et Sémantique m'accueille doublement : mon expérience de l'analyse de corpus et des outils informatiques d'analyse et de catégorisation de textes me permettent de trouver ma place au sein de l'opération *Traitement Automatique des Langues : terminologie et organisation conceptuelle* (maintenant *Sémantique et corpus : méthodes, outils et applications*) ; la visée de mes travaux justifie mon appartenance à l'opération *Discours*.

Ce parcours passablement bigarré m'a fait rencontrer des situations professionnelles, des cultures universitaires, des influences intellectuelles très diverses, qui se reflètent dans les thèmes de mes travaux et dans les approches adoptées. À travers cette diversité, toutefois, il n'est pas difficile d'identifier cinq axes principaux qui organisent les recherches présentées ici sur le plan des thématiques, de la méthodologie, des objectifs :

1) Le thème commun est l'identification de marques formelles de structuration des textes. Entre les contraintes syntaxiques qui pèsent sur la rédaction et les impondérables d'ordre pragmatique qui infléchissent l'interprétation, quels sont les outils dont dispose le scripteur pour guider le lecteur dans la construction d'un modèle interprétatif structuré ? Choix lexicaux, agencements syntaxiques, configurations dispositionnelles sont envisagés dans la perspective de leur fonctionnement en situation de communication, et donc de l'impact de ces choix de surface sur la fonction discursive. Une préoccupation centrale et constante est de faire le lien entre des "micro-fonctionnements" linguistiques et un niveau global, à la fois en ce qui concerne l'unité envisagée – le texte –, et la perspective adoptée – une optique fonctionnelle en discours.

2) Cette démarche fonctionnelle se distingue cependant de celle qui constitue la base de nombreux travaux sur les marqueurs discursifs, dans la mesure où ceux-ci prennent comme point de départ un ou des marqueurs – lexicaux pour la plupart – pour en étudier le fonctionnement en discours. Mon approche a plutôt été d'identifier une fonction – ou relation – et de rechercher les marqueurs pouvant y être associés. En effet, s'il est souhaitable de comprendre le fonctionnement de *mais* ou *donc*, on ne peut ignorer qu'il est possible dans de nombreux cas de "faire ce qu'ils font" avec un autre connecteur, ou sans connecteur du tout. Par ailleurs, me distanciant des approches qui privilégient le "tout-pragmatique" et minimisent le rôle des marques formelles, je pars d'une double hypothèse : a) que l'organisation des textes fait bien l'objet d'une signalisation explicite ; b) que celle-ci est identifiable à condition de ne pas limiter son regard aux marques lexicales, mais d'envisager au contraire des configurations qui conjuguent des marques lexicales, syntaxiques, ponctuationnelles, dispositionnelles.

3) La centralité des analyses de corpus représente une troisième constante. Ce choix de mode de constitution d'observables s'explique d'abord parce que le fonctionnement au plan du texte est sans doute plus difficilement accessible à l'intuition que le fonctionnement au plan de la phrase. Et il ne s'agit pas seulement d'étendre l'unité d'analyse mais surtout d'aborder le texte comme unité fonctionnelle. La perspective discursive me semble nécessiter le travail sur corpus dans la mesure où elle s'attache à élucider des fonctionnements linguistiques en contexte. Que les corpus soient formés de textes produits dans des conditions expérimentales comme dans la Partie I ou de textes préexistants comme dans la Partie II, je mets l'accent sur la nécessité de travailler sur des corpus situés, de manière à ce que les marques de structuration textuelle puissent être mises en relation avec une situation d'énonciation. L'analyse de corpus fait également partie intégrante de l'approche fonctionnelle définie au point 2), dans la mesure où il s'agit de modéliser sur le plan linguistique, par des méthodes exploratoires, des fonctions ou des relations identifiées dans des textes.

4) L'approche discursive et le travail empirique sur corpus aiguisent tout naturellement la conscience de la variation dans les réalisations langagières. Dans la mesure où les marqueurs recherchés ont trait au fonctionnement discursif, et non au système de la langue, on peut s'attendre à ce qu'ils varient en fonction de paramètres liés au genre discursif (visée discursive, relations entre les participants, canal, etc.). Le domaine d'expérience sera également pris en compte, et, pour les analyses mettant en jeu des langues différentes, la possibilité de fonctionnements différents de marqueurs "équivalents", ou de normes rhétoriques différentes.

5) Finalement, mes travaux abordent régulièrement des questions linguistiques reliées à des problématiques applicatives. Ce choix d'orientation a une double origine : l'intérêt

pour les études linguistiques ancrées dans des situations concrètes certes, mais aussi la conviction que cette "intrusion" de problématiques externes fait avancer la pensée linguistique en obligeant à jeter des ponts entre modèles et approches. Les domaines d'application, qui ont évolué en lien avec mes différents environnements professionnels, ont tous trait à l'organisation des textes et à sa signalisation formelle :

– enseignement et évaluation de l'écrit en langue étrangère ;

– apport de l'analyse des modes de construction de textes pour la modélisation des connaissances du scripteur dans le cadre d'un système tuteur intelligent ;

– modélisation de la structuration et de la signalisation d'objets textuels pour la recherche d'information, l'acquisition de connaissances à partir de documents, ou pour la génération automatique.

La première partie de cette synthèse est consacrée à l'élaboration et à l'illustration d'un cadre d'analyse pour des textes entiers, principalement produits dans le contexte de l'apprentissage de l'écrit. Dans la deuxième partie, les questionnements sur les niveaux de structuration des textes et le jeu des marques formelles se poursuivent avec la mise en relation de nouveaux modèles et la focalisation sur un objet textuel – la définition – dans des corpus de textes scientifiques ou techniques. La troisième partie prend la forme d'une réflexion sur trois thèmes doublement significatifs. Ce sont ceux qui à la fois parcourent l'ensemble des travaux présentés et motivent mes chantiers actuels et mes projets : d'abord, les niveaux d'organisation textuelle et leur articulation, ensuite la notion de marqueur, et pour finir, en relation avec la méthodologie d'analyse de corpus, les notions de variation et de genre discursif.

Partie I

Structures de texte : modèles et marqueurs

Chapitre 1

Aborder le texte

1.1 Texte, discours et pragmatique

Au départ, ce sont des questions motivées par des problèmes d'ordre pédagogique qui m'ont amenée à vouloir travailler sur le texte. Comment analyser de façon pertinente, de manière à guider utilement les apprenants, des productions écrites acceptables sur le plan grammatical mais difficilement interprétables en tant que textes ? Un exemple, qui sera traité dans la section 2.2.2, illustrera le problème :

> *Il y a certains qui sont tout à fait contre cette raison que donne le gouvernement, surtout celui de la Grande-Bretagne. On dit que le gouvernement paie bien d'autres choses /.../. On dirait que ce n'est pas aussi important que l'enseignement, /.../ On dit que les recherches nucléaires sont inutiles /.../. /.../*
>
> *Cependant, il y a d'autres qui disent qu'une expansion (...) est tout à fait impossible parce que ça coûte trop cher. On dirait que ce n'est pas juste de dépenser beaucoup d'argent ...*

Mon analyse de ce texte concernera principalement l'impact de l'usage répété de *on* sur sa structuration thématique.

M. Charolles (1978) exprime bien la frustration ressentie par les enseignants – en français langue maternelle en l'occurrence – devant l'inadéquation des corrections qu'ils inscrivent, faute de mieux, dans les marges des copies d'élèves. Ces corrections ponctuelles laissent complètement de côté les erreurs ou maladresses sur le plan de la "mise en texte", que les enseignants perçoivent pourtant comme constituant les vrais problèmes. Selon M. Charolles, les rares commentaires qui ont trait à ce niveau de fonctionnement (ex. "coq à l'âne") sont beaucoup trop imprécis pour pouvoir guider les apprenants, et manifestent bien la pénurie d'outils conceptuels et analytiques pour aborder le texte. La linguistique, qui commence seulement dans les années 70 à s'extraire de sa focalisation sur le fonctionnement syntaxique, ne s'est en effet pas beaucoup préoccupée de rendre compte de l'interphrastique et encore moins des régularités structurelle des textes ou de leur signalisation. Les théories du texte issues des études littéraires ne sont pas d'un grand secours. Les "grammaires du texte" non plus, qui se situent dans une perspective interprétative peu concernée par les choix de formulation au ras du texte. Au delà des enjeux applicatifs s'ouvrent cependant des questions linguistiques fondamentales sur les liens entre langue et parole, syntaxe et discours.

Dans son article de 1978, M. Charolles pose l'existence d'une compétence textuelle qui sous-tend la production de textes "bien formés" et les jugements que l'on peut porter sur la cohérence des textes. Plus tard (1983; 1986; 1995), il s'intéressera davantage à la dimension pragmatique. La linguistique du texte se rattache en effet inévitablement aux linguistiques du discours. Les régularités formelles que l'on cherche à identifier dans la signalisation de la mise en texte ne sont pas entièrement déterminées par le système de la langue, les méthodes mises au point pour l'analyse syntaxique – recours à l'intuition pour la construction d'exemples, formulation de jugements de grammaticalité ou d'acceptabilité – risquent fort de s'avérer inadaptées. L'approche ne peut être que fonctionnelle et contextuelle, intéressée par l'occurrence régulière de telle forme pour remplir telle fonction en discours. La référence à la pragmatique risquerait pourtant de conduire à un paradoxe : en mettant en avant le rôle des apports externes au texte dans l'interprétation, la pragmatique a tendance à réduire l'importance des choix de formulation. Ainsi, directement ou indirectement influencés par l'impact des propositions de H.P. Grice (1975), et s'opposant à l'idée d'une cohérence fondée sur la cohésion, de nombreux auteurs (Charolles, *op.cit*; Reinhart, 1980; Cornish, 1986; Mann et Thompson, 1988; *inter alia*) posent un principe premier de cohérence, qu'ils identifient comme ce qui oriente le lecteur vers une lecture cohésive des marques dans les textes. Cette perspective pragmatique, fondamentale dans la compréhension des fonctionnements discursifs, peut toutefois s'accompagner du risque de détourner les recherches de l'examen précis des choix de réalisation linguistique. Tout en intégrant les acquis de la pragmatique linguistique, les travaux présentés ici se focalisent, pour le revaloriser, sur le détail de la surface des textes. Ils cherchent à élucider comment des choix d'agencements de constituants qui ne sont pas déterminés par la syntaxe sont contraints par des facteurs contextuels et vont orienter différemment la construction de l'interprétation. Déjà en 1981, R. de Beaugrande et W. Dressler s'insurgeaient contre la tendance à minimiser les choix de réalisation linguistique :

> We must guard against allowing the text to vanish away behind mental processes. Recent debates over the role of the reader point up to the dangers of assuming that text receivers can do whatever they like with a presentation. If that notion was accurate, textual communication would be quite unreliable, perhaps even solipsistic. There must be definitive, though not absolute, controls on the variations among modes of utilising a text by different receivers. (de Beaugrande et Dressler, 1981:35)

Ce sont ces "moyens de contrôler les variations d'interprétation" que je vais m'efforcer d'identifier et de caractériser.

1.2 L'unité texte

Il est temps de préciser la notion de "texte", en particulier dans sa relation avec celle de "discours". Dans une définition récente, F. Cornish distingue les deux notions en termes clairs :

> (…) I view *text* (as a non-count noun) as denoting a typical instance of language *cum* other semiotic devices in use – i.e. occurring in some context and with the intention by the user of achieving some purpose or goal thereby. The term designates the connected sequence of verbal signs and non-verbal signals, vocal as well as non-vocal (i.e. visual, auditory, etc.) signals produced within the context of some utterance act. (…)
>
> *Discourse*, on the other hand, designates the hierarchically structured, mentally represented sequences of utterance and indexical acts which the participants are engaging in as the communication unfolds. Such sequences have as their *raison d'être* the accomplishment of some particular overall communicative goal (…) (Cornish, 1999:33-34).

Je conçois moi aussi le texte comme la trace – signes verbaux et signaux non-verbaux – d'un discours ancré dans un contexte. Le texte écrit, sur lequel se concentrent mes

travaux, présente par rapport à l'oral des spécificités qui méritent d'être notées dès à présent, même si elles font l'objet de plus amples développements par la suite :

– la première de ces spécificités est la *distanciation.* L'écrit, communication distanciée, implique presque toujours que les participants ne partagent pas le contexte. La "co-construction" du discours se fait donc en deux temps distincts. Les représentations que le scripteur[1] se fait du lectorat visé sont certes déterminantes dans la façon dont il gère "l'interaction", mais il n'en reste pas moins qu'il est seul maître à bord pendant la rédaction. La situation se renverse lors de la lecture. L'absence de contexte partagé implique dans la plupart des cas une exigence d'explicitation de la signalisation des différents niveaux de structuration. Le texte écrit, trace d'un acte de communication, doit comporter en lui-même les éléments qui, dans des contextes divers, permettront de recréer *du discours*.

– la deuxième spécificité, qui découle de la première, est le caractère *monologal* de la plupart des écrits. Il entraîne que l'introduction, la continuation ou l'abandon des thèmes se fait non pas par négociation entre les participants du discours mais sur la seule base des représentations et des intentions du scripteur. C'est donc au texte qu'incombe la tâche de guider le lecteur dans cet aspect de l'interprétation.

– finalement, les textes écrits sont des réalisations langagières matérialisées, inscrites sur un support, qu'il soit papier ou support informatique. Ce sont des objets *visuels* dont les propriétés visuelles sont partie prenante dans la construction du sens. Ce sont ces propriétés visuelles qui constituent à l'écrit les "signaux non-verbaux" de la définition de F. Cornish.

Par ailleurs, j'entends aussi par *texte* l'unité d'analyse par rapport à laquelle les unités constituantes seront envisagées. Il s'agit d'une unité fonctionnelle, et non pas formelle, bien définie par M. Halliday et R. Hasan comme "a unit of language in use" (1976:1). L'intérêt porté par la linguistique du texte aux relations entre segments textuels ne doit pas amener à confondre linguistique du texte et linguistique de l'interphrastique. D'une part parce que la linguistique du texte est concernée par des relations entre segments dépassant la phrase, mais aussi parce qu'un texte peut ne comporter qu'une phrase ou même qu'un mot. Ainsi, le mot *stop* peut constituer un texte, et permettre la reconstitution d'un acte énonciatif, si certaines conditions sont satisfaites. Ces conditions qui lui permettront de fonctionner comme "a unit of language in use" ont trait ici aux aspects visuels du texte et au contexte. En vertu de cette caractérisation fonctionnelle, un texte constitue un tout qui se tient, une "séquence connectée" dans la définition ci-dessus. La recherche des corrélats linguistiques de cette qualité de cohérence constitue un des enjeux principaux des travaux relevant de la linguistique du texte.

1.3 Outils pour aborder l'organisation textuelle

Un des auteurs qui a sans doute le plus influencé l'approche de l'unité texte est M. Halliday (1967/68; 1980; 1985; Halliday & Hasan, 1976). M. Halliday construit son modèle du système linguistique autour de trois composantes fonctionnelles-sémantiques, appelées métafonctions : les composantes *idéationnelle*, *interpersonnelle* et *textuelle*. Il en donne les définitions suivantes :

– la composante *idéationnelle* : la partie du système linguistique concernée par l'expression du "contenu" ; parmi les fonctions du langage, celle d'être "au sujet de" quelque chose. C'est le domaine de l'expérience, des relations logiques. Le "locuteur"[2] est vu dans son rôle d'*observateur*.

[1] C'est par souci de simplicité que j'écris "scripteur"; il est clair que la production de textes implique souvent plusieurs acteurs.

[2] Halliday s'intéresse principalement à la langue parlée.

– la composante *interpersonnelle* : les fonctions sociale, expressive et conative du langage. C'est le domaine des attitudes, des jugements, des relations de rôles en situation, des intentions qui sous-tendent le discours. Le locuteur est vu dans son rôle d'*intrus* ("intruder").

– la composante *textuelle* : la composante du système linguistique qui a trait à la construction de textes, concernée par les ressources langagières permettant de créer des textes. Le texte est défini comme une unité pertinente sur le plan opérationnel, cohérente en elle-même et par rapport au contexte situationnel[3]

Cohesion in English (Halliday & Hasan, 1976) est un ouvrage qui cherche de façon explicite et systématique à faire le lien entre des caractéristiques formelles de la surface textuelle et la qualité globale de cohérence, et fournit un examen approfondi des notions très productives de *cohésion* et de *texture* et de leurs réalisations linguistiques. A travers la notion de cohésion, définie comme "the means whereby elements that are structurally unrelated to one another are linked together, through the dependence of one on the other for its interpretation", les auteurs élaborent leur conception de la composante textuelle, et proposent un inventaire des ressources linguistiques pour créer du texte. Ces ressources se déclinent en cinq classes d'éléments : la *référence* (anaphores et cataphores), la *substitution lexicale*, l'*ellipse*, la *conjonction* (coordinateurs et connecteurs), la *cohésion lexicale*. Ce sont ces éléments qui, par les liens qu'ils établissent entre segments, confèrent au texte sa texture.

Le rôle de ces procédés de cohésion dans la construction de sens à partir d'un texte a fait et fait encore l'objet de nombreuses recherches. Il est impossible toutefois de s'en tenir à la cohésion pour expliquer la cohérence. D'abord parce que, comme l'ont montré de nombreux auteurs, l'interprétation des procédés de cohésion dépend elle-même du modèle interprétatif global en cours de construction : ce serait donc la cohésion qui dépendrait de la cohérence et pas l'inverse. Ensuite parce que *texture* n'est pas *structure*[4] : certes, la référence et la cohésion lexicale sont pertinentes par rapport à la structuration des référents, la conjonction par rapport aux relations entre segments, mais les travaux sur la cohésion ne constituent pas un modèle de l'organisation des textes.

Cette première partie prend donc comme un acquis les résultats des travaux sur la texture pour se concentrer sur la structuration textuelle, à travers un double examen des marques de mise en texte :

1) produire un texte, c'est évoquer des *objets de discours* (cf. Berthoud et Mondada, 1991) de telle façon qu'ils soient perçus comme thèmes dans le discours. La forme que ces objets et leurs relations donnent au texte constitue sa *structure thématique*. La dimension interactive est primordiale puisque la présentation de ces objets est fonction de la représentation que se fait le scripteur des connaissances et des croyances du lecteur ; la dimension dynamique l'est également puisque ces connaissances et croyances sont appelées à évoluer au fur et à mesure du processus d'interprétation ;

2) produire un texte, c'est aussi structurer des prédications concernant les objets présentés comme thèmes en fonction des objectifs discursifs. Le terme *structure rhétorique* sera utilisé (à la suite de W. Mann et S. Thompson, 1986; 1988) pour désigner deux formes d'organisation des propositions : leur hiérarchisation et le réseau de relations, sémantiques, pragmatiques, et textuelles qui les unissent.

---

[3] Bien que sans doute inspirée de celle formulée par les linguistes de l'Ecole de Prague (structure grammaticale des phrases, structure sémantique des phrases, organisation de l'énoncé), cette tripartition s'en distingue par le fait que l'unité envisagée est le texte et non la phrase.

[4] J'oppose ici *texture* et *structure de texte*, et non *texture* et *structure de phrase*, comme le font M. Halliday and R. Hasan (1976) pour distinguer ce qui est du ressort de la syntaxe de la phrase (structure) de ce qui a trait au texte.

Les choix formels à ces deux niveaux de fonctionnement peuvent être appréhendés (après W. Chafe, 1976) comme des choix de présentation, des "emballages" motivés, contraints par l'interaction : "emballage" de l'information en phrases et "emballage" des phrases en texte[5].

1.4 Domaine de l'étude et corpus

Les termes *mise en texte*, *production*, *interprétation*, *interaction* qui sont intervenus à plusieurs reprises dans cette présentation pourraient conduire à penser que j'ai poursuivi une approche de type psycholinguistique des processus de rédaction ou d'interprétation des textes. En fait, ce que j'ai cherché à faire, c'est examiner les textes en tant que traces du processus de rédaction et ensembles de signaux pour l'interprétation. Il s'agissait donc, dans une perspective discursive, d'identifier et de caractériser les choix de formulation liés à la mise en texte. Le plus gros du travail était de se donner les outils théoriques et méthodologiques, dans ce domaine encore balbutiant, pour aborder l'analyse. S'ajoutait à cet objectif, pour l'essentiel des travaux présentés dans cette Partie I, une visée contrastive motivée par mon insertion dans un département d'études françaises en Grande-Bretagne : en quoi ces choix de formulation significatifs des stratégies de mise en texte diffèrent-ils en langue maternelle et en langue étrangère, en français et en anglais ? J'ai donc opté pour un corpus expérimental qui me permettait de mieux contrôler la comparabilité des textes. Le corpus *Étudiants* se compose de trois séries de textes (tableau 1.1) :

Sous-corpus	**Langue**	**Sujets**	**Taille**
ALM	anglais langue maternelle	étudiants anglophones en 2ème année d'études de français dans une université britannique	15 textes ≅ 9 000 occ.
FLE	français langue étrangère	les mêmes étudiants que ALM	15 textes ≅ 6 500 occ.
FLM	français langue maternelle	étudiants francophones en 2ème année de LEA dans un Institut Universitaire de Technologie français	13 textes ≅ 6 000 occ.

Tableau 1.1 : Le corpus Étudiants.

Le recueil des données s'est fait dans des conditions semblables pour les trois séries et à partir du même matériau (auquel je ferai référence dans ce qui suit par la dénomination *matériau Étudiants*) :

– quatre tableaux simples (en anglais pour ALM, en français pour FLM et FLE) fournissant pour sept pays des données chiffrées sur quatre points (ex. le taux de scolarisation des 19-24 ans, le pourcentage du PIB consacré à l'enseignement supérieur) ;

– un sujet (en anglais pour ALM, en français pour FLM et FLE) qui exigeait explicitement une argumentation s'appuyant sur les données statistiques fournies.

A ce corpus de référence s'ajoute un corpus en anglais issu d'une expérience pédagogique dans le cadre de l'enseignement de l'écrit "universitaire" (*academic writing*) : des textes produits en anglais par des étudiants étrangers à partir d'une série de propositions

[5] Le terme *emballage de l'information* traduit la formule *information packaging* empruntée à W. Chafe, qui la définit ainsi : "The kind of phenomena at issue here (...) have to do primarily with how the message is sent and only secondarily with the message itself, just as the packaging of toothpaste can affect sales in partial independence of the quality of the toothpaste inside" (Chafe, 1976:28).

ont été reformulés par des locuteurs natifs dans le but d'améliorer la lisibilité du texte en en respectant au maximum les intentions (cf. Allwright *et al.,*1988). Les références à ce corpus (corpus *Reformulation*) dans ce qui suit seront purement illustratives.

Les textes dont sont extraits les exemples illustrant les analyses sont donnés *in extenso* dans les annexes 1 (corpus Étudiants) et 2 (corpus Reformulation).

Chapitre 2

Construire ce dont on parle : la structure thématique

Mon objectif était de cerner suffisamment la fonction de thème[6] pour pouvoir identifier des marques formelles associées à cette fonction. Je vais résumer les grandes lignes du cadre élaboré à partir des travaux existants, pour me concentrer sur les analyses réalisées dans ce cadre.

2.1 La paire thème-rhème

La notion de thème est d'autant plus problématique que le terme est utilisé à la fois comme mot de la langue courante et dans un sens technique par les linguistes. Notion relationnelle, le thème ne peut être envisagé que par rapport à un segment. Pour la langue courante, pour laquelle le thème est "ce sur quoi porte le texte", celui-ci est généralement le texte dans son ensemble. Pour les linguistes, l'unité de référence est la proposition ou la phrase. En ce qui me concerne, dans la mesure où je m'intéressais au marquage syntaxique du thème, et en particulier au rôle de la syntaxe complexe, j'ai opté pour une troisième unité : l'*unité syntaxiqu*e, constituée par une proposition principale plus toute proposition subordonnée ou proposition réduite qui s'y trouve attachée ou enchâssée[7]. On verra que la relation entre le thème local des linguistes et le thème global de la langue courante ne va pas de soi...

Dans la littérature, le thème fait l'objet d'une caractérisation en termes des aspects suivants :

---

[6] Je m'en tiens ici au terme *thème*, qui était celui utilisé dans les travaux résumés dans ce chapitre. Au chapitre 6, je reprendrai cette notion dans un cadre différent et sous le terme de *topique*.

[7] Cette définition correspond à celle donnée par W. Hunt pour le *T-unit* ou *minimal terminable unit* (cf. 2.3.1) : "one main clause plus any subordinate clause or non-clausal structure that is attached to or embedded in it" (Hunt, 1970 : 4)

1) sa fonction : le thème est ce sur quoi porte la proposition (*aboutness* ou *à propos*, cf. Berthoud (1996) pour le terme français), le point de départ de l'énoncé (Halliday 1967/8; van Dijk 1977; Reinhart 1980; Lambrecht 1981; 1994) ;

2) la nature de son référent : le thème a un référent donné, connu, disponible (Chafe 1970; 1976; Prince, 1981; Lambrecht, 1987; 1994; Gundel, 1985) ;

3) sa position dans la proposition ou la phrase : la position initiale, et *a fortiori* la dislocation à gauche, sont intimement liées à la mise en place du thème (Halliday 1967/8; Givón, 1979; Lambrecht 1981; 1994; Gundel, 1985).

La question se pose du statut de ces différentes caractérisations et de leurs relations. M. Halliday (1967) s'insurge contre l'amalgame souvent fait entre les deux premières, et donne l'exemple suivant, dans lequel le sujet *JOHN* est thème, selon sa définition, sans qu'il puisse être perçu comme donné :

(1) JOHN saw the play yesterday.

énoncé dans un contexte de contradiction, par exemple en réponse à :

(1') Nobody saw the play yesterday, did they?

ou en réponse à :

(1") Tony saw the play yesterday.

Mais son argumentation est liée au fait qu'il définit au départ le thème comme l'élément initial. Pour K. Lambrecht (1982; 1994), qui met en avant l'*à propos*, *JOHN* dans l'exemple précédent n'est pas un thème (*topic* dans sa terminologie), mais bien un constituant du rhème (*focus*). Ce statut est indiqué par le statut prosodique de *JOHN*, noté par les majuscules, dans l'exemple (1). Le français parlé, qui ne permet pas le marquage prosodique du sujet pour changer l'articulation pragmatique de l'énoncé, aurait recours pour introduire un sujet rhématique à une construction qui en ferait d'abord le rhème d'une structure présentationnelle et préserverait ainsi la structure d'information normale, où le donné précède le nouveau (cf. Lambrecht, 1994) :

*(2) **Ya** John qui a vu la pièce hier (en réponse à (1'))*

*(2') **C'est** John qui a vu la pièce hier (en réponse à (1''))*

Notons qu'il s'agit là de langue parlée, et de marques absentes de la langue écrite (prosodie) ou qui en sont pour certaines exclues par la norme (construction présentationnelle en "ya").

Pour K. Lambrecht, dont je reprendrai la définition pour mon étude, le thème (*topic*) a la double propriété d'encoder une relation (ce sur quoi porte la proposition, son domaine) et d'encoder une propriété du référent ("identifiabilité" pragmatique)[8]. Le rhème (*focus*), quant à lui, est défini de façon purement relationnelle, le rhème de l'assertion, la partie la plus saillante de l'information nouvelle (indépendamment du statut du référent). Cette formulation est celle qui va servir de base à mon exploration des marques formelles du thème. Les marques formelles relevées dans la littérature ont principalement trait à l'agencement des éléments dans la proposition, parfois aussi aux propriétés référentielles des lexèmes. Je vais me concentrer sur les agencements, en mentionnant les propriétés référentielles au détour du chemin.

2.2. Marquage du thème

[8] C'est la définition qu'il en donnait dans les publications qui ont informé les travaux présentés ici. Il a apporté de nombreuses modifications et précisions à cette définition dans un important ouvrage ultérieur (Lambrecht, 1994). Ces apports constituent un nouvel éclairage qui sera évoqué à plusieurs reprises au cours de l'exposé et permettra de réévaluer certains de mes résultats.

Ce que j'ai appelé plus haut agencement des éléments recouvre deux types de réalisations : d'abord l'ordonnancement des constituants là où cet ordonnancement est indépendant de la syntaxe, ensuite un petit nombre de constructions syntaxiques "marquées"[9].

2.2.1 Placement des constituants mobiles

• Ordre des constituants et poids informationnel

Pour les linguistes fonctionnalistes, à commencer par ceux de l'École de Prague (voir en particulier J. Firbas, 1972; 1986), le choix de placement des constituants le long de l'axe syntagmatique fait partie des procédés permettant d'influer sur leur poids informationnel. Comparer pour s'en convaincre :

(3) Elle donne une pomme à l'enfant.

(3') Elle donne à l'enfant une pomme.

ou

(4) Il boit pour oublier.

(4') Pour oublier il boit.

Bien sûr la liberté de choix n'est pas totale, ne serait-ce qu'à cause des contraintes syntaxiques, plus ou moins rigides selon les langues. Plusieurs linguistes envisagent l'ordre des mots dans la phrase comme la résultante de forces diverses, parmi lesquelles, outre la syntaxe, la structure sémantique (Firbas, 1972), l'iconicité expérientielle, les contraintes liées au traitement en temps réel dans l'oral spontané (Enkvist, 1985). La plupart s'accordent pour établir un lien fort entre position initiale et établissement explicite d'un thème (Givon, 1979; Lambrecht, 1981; Gundel, 1985), sinon pour sa continuation par un élément anaphorique. En outre, des études psycholinguistiques mettent l'accent sur l'avantage cognitif – rapidité de lecture, rappel – de la présentation des éléments connus avant les éléments nouveaux, c'est-à-dire du thème avant le rhème (Clark & Haviland, 1977; Kieras, 1981)[10].

• Circonstants en début ou en fin de phrase

Ce qui va m'intéresser dans cette optique, c'est d'observer en contexte le cas illustré de façon très artificielle par la paire (4)-(4'), c'est-à-dire le placement initial ou final d'éléments circonstanciels, propositions ou syntagmes prépositionnels. La paire minimale fabriquée pour les besoins de l'illustration donnerait à croire que le placement de la circonstancielle est "libre". C'est là le point de vue grammatical habituel : la place de ces constituants n'étant pas déterminée par la syntaxe, elle est généralement considérée comme relevant du style, de la "parole", et hors du champ de la linguistique. On verra cependant que la perspective fonctionnelle révèle des contraintes tout aussi réelles, mais d'un autre ordre.

Ma démarche est éclairée par une étude de S. Thompson (1985) sur les propositions circonstancielles de but dans un corpus diversifié, qui l'amène à proposer que la fonction discursive de cette construction diffère nettement selon qu'elle est placée en début ou en fin de phrase. Les circonstancielles de but initiales peuvent avoir une portée (*scope*) qui dépasse

---

[9] Je n'élaborerai pas ici la notion de construction marquée, qui a fait l'objet d'un développement dans ma thèse (pp. 59-60).

[10] K. Lambrecht (1994) considère que la position initiale est trop saillante sur le plan cognitif pour n'être exploitée que par une fonction, celle de marquer le thème. Elle est utilisée également en anglais parlé, on l'a vu avec l'exemple (1), pour un rhème marqué (sujet rhématique). Il convient que le principe du thème initial peut être préservé comme principe universel d'ordonnancement s'il est réservé aux expressions thématiques lexicales et pronominales accentuées associées à la fonction d'annonce du thème.

la phrase où elles apparaissent : elles guident l'attention du lecteur en nommant un problème issu d'attentes créées par le texte précédent (ou par inférence à partir de celui-ci), auquel la suite, qui peut consister en plusieurs phrases, fournit la solution. Les circonstancielles de but finales, au contraire, ont une portée limitée à leur proposition principale, elles se contentent de poser le but dans lequel est entreprise l'action nommée dans la principale. L'observation de cette différence, et en particulier du fait qu'une subordonnée peut être associée fonctionnellement avec beaucoup plus que la principale avec laquelle elle forme une phrase ponctuée m'a paru très pertinente dans ma recherche des marques formelles de structuration des textes. J'ai repris cette analyse pour le français, sur les circonstancielles de but d'abord pour l'élargir ensuite à d'autres types, en m'efforçant d'en tirer des conclusions sur le plan de la structure thématique (Péry-Woodley, 1993a; 1996b). Un exemple en français illustrera la conformité du fonctionnement des circonstancielles de but au modèle proposé par S. Thompson :

(5) (extrait d'un mode d'emploi pour l'accès aux services Minitel d'une banque)

[1] ***Pour vous connecter sur le service BNPTEL***
[2] Après avoir allumé votre Minitel :
[3] — Composez sur votre téléphone le 3616.
[4] — Au bip sonore, appuyez sur la touche CONNEXION (...)
[5] — puis raccrochez (...).
[6] — Tapez sur le clavier (...)
[7] — Appuyez alors sur la touche SOMMAIRE
[8] — ***pour créer votre mot de passe****.*
(...)
[9] Vous pouvez ainsi obtenir les opérations effectuées depuis la date d'établissement de votre dernier relevé de compte.
[10] ***Pour toutes les consulter****,*
[11] appuyez sur la touche SUITE
[12] ***afin de visualiser tous les écrans****,*
[13] ou RETOUR
[14] ***pour revenir en arrière****.*

Les propositions ont été numérotées[11] pour faciliter l'analyse. On constate que la portée de la proposition [1] s'étend sur les propositions [2]-[8], qui composent la liste des actions à exécuter pour réaliser le but nommé en [1]. De même, la portée de [10], qui formule le "problème" issu de [9], recouvre les propositions [11]-[14]. A l'inverse, les circonstancielles de but finales [8], [12], [14] ont, comme l'observait S. Thompson pour l'anglais, une portée purement locale, limitée à leur proposition principale. Le rôle du placement à l'initiale de constituants extérieurs à la prédication principale sera repris dans le chapitre 6 avec les expressions introductrices d'univers de discours.

• Saillance thématique et saillance rhématique

S. Thompson (1985) décrit les circonstancielles de but initiales comme des thèmes marqués. Dans cette optique, leurs homologues en finale doivent être considérés comme rhématiques. On a jusqu'à présent envisagé la fonction de thème et de rhème dans le cadre de la proposition, au plus de la phrase. Les observations de S. Thompson sur la portée des circonstancielles initiales, qui semblent s'appliquer aussi au français, font éclater ce cadre pour aborder la question de la structuration textuelle à proprement parler. Je voudrais maintenant les réexaminer en relation avec une notion importante dans le domaine de la

[11] La proposition [13] comporte une ellipse : "RETOUR" est lu comme "appuyez sur la touche RETOUR".

linguistique du texte : la *saillance*. Il s'agit d'une notion cruciale pour les modèles d'organisation textuelle, pour les applications telles que la génération de texte et tout particulièrement pour le résumé automatique, ainsi qu'en didactique de l'écrit. Mais une certaine confusion règne quant à sa définition. Pour les auteurs adoptant une perspective cognitive (Chafe 1976; Prince 1981), est *saillant* un référent que le locuteur suppose présent à l'esprit de l'interlocuteur ; on retrouve là un des aspects du statut "donné" d'un référent, et qui s'applique par conséquent le plus souvent au *thème*. Mais le *rhème* est lui aussi régulièrement caractérisé par la saillance : c'est l'élément nouveau, l'élément le plus informatif. P. Werth (1984) définit le rôle de l'articulation thème-rhème comme étant précisément de concentrer le matériau le plus important à la fin de l'énoncé pour le rendre plus saillant. Pour K. Lambrecht (1981), thème (*topic*) et rhème (*focus*) sont tous deux saillants, mais de façons différentes : le thème dans un sens référentiel, le rhème parce qu'il reçoit l'accent[12]. Ces définitions concernent l'unité phrase. Dans la perspective non plus de la structure thématique de la phrase mais de la construction d'un texte, et en reprenant le cas des circonstancielles à portée étendue, on s'aperçoit que la différence fondamentale entre la *saillance thématique* et la *saillance rhématique*, c'est que la première s'exerce sur le plan du texte alors que la seconde est limitée au plan local. Les circonstancielles de but finales sont saillantes au niveau de la phrase, où elles indiquent le but dans lequel une action est entreprise. Les circonstancielles initiales sont saillantes au niveau du texte parce qu'elles contribuent à lui conférer une structure en rassemblant, grâce à leur portée étendue, un ensemble de segments sous un "drapeau" thématique.

• Circonstants thématiques et visée discursive

Si les propositions circonstancielles de but fonctionnent différemment en début et en fin de phrase, qu'en est-il d'autres éléments circonstanciels, et en particulier des propositions et syntagmes prépositionnels spatiaux et temporels ? Le corpus Étudiants se prêtait à une telle étude puisque la tâche de rédaction qui en était l'origine exigeait qu'une argumentation soit construite à partir de données statistiques, sous forme de tableaux concernant différents pays à différentes périodes (matériau Étudiants, cf. 1.4). Dans un premier temps, j'ai observé le fonctionnement des circonstants initiaux et finaux. Les exemples ci-dessous visent à montrer qu'on y retrouve certains aspects du fonctionnement des circonstancielles de but :

(6) ***En 65 et 70****, c'est en France que le pourcentage est le plus bas, (…) ; mais* ***en 74****, c'est l'Italie qui détient le plus faible pourcentage, suivie du Japon … (FLM-13 : 34-35)*

(7) ***En Grande-Bretagne*** *moins de la population âgée de 19 à 24 ans font l'enseignement supérieur qu'en France (…).* ***Au Japon*** *14,7% de la population âgée de 19 à 24 ans ont été admis au système d'enseignement supérieur … (FLE-8 : 2-4)*

(8) Le pourcentage du PNB consacré aux dépenses publiques dans l'enseignement supérieur n'a que très peu augmenté ***entre 1970 et 1974****. (FLM-11 :7)*

(9) La Faculté semble donc beaucoup plus accessible ***aux USA.*** *(FLM-13 : 4)*

Ce qu'on retrouve dans le cas des circonstants initiaux, c'est la structuration thématique d'un segment de texte d'un empan supérieur à la phrase. Ici toutefois, ce n'est pas grâce à la portée étendue du circonstant initial, mais par le biais de la répétition, de phrase en phrase, du même type de circonstant. On a donc des segments, voire des textes entiers, organisés par des circonstants chronologiques, d'autres structurés par la comparaison

[12] Le corollaire est la confusion terminologique autour du terme *focus*, qui désigne dans la paire *topique-focus* l'élément d'information nouvelle présenté comme le plus important, mais qui est parfois employé pour désigner le référent de discours psychologiquement prééminent. J. Gundel (2000) dénomme le premier "focus sémantique" et le second "focus psychologique". La théorie du centrage (chap. 6 *infra*) échappe à cet imbroglio en utilisant le terme de *center* pour le focus psychologique.

entre pays. Les circonstants finaux ne sont pas typiquement intégrés dans des chaînes, et n'ont pas ce rôle d'organisation textuelle.

Ces observations m'ont conduite à faire l'analyse de tous les éléments circonstanciels initiaux de mon corpus, et à établir un classement en trois grands types : les circonstanciels *lieu/temps* situent la prédication dans un cadre spatial ou temporel (ex. (6)-(7)). Les circonstanciels *inter-texte*[13] lui donnent pour cadre l'un ou l'autre des tableaux statistiques du matériau Étudiants (les tableaux eux-mêmes, et non plus leurs éléments, fournissent le cadre : ex. (10)). Les circonstanciels appelés *condition/concession* sont des conditionnelles ou des concessives qui servent généralement de cadre à l'expression d'une opinion ou d'un argument (marqueurs d'implication, ex. (11)).

(10) ***D'après le premier tableau,*** *nous pouvons constater qu'en général, peu de jeunes de 19 à 24 ans suivent des études supérieures. (...) Le second tableau nous renseigne sur ....* ***En nous référant au quatrième tableau*** *... (FLM-5 :2, 8,18)*

(11) ***If cuts must be made****, then, the government must examine carefully the areas which will be affected (ALM-3 : 20-21)*

Cette fois j'ai également abordé la description de mon corpus de façon quantitative pour donner une image sans doute grossière mais néanmoins révélatrice de l'emploi de ces circonstanciels dans mes trois sous-corpus :

corpus	lieu/temps	inter-texte	condition/ concession
FLM (n=116)	15,5%	11,2%	13,8%
ALM (n=195)	4,6%	3,6%	**30,3%**
FLE (n=167)	**33,5%**	7,8%	10,2%

n = nombre de circonstanciels initiaux

Tableau 2.1: Trois fonctions des circonstanciels initiaux (% de tous les circonstanciels initiaux).

Si les circonstants jouent effectivement un rôle important en organisant des segments de texte autour de thèmes, on constate ici des différences nettes dans l'organisation thématique des trois groupes de textes. Le résultat le plus frappant concerne les corpus ALM et FLE. Rappelons qu'il s'agit des mêmes sujets qui ont composé en anglais langue maternelle (ALM) et en français langue étrangère (FLE). On voit que les textes en langue maternelle privilégient les circonstants de type *condition/concession* alors que leurs homologues en langue étrangère ont une forte majorité de circonstants de lieu et de temps. Il en résulte que, globalement, et malgré les consignes identiques, on a l'impression d'avoir affaire à des textes radicalement différents : argumentations pour ALM (ce qui était demandé), exposés commentés pour FLE. Les textes en français langue maternelle, FLM, représentent un moyen terme, avec une distribution équilibrée des différents types.

Il serait tentant d'interpréter ces résultats dans une optique de rhétorique contrastive pour la comparaison entre les deux corpus en L1 : les francophones et les anglophones exhibent des stratégies de textualisation assez nettement distinctes, qui pourraient être représentatives de "modes" rhétoriques différents (Cf. Clyne, 1981; Hinds, 1983; Béacco, 1992; Régent, 1985; 1992; Rodrigues Faria Coracini, 1992). Le faible volume des données incite néanmoins à la prudence. Les études de rhétorique contrastive posent un difficile problème de comparabilité : on ne peut utilement comparer les stratégies de textualisation

[13] Cf. Bronckart *et al.* (1985).

que dans des textes appartenant au même genre discursif. Or la définition de ces genres[14], et *a fortiori* l'établissement de correspondances entre genres dans des langues/cultures différentes, est un problème épineux, si ce n'est dans les cas de forts contacts et/ou de soumission à un mode rhétorique dominant, comme pour les articles scientifiques par exemple. Ce que j'ai voulu montrer par ces chiffres bruts et par ces exemples, c'est le rôle majeur que jouent ces thèmes marqués dans l'organisation textuelle, et par conséquent dans la visée discursive que l'on perçoit à la lecture de ces textes.

Pour conclure cette section sur le placement des éléments mobiles, un exemple tiré du corpus Reformulation (Péry-Woodley, 1993a). Il s'agit du début d'un texte d'étudiant et de sa reformulation par un enseignant "natif". Le texte de gauche, l'original rédigé par l'étudiant non-natif, place le circonstant de temps *twenty five years ago* en position rhématique dans la première unité syntaxique du texte. La reformulation le déplace en position thématique, créant un effet de structuration tout à fait différent. On a alors, de phrase en phrase, une structuration chronologique : *A quarter of a century ago*, *Now*, qui semble mieux préparer le terrain pour le thème suivant : *The change*.

(12) Texte original	Reformulation
[1] Leyford, town close to London, was a declining industrial town ***about twenty five years ago,*** *[2] but now our town is a beautiful thriving tourist and commercial centre. [3] What made Leyford change so completely ?*	*[1]* ***A quarter of a century ago****, Leyford, a small town situated close to London, was a declining industrial town. [2] Now it is a beautiful and thriving tourist and commercial centre. [3] What has caused Leyford to change so radically ?*
[4] The trigger of the change was the world recession.	*[4] The change was triggered, ironically, by the world recession.*

Ainsi se termine cette première partie de l'exploration de la structure thématique, qui a montré comment des constituants phrastiques dont la place n'est pas fixée par la syntaxe acquièrent une fonction de structuration textuelle fondamentalement différente selon qu'ils sont "emballés" comme thème ou comme rhème. Je reviendrai sur ces questions dans le chapitre 6, avec l'étude de certaines expressions introductrices d'univers de discours. Je vais maintenant examiner des constructions syntaxiques dont la fonction semble être spécifiquement liée à la structuration thématique. Il s'agit donc, là encore, d'un regard textuel sur les agencements phrastiques.

2.2.2 Syntaxe marquée

Mon étude du rôle de certaines constructions syntaxiques dans le marquage du thème a été considérablement inspirée par les travaux de trois chercheurs : T. Givón (1983), A. Davison (1984) et K. Lambrecht (1987). Le premier coordonne un ouvrage consacré au codage grammatical du thème dans diverses langues. L'objectif central de ces recherches est la caractérisation des "structural correlates of the functional domain of topic identification, topic maintenance and topic continuity in discourse" (Givón, 1983:9). Il y propose une échelle qui ordonne les constructions et les types de référents selon la facilité d'identification du thème. On y trouve, du plus accessible au moins accessible, les anaphores zéro (pronoms zéro en fonction sujet dans des langues comme l'italien ou l'espagnol), les anaphores pronominales non accentuées (clitiques du français), les anaphores pronominales accentuées, les syntagmes nominaux définis disloqués à droite, disloqués à gauche, thématisés, les constructions clivées, les syntagmes nominaux indéfinis. Il aborde également le rôle du passif et de la subordination.

[14] J'ai esquissé une définition dans le préambule, en termes de visée discursive, de relations entre participants et de canal. La notion de genre discursif sera le thème principal du chapitre 8.

A. Davison se focalise sur la syntaxe, et s'attache à montrer son rôle dans la signalisation du thème. Les langues ne semblent pas avoir de marques grammaticales spécialisées pour indiquer le thème (le *wa* du japonais, souvent cité dans ce contexte a également une interprétation contrastive), mais certaines relations syntaxiques peuvent faire l'objet d'une "exploitation pragmatique" pour signaler le thème : "Surface syntactic structure has properties which are perceived wben the sentence is related to discourse context" (Davison, 1984:801). Elle propose une échelle de constructions marquées où l'on retrouve le passif, les dislocations à droite et à gauche, la thématisation.

C'est aussi ce regard pragmatique sur la syntaxe qui motive les travaux de K. Lambrecht : certains choix syntaxiques ne s'expliquent qu'en référence à la structure d'information. K. Lambrecht se concentre sur le français parlé et ses écarts syntaxiques par rapport à la syntaxe canonique décrite par les linguistes. Il analyse finement les dislocations, les thématisations, les constructions présentationnelles et les clivées si typiques du français parlé. Ci-dessous quelques uns des exemples qui fondent son analyse (les conventions de transcription sont les siennes, j'ai souligné les constituants impliqués) :

– Dislocations à droite et à gauche :

(13) (Un mari à sa femme, se plaignant du contenu de son assiette)

A— Ça n'a pas de goût, ***ce poulet***

B— ***Le veau****, c'est pire (1985, ex. 18)*

– Constructions présentationnelles :

(14) A l'heure actuelle, j'm' plains pas, ***ya*** *un camarade d'usine* ***qui*** *m'ramène en voiture (1985, ex. 6)*

(15) Moi ***j'ai*** *encore un formulaire* ***que*** *j'ai pas (1985, ex.9)*

– Constructions clivées :

(16) Tous ceux qu'ya dans le quartier, ***c'est moi qui*** *leur a donné des bouts. (1985, ex. 2)*

K. Lambrecht explicite les contraintes liées à la structure d'information qui font que ces constructions sont employées de préférence à la version "canonique" SVO correspondante. Le fait que ses travaux portent sur des constructions typiques de l'oral, dont certaines sont totalement exclues de l'écrit normé, n'enlève rien à la pertinence de l'approche. Et c'est dans cette optique que j'ai abordé l'analyse des constructions clivées et des passifs dans mes corpus.

• Constructions clivées

Les clivées constituent un groupe de constructions particulièrement intéressantes dans une perspective discursive sur la syntaxe dans la mesure où on ne peut s'expliquer le choix d'une clivée de préférence à une non-clivée, ni le fonctionnement de la clivée, sans faire appel à des considérations fonctionnelles dépassant le cadre de la phrase. Je ne m'intéresserai pas ici à la pseudo-clivée (ex. *Ce qu'il me faut, c'est une bière bien fraîche*), mais seulement aux clivées proprement dites, que plusieurs auteurs s'accordent à répartir en deux types selon le statut présuppositionnel de la proposition en QU. On a longtemps en effet considéré que celle-ci était obligatoirement présupposée. C'est le cas des clivées d'identification ou de contraste typiques de l'oral qui permettent en français d'éviter de placer un élément rhématique dans la position normalement dévolue au thème (cf. l'exemple anglais de 2.1, qui pourrait se traduire, en réponse à *Tony saw the play yesterday* : *C'est John (et pas Tony) qui a vu la pièce hier*). Elles semblent viser avant tout la "mise en saillance" de l'élément introduit par le présentatif *c'est*, qui en fait le rhème de la première partie de cette phrase "divisée", avant qu'il devienne ensuite thème de la deuxième partie. Appelons-les *clivées de saillance*. On en trouve également des exemples à l'écrit :

(17) ... on dit que l'opinion compare l'économie carterienne et l'économie reaganienne ; elle compare en réalité la récession reaganienne et la reprise reaganienne. ***C'est dans ce cadre, et dans ce cadre seulement, qu****'il y a amélioration réelle.*

Toutefois on recueille sans peine, à l'écrit tout au moins, des exemples de clivées dont la proposition en QU, comme certaines relatives appositives, ne semble pas présentée par le scripteur comme "connue" du lecteur. Comme pour les relatives appositives "informatives", elles sont présupposées dans le sens où elles ne sont pas assertées, mais pas dans le sens où elles sont traitées comme connues. On les appellera *clivées d'information* :

(18) Il y a deux mois encore, on ne connaissait à cette planète de gaz près de soixante-dix fois plus volumineuse que la Terre que cinq satellites (...) dont le dernier fut découvert en 1948.

En moins d'un mois, ***ce sont neuf autres qui*** *sont identifiés sur les images envoyées par la sonde pourtant distante de la terre de trois milliards de kilomètres.*

A. Borkin (1984) propose une intéressante analyse discursive des clivées : l'élément qui, au niveau de la phrase, détient le plus haut statut informationnel, se trouve dans les clivées occuper la position – en début de phrase – normalement associée avec l'élément connu. La proposition subordonnée, normalement présupposée, occupe elle la position finale, normalement associée au pic informationnel de la phrase. On a donc une construction dynamique où des forces opposées exercent une tension. Il en résulte, autant qu'une mise en valeur d'un élément dans la clivée, une mise en valeur de la clivée dans son contexte. On aurait donc, là aussi, comme on l'a vu pour certains circonstants thématiques, un élément saillant dans un segment de texte, et qui contribue ainsi à organiser l'ensemble.

Mes corpus étant limités en taille, et les clivées rares, le faible nombre d'occurrences analysées impose la plus grande prudence interprétative. Envisager d'aussi rares occurrences au niveau du groupe, alors que d'importants écarts existent d'un sujet à l'autre, n'est pas non plus sans poser problème. Les décomptes ne sont toutefois pas utilisés ici avec une visée généralisatrice mais uniquement comme contribution à un faisceau d'observations. Dans le corpus Étudiants, on est d'emblée frappé par la différence de fréquence des clivées entre les sous-corpus (tableau 2.2) :

	fréquence moyenne	**clivées de saillance**	**clivées d'information**
FLM (n=23)	**1,77/texte**	0	**23**
ALM (n=10)	0,67/texte	4	6
FLE (n=10)	0,67/texte	2	8

n = nombre total de clivées

Tableau 2.2: Clivées : Fréquence moyenne et par type(corpus Étudiants).

Les francophones utilisent ces constructions presque trois fois plus que les anglophones, qui les utilisent d'ailleurs aussi peu en anglais qu'en français. Une analyse en termes des deux types définis plus haut révèle que toutes les clivées des francophones se rangent clairement dans la catégorie des clivées dites "d'information", alors que les clivées des anglophones, quelquefois d'ailleurs difficiles à caractériser, se répartissent dans les deux catégories. L'analyse de A. Borkin m'avait amenée à attendre un effet de saillance, un rôle structurant de la clivée, qui dominerait un certain nombre de propositions. Ce n'est pas du tout le cas pour les clivées du corpus FLM, qui sont presque systématiquement employées pour introduire des données extraites des tableaux statistiques du matériau Étudiants :

(19) En 65 et 70, ***c'est en France que*** *le pourcentage est le plus bas, avec une très faible progression, mais régulière et qui se poursuit en 74 ; mais en 74,* ***c'est l'Italie***

***qui** détient le plus faible pourcentage, suivie du Japon : on peut noter que c**'est le Japon qui** connaît la plus faible progression des pourcentages au fil des ans. (FLM-13 : 34-36)*

Le fait que, comme dans l'exemple (19), ces clivées peuvent se succéder à très faible intervalle exclut la notion d'une mise en valeur d'une proposition dans le discours. Dans l'exemple (19), on semble avoir un effet de double structuration thématique : chronologique par le circonstant à l'initiale, puis géographique par le nom de pays dans la clivée. Il n'en va pas du tout de même dans les clivées des anglophones, particulièrement en L1. Elles se situent pour la plupart dans la dernière phrase d'un paragraphe ou du texte, quelquefois dans la première, et semblent avoir un rôle important dans l'orientation argumentative de ces textes. L'exemple (20) montre l'utilisation systématique de clivées pour conclure des paragraphes consacré à l'examen des différents tableaux statistiques :

*(20) Table one shows the rate of scholarization (...). Compared with the US and Italy, (...), Great Britain's percentage (...). In fact, of all the countries mentioned, **it is Great Britain which** has the lowest percentage (...) in higher education.*

*The average age (...) is represented in table two. (...) The United States' figures (...). However, Great Britain's figures (...). So **it is again the British education system which** seems to be contracting the most compared with the other countries.*

*The average rate of annual increase of public expenditure on higher education is shown in table three. The figures clearly reveal (...). Again **it is the British education system which** is contracting the most in view of the amount of money consecrated to it. (ALM-2)*

L'exemple suivant, issu du corpus Reformulation, va dans le même sens. Il reproduit l'introduction, déjà présentée dans la section précédente, avec cette fois la conclusion de l'extrait dans les deux versions (on trouvera le texte complet en annexe) :

(21) Texte original	Reformulation
Leyford, town close to London, was a declining industrial town about twenty five years ago, but now our town is a beautiful thriving tourist and commercial centre. What made Leyford change so completely? *(...)* *There were so many people in London that the government felt it necessary to decentralize overpopulated London.*	*A quarter of a century ago, Leyford, a small town situated close to London, was a declining industrial town. Now it is a beautiful and thriving tourist and commercial centre. What has caused Leyford to change so radically?* *(...)* *Eventually London became so overpopulated that the government felt it necessary to adopt a policy of decentralisation. **It is this decentralisation policy that** has enabled Leyford to recover so successfully from the period of industrial decline.*

Le reformulateur a ressenti la nécessité d'ajouter une phrase de conclusion, absente du texte original. Cette phrase fait le lien avec l'introduction et conclut ce paragraphe sur l'origine du changement survenu à Leyford au cours des vingt cinq dernières années. Il s'agit d'une phrase clivée, qu'il est intéressant, pour en dégager la fonction, de contraster avec son "allophrase"[15] non-clivée :

(22') This decentralisation policy has enabled Leyford to recover (so) successfully from the period of industrial decline.

[15] Le terme *allosentences* est repris par Lambrecht (1994) à Danes (1966) pour dénoter des paires de phrases qui, tout en étant équivalentes sémantiquement, diffèrent sur les plans formel et pragmatique.

La clivée est perçue comme clôturant la portion de texte qui apporte une réponse à la question posée à la fin de l'introduction : *What has caused Leyford to change so radically?* Notons que l'intensifieur *so*, rappel de la question, doit être supprimé dans la version non-clivée, qui ne semble en effet pas pouvoir jouer le rôle de rappel et de clôture joué par la clivée. On a donc ici un fonctionnement qui semble correspondre à ce que percevait Borkin : plutôt que la mise en saillance d'un élément à l'intérieur de la clivée, la mise en saillance de la phrase en discours.

Si on reprend la distinction entre clivées de saillance et clivées d'information, il semble que les secondes soient exclues de l'effet de saillance discursive décrit par A. Borkin, et qui m'intéresse particulièrement sur le plan de l'organisation du texte. De nombreuses questions se posent. Les seuls cas de saillance discursive ont été observés en anglais. Cela indique-t-il un fonctionnement différent en anglais et en français ? Les clivées de saillance en anglais semblent fréquemment associées aux segments conclusifs, fins de paragraphes ou du texte. Peut-on systématiser cette observation ? Ce rôle de la clivée est-il le propre de certains genres discursifs ? [16].

• Passifs

A la suite d'un certain nombre d'auteurs (Givón, 1979; Werth, 1984), je me suis d'abord intéressée au passif en tant que moyen d'harmoniser structure syntaxique et structure thématique, puisqu'il permet de placer en fonction de sujet grammatical et dans la position initiale (ou proche de l'initiale) normalement associée au thème un élément connu qui, étant donné la structure argumentale du verbe, serait normalement en position de rhème. On verra aussi qu'une autre fonction du passif est la mise en valeur du rhème. Pour illustrer, trois exemples où des passifs attestés (colonne de gauche) sont mis en regard de leurs allophrases actives (colonne de droite) :

Version originale passive	Version active
(22) Inculpation de deux responsables de la construction du télésiège de Luz-Ardiden *MM. Jean Berseille et Yves Estebenet, deux des responsables de la construction du télésiège de Luz-Ardiden, **ont été inculpé**s vendredi 31 juillet d'homicide et blessures involontaires par M. Christian Mésière, juge d'instruction à Tarbes. (…)*	*Inculpation de deux responsables de la construction du télésiège de Luz-Ardiden* *M. Christian Mésière, juge d'instruction à Tarbes, a inculpé vendredi 31 juillet d'homicide et blessures involontaires MM. Jean Berseille et Yves Estebenet, deux des responsables de la construction du télésiège de Luz-Ardiden*
*(23) (…) Aden, cité maudite où, dit-il, "les affrontements de ces dix derniers jours ont sans doute coûté la vie à dix mille personnes". Ce très lourd bilan lui **a été communiqué par** M. Abbas Zaki, le représentant local de l'OLP, homme très en vue et très introduit à Aden.*	*(…) Aden, cité maudite où, dit-il, "les affrontements de ces dix derniers jours ont sans doute coûté la vie à dix mille personnes". M. Abbas Zaki, le représentant local de l'OLP, homme très en vue et très introduit à Aden, lui a communiqué ce très lourd bilan*

[16] J'ai repris ce thème de recherche récemment grâce à trois étudiantes de maîtrise. Le projet vise la mise en relation d'une description syntaxique de clivées en corpus (constituants après *c'est*, type de "relative") avec la distinction entre clivée de saillance et d'information, et l'examen du fonctionnement en contexte des clivées par la comparaison systématique avec leur allophrase non-clivée.

(24) ***The Editorial Board***	***The Editorial Board***
Members of the Editorial Board ***are appointed by*** *the Committee of the Society (...). Two members* ***are nominated by*** *the British Biophysical Society.*	*The Committee of the Society appoints members of the Editorial Board. (...) The British Biophysical Society nominates two members.*
Normally a paper ***is read by*** *at least two people: either by two members of the Editorial Board, or (...)*	*Normally at least two people read a paper: either two members of the Editorial Board, or (...)*

Ces trois exemples, même peu contextualisés, devraient suffire à montrer que le choix d'un passif n'est pas gratuit sur le plan du texte. Les deux premiers sont extraits du journal *Le Monde*, le troisième d'une revue scientifique, le *Biochemical Journal*. En (22), les versions passive et active présentent deux organisations du récit, autour des inculpés (conditionnement habituel pour ce type de "brèves") ou autour du juge. En (23), le passif permet de mettre en fonction de sujet grammatical, fonction associée au rôle de thème, un groupe nominal résumant le rhème précédent (établissement d'un nouveau thème, référent connu), et l'auteur de la communication est présenté en position rhématique par rapport à ce thème. La version active rompt la continuité, et accorde à *M. Abbas Zaki* un rôle thématique, et donc une saillance discursive, que le reste du texte ne justifie peut-être pas. La version active de (24) présente un catalogue quasi prévertien d'agents au lieu d'être organisée par les objets appartenant au "frame" d'un comité de rédaction de revue scientifiques : "members, papers". Dans ces trois cas il est clair qu'il ne s'agit pas de variantes stylistiques "libres", mais que le choix d'une construction active ou passive oriente le texte de façon fondamentale.

J'ai procédé à une analyse détaillée des passifs du corpus Étudiants dans un triple objectif : parvenir à une meilleure compréhension du fonctionnement du passif en contexte, comparer le rôle textuel joué par le passif en français et en anglais, examiner dans quelle mesure des apprenants maîtrisent ce procédé de structuration thématique en français langue étrangère (Péry-Woodley, 1989; 1991a). Pour l'anglais, j'ai suivi la classification de E.L. Keenan (1985) et j'ai inclus :

– les passifs sans agent exprimé, ex. *Expenditure on universities has been cut back* ;

– les passifs avec agent, ex. *Expenditure on universities has been cut back by the government* ;

– les passifs impersonnels (avec ou sans agent), ex. *It is argued (by many) that there is a need to reduce expenditure*.

Pour le français, on retrouve les deux premiers types, ex. *D'autres pays ont été assez durement touchés (par la crise)*. En revanche, aucune occurrence de passifs impersonnels dans le corpus. Une construction proche a été incluse dans l'analyse, la construction dite "moyenne", ex. *Un tel pourcentage aux USA peut peut-être s'expliquer par un moindre coût des études supérieures*.

Après un bref aperçu comparatif de la fréquence du passif, j'illustrerai ses trois fonctions majeures dans le corpus, pour terminer sur un examen de fonctionnements propres à certains sous-corpus : passifs impersonnels en anglais, *on* comme équivalent du passif en français langue étrangère.

a) Fréquence

Le tableau 2.3 apporte une réponse quantitative nette à la question contrastive sur le rôle textuel joué par le passif en anglais et en français :

FRÉQUENCE

	moyenne par texte	% unités syntaxiques avec passif
FLM (n =20)	1,5	7,1%
ALM (n =123)	**8,2**	**31,1%**
FLE (n =11)	0,7	3,2%

n = nombre de passifs

Tableau 2.3 : Fréquence des passifs dans les trois sous-corpus (corpus Étudiants).

On ne peut que constater que si le choix du passif tient effectivement à des contraintes de structure thématique, c'est un procédé bien plus utilisé dans les textes en anglais que dans les textes en français, et on peut se demander par quoi les scripteurs "remplacent" en français ce procédé d'établissement et de continuation thématique. Le passif est encore plus rare dans les textes en FLE. Il y a lieu de s'interroger sur des écarts aussi marqués : autant qu'à une différence liée aux procédés régulièrement utilisés dans les deux langues pour marquer le thème, il se pourrait que ces écarts aient trait au fait que les trois groupes de textes sont assez distincts sur le plan de la visée discursive (cf. commentaire du tableau 2.1). Le passif pourrait être associé à la démarche argumentative surtout adoptée par les anglophones en ALM.

T. Givón (1979) insiste sur les variations de fréquence des passifs en fonction du genre discursif : il donne une fourchette allant de 4% des phrases dans le registre relativement informel des articles d'information et de sport du Los Angeles Times, à 18% dans un passage d'un ouvrage de N. Chomsky[17]. Ces analyses quantitatives soulèvent des questions méthodologiques intéressantes. Dans une approche fonctionnelle, et donc nécessairement sur corpus, l'analyse quantitative fait partie intégrante de la démarche. Il ne s'agit pas en effet de déterminer si une construction est grammaticale ou non, mais comment elle est utilisée en discours. Sa fréquence est un aspect essentiel de son utilisation[18]. Pour ce qui est des analyses contrastives, le chemin est semé d'obstacles. D'abord celui de la comparabilité des textes, évoqués plus haut. On voit ici que le fait de recueillir un corpus en situation expérimentale à partir d'un même instrument ne garantit pas la comparabilité entre les textes. Les sujets anglophones se sont en effet montrés moins capables de construire une argumentation en FLE qu'en ALM, et ont eu tendance à s'engluer dans les données statistiques du matériau Étudiants. Surcharge cognitive liée à la difficulté de rédiger en langue étrangère ? Façons différentes d'aborder la tâche d'écriture ? On constate en tout cas la difficulté de constituer des corpus de textes comparables. Pour finir, il est important de veiller à ne pas considérer d'emblée qu'une forme syntaxiquement "équivalente" dans les deux langues, telle le passif, fonctionne de façon semblable sur le plan du texte. C'est ce que je vais montrer dans ce qui suit.

b) Passifs, thèmes et rhèmes

Les fonctions textuelles du passif définies plus haut – établissement d'un thème, respect de la structure d'information – se retrouvent dans les trois sous-corpus, en dépit des différences de fréquence. Pour illustrer ces fonctions, le corpus ALM me fournit un bel exemple :

> *(25) The United States' figures represent the longest time for pupils to stay in the education system : that is **16.7 years on average are spent by** a pupil at school. (ALM-2 : 7-8)*

[17] Ces chiffres ne sont pas directement comparables avec les miens, l'unité envisagée étant la phrase dans l'analyse de T. Givón, l'unité syntaxique (cf. 2.1) dans la mienne.

[18] Je reviendrai sur cette question dans le chapitre 8, en présentant les résultats d'une analyse à grande échelle du passif en anglais (Francis et Kucera, 1982).

Formulation étrange étant donnée la préférence supposée à la fois pour les constructions actives et pour les sujets humains. On s'attendrait à trouver :

(25') (…) pupils spend on average 16.7 years at school.

Le passif est ici clairement motivé par des considérations de structure thématique : en faisant de la durée de scolarité, qui reprend en le spécifiant le rhème précédent, le sujet de la construction passive, on en fait l'élément qui sera perçu comme thème de l'unité syntaxique, et on assure la continuité thématique.

On retrouve également avec certains passifs un effet de structuration sur le plan du texte par l'établissement de thèmes liés, comme on l'avait vu pour les circonstants initiaux dans la section précédente (2.2.1). Les deux passifs de l'exemple (26) ouvrent de nouveaux paragraphes, à l'intérieur d'une partie consacrée aux tableaux statistiques, en présentant le titre de chaque tableau comme thème d'un paragraphe :

(26) The average age at which pupils enter and leave the education system ***is represented*** *in the figures of table two (…).*

The average rate of annual increase of public expenditure on higher education ***is shown*** *in table three (…). (ALM-2 : 6 et 11)*

En retournant les choses par rapport à l'ordre "normal" des arguments, le passif touche non seulement le thème mais aussi le rhème. T. Givón (1979), P. Werth (1984), E.L. Keenan (1985) *inter alia* se sont intéressés à l'expression de l'agent, et à la valeur rhématique qu'il prend dans cette construction. T. Givón constate que parmi la faible proportion de passifs avec agent (20% dans son corpus), ceux-ci sont presque toujours indéfinis, et en conclut que le passif sert aussi à placer des référents nouveaux en position rhématique. L'examen de mon corpus révèle des différences considérables à ce propos. Les passifs avec agents sont en effet rares dans les textes rédigés par des anglophones (7,3% en L1 et 9,1% en L2). Ils sont en revanche beaucoup plus fréquents que dans le corpus de T. Givón chez les francophones (35%). Pour ces trois groupes toutefois, l'analyse de T. Givón ne semble pas pertinente puisque la plupart de ces agents sont définis, et donc connus (ou traités comme tels). Une autre explication, proposée par R. Quirk *et al*. (1985), serait que le passif est utilisé pour placer en fin de phrase un constituant très développé (*principle of end-weight*). L'exemple (27), issu du corpus FLM, semble procéder de ce principe :

(27) Un tel pourcentage aux USA peut alors peut-être s'expliquer par un moindre coût des études supérieures ou alors par le fait que les familles des étudiants sont plus aisées financièrement en moyenne et donc en mesure de subvenir aux besoins de leurs enfants si ceux-ci désirent poursuivre leurs études.

Ce principe a l'utilité de rappeler que l'ordonnancement des constituants est la résultante de l'application de multiples facteurs. Sur le plan de la structure thématique, il semblerait en tout état de cause que ce soit les fonctions liées au thème qui prédominent dans ce corpus, sauf pour les passifs impersonnels, analysés ci-dessous.

c) Passifs impersonnels anglais

On a vu que le sous-corpus anglais est le seul à comporter des passifs impersonnels. Ceux-ci constituent 8% des occurrences de passifs dans ce corpus, avec une distribution très inégale (0 à 5 par texte). Ces constructions semblent paradoxales : en effet, elles ne peuvent participer du rôle majeur du passif, établir un thème, dans la mesure où leur sujet, l'impersonnel *it*, n'a pas de référent et en conséquence aucun potentiel thématique. Ces passifs sont toujours construits avec une complétive qui occupe la fonction de rhème. Elles semblent avoir un rôle d'encadrement énonciatif, soit avec des verbes *dicendi* qui présentent une assertion que le rédacteur n'assume pas, soit pour émettre une opinion sous une forme impersonnelle :

(28) However ***it is argued by*** *many that there is an inevitable need to reduce the rate of expenditure on higher education. In the bleak economic climate of today* ***it must be admitted*** *that many young people graduating from higher education are still unable to find employment. (ALM-12 :13-14)*

Ces passifs, qui effacent l'agent ou le placent dans une position non saillante, ont pour effet de créer une distance énonciative. On peut les rapprocher de trois constructions du français :

(29) ***Il convient tout d'abord de noter*** *que ces statistiques ne concernent que des pays industrialisés, ce qui permet un recoupement plus facile, mais restreint les éléments de comparaison entre différents systèmes économiques. (FLM-7 :5))*

(30) (…) ***nous pouvons constater*** *qu'en général, peu de jeunes de 19 à 24 ans suivent des études supérieures. (FLM-5 : 2)*

(31) Cependant, pour ce qui est de la France, ***on remarque*** *que pour la période 65 à 70 elle fait partie des pays qui ont fait le plus gros effort en faveur de l'enseignement supérieur. (FLM-2 :11)*

Ce rapprochement, ainsi que la remarquable fréquence de *on* dans les textes en français rédigés par des anglophones (18,4% des sujets de propositions principales, contre 11,3% en FLM et 3,5% pour *one* en ALM), m'ont conduite à examiner la fonction de *on* dans le corpus en relation avec le passif.

d) Passifs et on

Cette analyse fait référence à un aspect de la "mise en thème" brièvement évoqué dans l'introduction à cette section : l'impact des propriétés référentielles d'un syntagme nominal sur son *potentiel thématique* – sa plus ou moins grande aptitude à jouer le rôle de thème –, et le potentiel thématique des sujets grammaticaux (Péry-Woodley, 1991a; 1993a). A. Davison (1984) classe ainsi les SN depuis ceux qui font de bons thèmes grâce à leur aptitude à désigner un référent, par exemple les noms propres, jusqu'à ceux dont le comportement référentiel en fait des thèmes improbables : SN indéfinis, ou parties de locutions. Les propriétés référentielles d'un SN ont trait à la fois à la détermination et au sémantisme des lexèmes. Je n'approfondirai pas ici l'aspect lexical de ce potentiel, qui va éclairer le fonctionnement des constructions identifiées précédemment comme pouvant "équivaloir" en français à certains passifs anglais.

Les remarques qui suivent partent d'une distinction initiale entre deux usages de *on* et de *nous* : un usage que j'appellerai *interne* parce qu'il désigne les participants à l'interaction, se situant au plan interpersonnel ; et un usage *externe* qui fait référence à des acteurs au plan idéationnel. Il faudrait creuser davantage l'association de ces deux usages avec des classes de verbes : il semble que premier soit clairement associé aux verbes *dicendi*. Les exemples (30) et (31) ci-dessus illustrent l'usage interne, (32) et (33) ci-dessous l'usage externe :

(32) (…) en Angleterre, ***nous*** *dépensons plus d'argent que les autres pays (sauf les Etats-Unis), mais très peu de jeunes gens reçoivent l'enseignement supérieur. (FLE-14 : 10-11)*

(33) Le problème du manque de crédits se pose dans la majorité des facultés : ***on*** *supprime des cours,* ***on*** *ferme les bibliothèques … (FLM-11 : 20-22)*

On constate immédiatement que c'est dans leur usage interne que *nous* et *on* peuvent correspondre au passif impersonnel anglais (*on remarque…, it must be noted…*). Les deux sous-corpus en français présentent des exemples de cet usage, qui semble bien maîtrisé par les apprenants. Ils diffèrent en revanche fortement en ce qui concerne les taux d'utilisation : l'usage interne est fortement majoritaire en FLM (exclusif même pour le cas de *nous*), alors qu'il est minoritaire en FLE. Je vais à ce stade délaisser *nous*, très peu fréquent, pour me focaliser sur *on*, dont la fréquence comme sujet de proposition principale est considérable,

surtout en FLE. Le tableau 2.4 donne le détail de l'utilisation de *on* dans les sous-corpus FLM et FLE.

	EXTERNE		INTERNE		total %
	% *on*	% sujets	% *on*	% sujets	sujets
FLM (n= 32)	31,3%	3,5%	**68,7%**	7,7%	11,3%
FLE (n= 63)	**57,1%**	10,5%	42,9%	7,9%	**18,4%**

n = nombre de *on* sujets de propositions principales

Tableau 2.4 : Fréquence relative de on *externe et interne (corpus Étudiants).*

L'usage externe de *on* représente en FLE 10,5% des sujets de propositions principales. Si on reprend l'exemple (33), on peut émettre l'hypothèse que ces formulations, proches d'un passif sans agent (*des cours sont supprimés, les bibliothèques sont fermées*), compensent en quelque sorte le "manque" de passifs dans ce sous-corpus. A y regarder de plus près, on aperçoit cependant un problème. F. Atlani (1984) fait état d'une contrainte importante dans l'usage de *on* que j'ai appelé externe : le seul usage anaphorique possible de *on* est dans le sens du *ils* indéterminé. Or cette contrainte se trouve très souvent violée dans les textes d'apprenants, donnant lieu à des problèmes de résolution d'anaphore, mais aussi plus globalement à des problèmes de structuration thématique. Quelques exemples :

(34) Il y a certains qui sont tout à fait contre cette raison que donne le gouvernement, surtout celui de la Grande-Bretagne. ***On*** *dit que le gouvernement paie bien d'autres choses /.../.* ***On*** *dirait que ce n'est pas aussi important que l'enseignement, /.../* ***On*** *dit que les recherches nucléaires sont inutiles /.../. /.../*

Cependant, il y a d'autres qui disent qu'une expansion (...) est tout à fait impossible parce que ça coûte trop cher. ***On*** *dirait que ce n'est pas juste de dépenser beaucoup d'argent ... (FLE-6)*

Les quatre *on* de (34) sont utilisés comme anaphoriques avec des référents spécifiques, en l'occurrence les groupes définis dans les phrases précédant *on* : "*certains qui sont tout à fait contre…",* et "*d'autres qui disent qu'une expansion*…". Cela les rend déviants, et presque ininterprétables[19]. Ces usages anaphoriques déviants expliquent le fait qu'une majorité de *on* en FLE sont externes. Pour reprendre la question posée plus haut (commentaire du tableau 2.3), cet usage de *on* peut être envisagé comme une sorte de compensation pour le passif presque absent des textes FLE. Il ne constitue cependant pas un bon substitut au passif sur le plan textuel puisque d'une part il ne peut créer l'effet de distance énonciative auquel contribue le *on* interne caractéristique de FLM, et d'autre part sa référentialité déviante l'empêche d'être performant pour introduire ou rétablir des thèmes.

Cette analyse montre comment des structures qui se correspondent d'une langue à l'autre sur le plan syntaxique peuvent avoir, si on les examine dans une optique textuelle, des fonctionnements assez différents. J'ai proposé le nom de "faux-amis textuels" pour ces pièges mis en évidence par l'analyse comparative : la ressemblance formelle cache une différence fonctionnelle. S'il est possible d'arriver à des généralisations interlinguistiques quant à la signalisation de l'organisation textuelle, c'est sans doute uniquement sur le plan relativement abstrait des types de constituants impliqués, ou peut-être de types de transformations, telles l'union d'éléments paratactiques en une seule unité avec un élément hypotactique. C'est le rôle de ces variations syntaxiques dans la structuration thématique qui va maintenant être abordé.

[19] Sur un autre plan d'analyse, il semble qu'on soit ici en face d'une difficulté à gérer la polyphonie en langue étrangère.

2.3. Syntaxe complexe et structuration thématique

On a vu comment le positionnement de certains constituants ainsi que certaines structures syntaxiques dites "marquées" peuvent représenter et signaler des choix de structuration thématique. On va maintenant s'intéresser à l'impact thématique d'un autre type de choix syntaxique : entre expression paratactique ou hypotactique du lien entre plusieurs propositions (Péry-Woodley, 1990b; 1991b; 1993a).

2.3.1 Maturation syntaxique

Mon approche initiale des effets textuels de la syntaxe (phrastique) complexe a été impulsée par les travaux américains – passablement controversés d'ailleurs – sur la *maturité syntaxique*. K.W. Hunt (1965; 1970) montre, dans un cadre très marqué par la grammaire générative d'alors, l'évolution de la syntaxe au cours de la scolarité sur des échantillons correspondant à nos CM2, troisième et terminale. Plutôt que la phrase, il choisit pour son analyse une unité proprement syntaxique le T-unit[20],, qui lui permet de différencier un allongement dû à des coordinations d'un allongement dû à des subordinations. Il mesure :

– la longueur moyenne des T-units ;
– le nombre de propositions par T-unit ;
– la longueur des propositions ;
– le nombre de T-units par phrase ;
– la longueur des phrases.

S'il y a une progression pour toutes les mesures, la fiabilité des mesures varie : la longueur des phrases se révèle être l'indice le moins fiable, la longueur des T-units et des propositions les indices les plus fiables. Cet effet de "maturation" syntaxique, qui consiste en une aptitude croissante à rattacher deux propositions ou plus à une principale par subordination ou réduction, est illustrée par l'exemple suivant. Pour les deux phrases de base : *J'ai un fils. Il a dix ans*, l'évolution typique sera :

coordination ======> subordination =====> réduction

J'ai un fils et il a dix ans ==> J'ai un fils qui a dix ans ==> J'ai un fils (âgé) de dix ans.

Sans entrer dans la controverse sur la relation entre maturité syntaxique et "qualité" de l'écrit, il est intéressant de s'interroger sur ce que fait la syntaxe complexe sur le plan textuel, et si ces fonctions peuvent être aussi bien réalisées par des phrases simples. Mon intuition de départ était que les processus complexes de textualisation, qui passent par la hiérarchisation des propositions et l'établissement de liens entre elles, ne pouvaient être réalisés dans les textes que par des constructions syntaxiques complexes. Je pensais par ailleurs que les modèles du texte centrés sur la structure thématique pouvaient fournir des outils pour cet examen textuel de la syntaxe complexe.

2.3.2. Syntaxe complexe et sélection thématique

R. Lakoff (1984) distingue quatre façons de relier des propositions :

i) parataxe pure :

(35a) L'enseignement supérieur est en pleine restructuration. Les jeunes ne savent plus où ils en sont.

ii) type mixte

(35b) L'enseignement supérieur est en pleine restructuration et les jeunes ne savent plus où ils en sont.

iii) quasi-hypotaxe

[20] Cf. définition en note 7 (2.1).

(35c) Comme l'enseignement supérieur est en pleine restructuration, les jeunes ne savent plus où ils en sont.

iv) hypotaxe pure (proposition réduite)

(35d) L'enseignement supérieur étant en pleine restructuration, les jeunes ne savent plus où ils en sont.

Avec les types i) et ii) on a deux prédications, une concernant l'enseignement supérieur, l'autre les jeunes, reliées en i) seulement par la juxtaposition, en ii) par un lien non spécifique, qui équivaut à dire au lecteur : "*ces deux idées sont reliées, à vous de deviner comment*" ! Les types iii) et iv) ont deux effets supplémentaires : ils ajoutent un élément d'information – la nature du lien entre les propositions –, et ils établissent une structure hiérarchique – il n'y a plus qu'une prédication et la première proposition devient présupposée. Un joli exemple tiré de T. Shopen et J.M. Williams (1981) éclaire un autre aspect du même phénomène :

(36) Tous les hommes sont mortels ;
or Socrate est un homme ;
donc Socrate est mortel.

Si l'on forme une seule unité syntaxique à partir des trois propositions de ce syllogisme connu, diverses formulations sont possibles, qui sont toutes caractérisées par le fait que *Socrate* est sujet de la proposition principale. Le but discursif d'un syllogisme étant sa conclusion, celle-ci est exprimée systématiquement dans la proposition principale[21]. Plusieurs observations ressortent de ces exemples :

– la syntaxe complexe crée un effet de hiérarchisation ;

– la proposition principale bénéficie d'un effet de saillance par rapport aux subordonnées ;

– si différents syntagmes nominaux sont en concurrence pour la fonction de thème, l'entité réalisée par le sujet de la proposition principale d'une phrase complexe acquiert un statut que cette même entité n'aurait pas dans une formulation paratactique.

La complexité syntaxique semble donc jouer un rôle dans la hiérarchisation des thèmes et la sélection des thèmes principaux. Cette hypothèse reçoit une confirmation d'ordre psycholinguistique à travers l'étude de L. Lautamatti (1978; 1987) sur la simplification syntaxique visant la facilitation de la lecture de textes scientifiques en anglais pour des étudiants non anglophones. Son étude révèle en effet que les textes "simplifiés" sont paradoxalement d'une lecture plus difficile que les textes originaux, et que le rappel est moins bon! L'exemple (37), où j'ai souligné les sujets de propositions principales, permettra de saisir le problème :

(37) Texte original :

*When a human infan**t** is born into any community in any part of the world **it** has two things in common with any other infant, provided neither of them has been damaged in any way either before or during birth. Firstly, and most obviously, **new born children** are completely helpless. Apart from a powerful capacity to draw attention to their helplessness by using sound there is nothing **the new born child** can do to ensure his (sic) own survival. Without care from some other human being or beings, be it mother, grandmother, sister, nurse, or human group, **a child** is very unlikely to survive.*

Texte simplifié :

***All healthy, new-born babies**, in all countries of the world, share two characteristics.*

[21] Cette affirmation se fonde sur des expérimentations certes informelles mais nombreuses avec des groupes d'étudiants au fil des années.

The first characteristic *which all human babies share, is that they are completely helpless.* ***The only thing they can do*** *to persuade someone to look after them is to cry, and in this way* ***they*** *can draw attention to themselves.* ***A helpless baby*** *will only survive if another human being looks after it.* ***The other human being*** *need not necessarily be the mother.* ***A grandmother, sister, or someone who is not related to the chil****d, may care for it.*

Là où *infant* (réalisé par *infant* ou *new born child)* est maintenu comme thème des quatre phrases du texte original, le texte simplifié introduit d'autres thèmes d'une façon non structurée, laissant au lecteur un travail de construction du sens beaucoup plus considérable.

On peut observer ce même effet dans un extrait du corpus Reformulation, où le locuteur natif opte dans sa reformulation pour une expression plus complexe syntaxiquement, mais qui précise la nature sémantique de la relation entre les propositions paratactiques de l'original, et impose une hiérarchisation thématique :

(38) Texte original	Reformulation
After the Second World War, the industrial foreign competition became severe. Japan and West Germany began to produce many good cheap machines such as cars and typewriters.	*After the Second World War, foreign industrial competition became severe as Japan and West Germany, in particular, began to produce good cheap manufactured goods (cars, typewriters, etc.).*

En réponse à la question évoquée au début de cette section, il semblerait donc que certaines des choses que l'on fait, textuellement parlant, avec une syntaxe complexe ne peuvent être faites avec des phrases simples. La relation entre syntaxe complexe et structuration thématique n'est cependant pas immédiate. On peut d'une part envisager des contre-exemples, où une syntaxe simple ne conduirait pas à une prolifération de thèmes, tel un récit en phrases simples mais dont chaque phrase reprend le même thème (progression à thème constant de F. Danes (Danes, 1974; Combettes, 1983)). Par ailleurs je me suis bien gardée jusqu'à présent d'évoquer le cas décrit en détail dans la section précédente des circonstancielles antéposées, auxquelles j'ai attribué un rôle thématique. Je vais donc maintenant affiner ce regard sur la syntaxe complexe en réintroduisant la notion de saillance, définie plus haut précisément en lien avec les circonstancielles.

2.3.3. Syntaxe complexe et saillance

C. Matthiessen et S. Thompson (1988), dans un article qui annonce la théorie de la structure rhétorique (RST, cf. chap. 3 *infra*), proposent une vision plus précise et plus explicative de la relation entre syntaxe complexe et structure discursive. Ils commencent par exclure de leur champ les enchâssements, où une proposition est un constituant de la proposition principale, pour ne s'intéresser qu'à l'hypotaxe, où une proposition est dans une relation de dépendance par rapport à une "tête" mais ne peut en être considérée comme un constituant. Se situant dans un contexte cognitif, où un des processus centraux dans l'interprétation des textes est la constitution de groupements hiérarchisés d'unités, ils cherchent à définir la fonction discursive qui motive l'hypotaxe. Un concept essentiel dans leur modèle – essentiel également, on le verra, dans la RST – est la relation asymétrique noyau-satellite :

> in any multi-unit text, certain portions realize central goals of the writer, while others realize goals which are supplementary or ancillary to the central goals. (1988:20)

Le noyau réalise donc les objectifs centraux, le ou les satellite(s) des objectifs auxiliaires. La relation noyau-satellite, qui ne se limite pas au plan des constituants phrastiques, est selon ce modèle indépendante de toute signalisation, mais la combinaison de propositions par hypotaxe est expliquée comme une grammaticalisation de cette relation. Cette analyse n'est pas sans rappeler celle de S. Thompson (1985) sur la portée des

circonstancielles de but antéposées (cf. 2.2.1), portée qui peut dépasser la proposition principale pour s'étendre à toute une série de propositions. En effet, si l'hypotaxe est une grammaticalisation de l'organisation rhétorique du discours, il n'est pas étonnant que la portée d'une proposition puisse dépasser l'empan de l'unité syntaxique pour embrasser un segment discursif.

Pourtant, si l'on en revient à mon analyse des circonstancielles initiales comme étant une façon privilégiée de conférer à un texte ou à un segment une organisation thématique, il semble y avoir contradiction entre cette saillance discursive et le statut de but secondaire ou dépendant qui découle de l'analyse de C. Matthiessen et S. Thompson. Pour mieux poser le problème, je reproduis en (39) un extrait de l'exemple de circonstancielles de but traité en 2.2.1 (ex. (5)),

(39) *[10] Pour toutes les consulter,*
[11] appuyez sur la touche SUITE
[12] afin de visualiser tous les écrans,
[13] ou RETOUR
[14] pour revenir en arrière.

Les propositions principales "jumelles" [11] et [13] seraient donc les noyaux, réalisations de l'objectif central du scripteur, tandis que les propositions hypotactiques, [10] au même titre que [12] et [14], réaliseraient des objectifs secondaires. Il me semble regrettable que l'intuition intéressante représentée par l'apparition de la relation noyau-satellite ait en quelque sorte annulé l'intuition antérieure concernant le rôle structurant des circonstancielles antéposées à portée étendue. Ici, dans un mode d'emploi, les propositions qui formulent les actions à entreprendre sont en effet centrales, et dûment exprimées par les principales ; les buts dans lesquels ces actions sont entreprises sont en effet auxiliaires par rapport aux actions elles-mêmes.

Ce qui en revanche semble à ce stade avoir été perdu de vue, c'est la différence de fonction de ces circonstancielles selon leur position thématique ou rhématique, différence de fonction qui touche le potentiel de structuration sur le plan du texte. S. Thompson (1985) interprétait pourtant cette différence dans les termes des métafonctions hallidayennes, et en particulier des plans textuel et idéationnel : les circonstancielles initiales fonctionneraient sur les deux plans, alors que les finales ne fonctionneraient que sur le plan idéationnel. L'interaction complexe entre hiérarchisation syntaxique et agencement linéaire des constituants ne peut en effet être comprise qu'en termes d'un modèle également plus complexe de l'organisation des textes. On verra toutefois dans le chapitre suivant que la RST fournit un moyen de représenter sinon la distinction entre ces plans, tout au moins la portée multiple de la circonstancielle initiale par opposition à la portée purement locale des circonstancielles finales.

Il y a eu dans cet examen du rôle discursif de la syntaxe complexe un glissement depuis la hiérarchisation thématique jusqu'à la relation noyau-satellite. Si la convergence est facile à retrouver en principe dans la mesure où on peut s'attendre à ce que le thème du noyau soit hiérarchiquement supérieur au thème du satellite, le cas de propositions antéposées, donc thématiques, ayant un sujet différent de la proposition principale, peut poser problème[22]. La comparaison devient donc multidimensionnelle, comme essaient de l'illustrer les exemples fabriqués suivants :

(40) Alors que le Japon et les Etats-Unis augmentent régulièrement la part du PNB consacrée à l'enseignement supérieur dans la période 1970-1980, la Grande-Bretagne la réduit de 5%.

[22] Ce problème ne se pose pas dans l'exemple (39), puisque principales et subordonnées partagent le même sujet (non exprimé).

(40') Alors que la Grande-Bretagne réduit de 5% la part du PNB consacrée à l'enseignement supérieur dans la période 1970-1980, le Japon et les Etats-Unis l'augmentent régulièrement.

(40'') La Grande-Bretagne réduit de 5% la part du PNB consacrée à l'enseignement supérieur dans la période 1970-1980, alors que le Japon et les Etats-Unis l'augmentent régulièrement.

(40) et (40') s'opposent par la hiérarchisation syntaxique : dans le premier exemple, *le Japon et les Etats-Unis* (thème du satellite) est un thème secondaire dans une proposition présupposée, introduit par rapport à *la Grande-Bretagne* (thème du noyau), dans le second c'est le contraire. (40) s'oppose par ailleurs à (40'') sur le plan de l'agencement linéaire des constituants, qui confère à la subordonnée un rôle thématique en (40), rhématique en (40''). Dans l'exemple (40) toutefois, le fait que la subordonnée prise dans son entier ait un rôle thématique, quelquefois appelé de "mise en place de la scène" (cf. *scene setting topics*, cf. Chafe, 1976; Lambrecht, 1994; chap. 6 *infra*), ne change en rien le statut secondaire de son thème.

On a vu dans cette partie comment la syntaxe – position de constituants mobiles, structures marquées – constitue un moyen d'indiquer le statut thématique de certains éléments à l'intérieur de l'unité syntaxique. Plutôt que de syntaxe, il faudrait peut-être parler de jeu dans la syntaxe, puisqu'il s'agit de choix qu'elle ne détermine pas, et dont on a justement montré qu'ils étaient motivés par des considérations d'un autre ordre. On a vu également comment certains éléments thématiques ont une portée qui dépasse cette unité syntaxique et contribuent ainsi à la structuration du texte. Il faut cependant constater que cette structuration thématique ne constitue qu'un aspect de l'organisation textuelle, celui qui a trait à l'établissement de "ce dont on parle", et à la continuité qui caractérise des segments de textes. Elle ne peut rendre compte des relations d'ordre sémantique (idéationnel) ou pragmatique (interpersonnel) entre propositions, ni de l'organisation globale des textes, que l'on va aborder maintenant à travers une autre approche théorique et d'autres analyses.

Chapitre 3

Organiser les prédications : structure rhétorique

A la fin de la section 1.3, j'ai organisé mon approche autour de la distinction entre deux niveaux de structuration des textes : "emballage" de l'information en phrases et "emballage" des phrases en texte. Les notions de thème et de structure thématique constituaient l'essentiel de l'outillage conceptuel pour aborder le premier niveau. On a vu à plusieurs reprises que le thème, initialement envisagé par rapport à la structure d'information dans la proposition ou la phrase, débordait ces unités pour prendre une pertinence plus large. L'analyse des circonstants thématiques à portée étendue (2.2.1), de la hiérarchisation des thèmes par la syntaxe (2.3.2), préparait déjà le terrain pour ce qui va m'occuper ici : la représentation de l'organisation globale des textes, et l'identification de marques de cette organisation. Comme d'autres auteurs (Givón, 1983, en particulier), j'ai commencé par me poser la question en termes de regroupements d'unités de premier niveau (pour moi l'unité syntaxique) sous ce que j'ai appelé plus haut un même drapeau thématique (2.2.1). T. Givón parle ainsi de *paragraphe thématique*, où un thème, souvent codé par la fonction grammaticale sujet, représente le participant le plus impliqué dans la suite d'actions et le plus directement lié au thème de niveau supérieur. Cependant, dès qu'on met ce modèle de structure du paragraphe à l'épreuve des données, un certain nombre de problèmes apparaissent :

– il faut tout d'abord rendre compte de structurations thématiques complexes, telles que celles qui résultent dans une suite de phrases de la présence, en plus d'une expression thématique reprise de phrase en phrase en fonction de sujet, de circonstants initiaux au rôle thématique étendu (cf. chap. 6) ;.

– la description de T. Givón évoque immédiatement les récits ("the participant most crucially involved in the action sequence running through the paragraph" (1983:8) ; sa pertinence pour d'autres types de texte paraît problématique. Les textes argumentatifs avec leurs fréquents encadrements énonciatifs (*On constate que…, We have to admit that …*) posent des problèmes spécifiques si on accorde aux sujets de propositions principales le plus important potentiel thématique ;

– il n'est que trop facile de trouver des contre-exemples, des paragraphes parfaitement interprétables malgré une discontinuité thématique flagrante ;

– cette conception thématique d'un niveau intermédiaire d'organisation du texte ne résout pas le problème fondamental du lien entre thème local (ou semi-local) et thème global, celui-ci étant perçu par la plupart des auteurs comme propositionnel (van Dijk, 1981).

Le recours aux notions de thème et de structure thématique permet d'élucider un certain nombre de choix linguistiques à l'interface de la syntaxe et du discours, mais ces notions ne me semblent pas être l'outil conceptuel qui permettra de penser l'organisation globale du texte. C'est à partir de ce constat que j'ai fait appel à des modèles qui placent au centre de la cohérence textuelle l'existence de relations entre propositions ou groupes de propositions, et parmi ceux-ci en particulier à la Rhetorical Structure Theory.

3.1 Rhetorical Structure Theory (RST)

Ce n'est pas sans une certaine appréhension que je me suis orientée vers une approche inévitablement sémantique de questions que j'avais jusque là abordées sous l'angle de la relation forme-structure. Plusieurs raisons m'ont conduite à m'intéresser tout particulièrement au modèle élaboré par W. Mann et S. Thompson (1986; 1988; 1989; 1992). D'abord S. Thompson avait auparavant mené plusieurs études à la fois précises dans la description et d'un grand intérêt théorique sur la relation forme syntaxique-fonction discursive (cf. 2.2.1, 2.3.3). Ensuite, le modèle de la RST se présentait, quand je l'ai rencontré, comme l'héritier d'une longue lignée de travaux sur le discours (Beekman & Callow, 1974; Grimes, 1975; Longacre, 1976; 1979; Hoey, 1983; Martin, 1983) et sur la compréhension en TAL (Hobbs, 1985; Grosz & Sidner, 1986). Et ce modèle était conçu pour la génération de texte, ce qui signifiait qu'autant qu'à la structuration du discours, ses auteurs s'intéressaient à l'interface entre intention communicationnelle et surface textuelle. Cette finalité applicative imposait enfin un niveau de précision non encore atteint dans la formulation du modèle.

La présentation rapide du modèle cherchera à signaler au passage les points de convergence et de divergence avec mes propres intérêts et positions théoriques et méthodologiques. Dans la Partie II (chap. 5), la RST fera l'objet d'une mise en relation avec le modèle de représentation de l'architecture textuelle, mise en relation qui vise à pallier certaines de ses insuffisances.

3.1.1 Notions de base

Deux notions de base fondent la RST :

– la notion de proposition relationnelle ;

– la notion de nucléarité.

• Propositions relationnelles

Selon ce modèle, interpréter comme cohérent le petit texte de l'exemple (41) revient à (r)établir une relation entre ses éléments, relation qui est elle-même de nature propositionnelle : en l'occurrence il faut par exemple poser que la première proposition est le but dans lequel sont données les deux consignes.

(41) [1] Replacer le combiné sur le support mural.
[2] Insérer la base du combiné d'abord,
[3] puis enfoncer fermement la partie supérieure

Les propositions relationnelles sont définies comme des "propositions implicites qui se dégagent des combinaisons de propositions dans le texte" (Mann & Thompson, 1986:88). Elles sont présentées comme étant essentielles à la construction d'une interprétation

cohérente. Les segments reliés par une relation forment un schéma, qui est conventionnellement représenté comme suit :

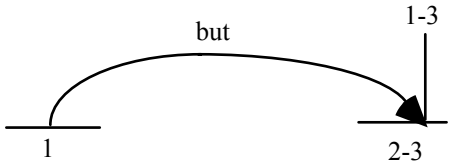


Figure 3.1 : Schéma de relation.

Les segments de texte reliés dans un schéma peuvent être des propositions (l'unité de base) ou des groupes de propositions, comme ici pour les propositions 2 et 3. Les mêmes relations unissent donc des segments de tailles différentes, un texte cohérent devant pouvoir être représenté par une seule relation englobante. Il s'agit donc de relations enchâssées, l'analyse si elle est faite de façon descendante permettant de déplier le texte progressivement jusqu'à l'unité de base.

• Nucléarité

Les schémas distinguent un segment noyau, marqué par une ligne verticale, et indiquent le sens de la relation, du satellite vers le noyau : ainsi, le segment [1], satellite, est dans une relation de but avec le segment [2]-[3], noyau du schéma. La relation noyau-satellite a déjà été évoquée en 2.3.3 en lien avec la relation syntaxique proposition principale-proposition hypotactique, qui en serait, selon C. Matthiessen et S. Thompson (1988), la grammaticalisation. On voit ici que cette relation a une portée plus étendue, puisqu'elle est posée comme caractérisant les schémas RST quelle que soit la nature des segments concernés. Cette asymétrie est une caractéristique fondamentale, à laquelle échappent seulement quelques relations.

• Les relations et leurs définitions

Les relations sont définies en termes de contraintes sur le noyau et le satellite, et d'effet sur le lecteur. Il s'agit donc d'une définition essentiellement pragmatique. Je donne ci-dessous la définition de la relation de but, illustrée par les exemples (41) *supra* et (42) *infra* :

Nom de relation : BUT
Contraintes sur le noyau (N) : présente une activité
Contraintes sur le satellite (S) : présente une situation qui n'est pas réalisée
Contraintes sur la combinaison N + S : S présente une situation qui sera réalisée par l'activité en N
Effet : le lecteur reconnaît que l'activité en N est entreprise de façon à réaliser S
Lieu de l'effet : N et S

Figure 3.2 : Définition de la relation de but dans la RST.

Dans la présentation initiale de la théorie, W. Mann et S. Thompson définissent vingt-trois relations, tout en insistant sur le fait qu'il ne s'agit pas d'un inventaire clos, mais des relations qu'il leur a été nécessaire de définir pour rendre compte de leur corpus. Ces relations sont alors classées en deux groupes – *subject matter* et *presentational* – qui ne sont pas sans rappeler la distinction faite par M. Halliday entre métafonction *idéationnelle* et *interpersonnelle* (Halliday, 1985), ou par T. van Dijk (1977) et d'autres entre fonction *sémantique* et *pragmatique*. Je donne ci-dessous la liste des relations avec ma traduction :

Subject matter relations	Presentational relations
Elaboration (*élaboration*)	Motivation (*motivation*) : accroit le désir
Circumstance (*circonstance*)	Antithesis (*antithèse*) : accroit la considération positive
Solutionhood (*solution*)	
Volitional Cause (*cause, action délibérée*)	Background (*arrière-plan*) : accroit la capacité
Non-Volitional Cause (*cause, action non-délibérée*)	
Volitional Result (*résultat, action délibérée)*	Enablement (*facilitation*) : accroit la capacité
Non-Volitional Result (*résultat, action non--délibérée*)	Evidence (*démonstration*) : accroit la croyance
Purpose (*but*)	Justify (*justification*) : accroit l'acceptation
Condition (*condition*)	Concession (*concession*) : accroit la considération positive
Otherwise (*autrement*)	
Interpretation (*interprétation*)	
Evaluation (*évaluation*)	
Restatement (*reformulation*)	
Summary (*résumé*)	
Sequence (*suite*)	
Contrast (*contraste*)	

Figure 3.3 :Classification des relations dans la RST.

S'ajoute à cette liste un schéma multinucléaire, *Jonction*, qui n'a pas de relation correspondante, parce qu'il s'applique précisément lorsqu'aucune relation n'existe entre les noyaux.

• Un exemple d'analyse RST

(42) [1] Vous pouvez ainsi obtenir les opérations effectuées depuis la date d'établissement de votre dernier relevé de compte.

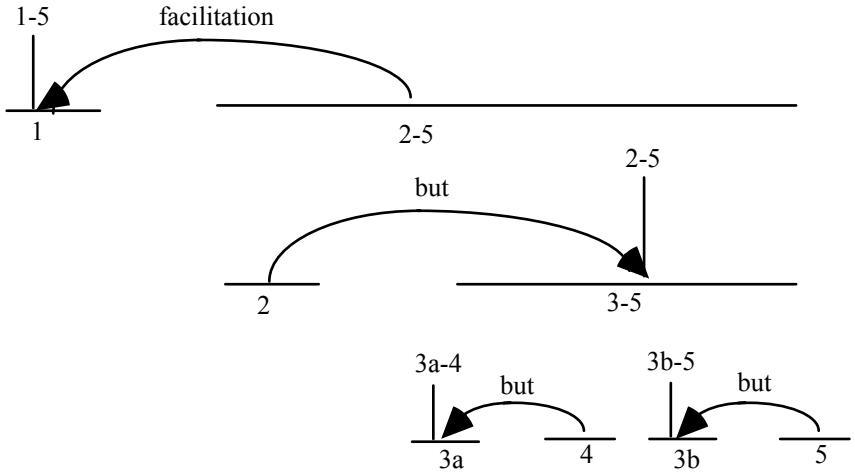
[2] Pour toutes les consulter,

[3a] appuyez sur la touche SUITE

[4] afin de visualiser tous les écrans,

[3b] ou RETOUR

[5] pour revenir en arrière[23].



[23] La numérotation des propositions est différentes de celle de la présentation initiale de cet exemple, de façon à se conformer aux règles de l'analyse RST.

Figure 3.4 : Représentation RST de l'exemple (42).

La seconde partie du petit texte examiné en 2.2.1 (ex. (5)) et reproduit ci-dessus est représentée au premier niveau par une relation englobante de *facilitation* : le segment 2-5 accroît la capacité du lecteur à réaliser l'action nommée en [1]. Le segment [2]-[5] s'analyse ensuite en un satellite (la circonstancielle de but initiale), relié par une relation de but à son noyau ([3]-[5]). Les actions dans[3a] et [3b] sont chacune noyau d'une relation de but dont les circonstancielles finales ([4] et [5]) sont les satellites.

3.1.2 Méthode d'analyse

• Jugements de plausibilité

W. Mann et S. Thompson posent la question du rôle interprétatif de l'analyste dans l'élaboration d'une représentation de la structure rhétorique d'un texte. On l'a vu, la définition des relations repose en partie sur la notion d'effet sur le lecteur, et indirectement sur celle d'intention du scripteur. En l'absence de ces participants, les analystes travaillant sur les textes ne peuvent, disent les auteurs, que proposer des "jugements de plausibilité". Cet aspect interpétatif est d'autant plus important que les auteurs rejettent le recours à des marqueurs pour l'identification des relations[24].

• Signalisation des relations

W. Mann et S. Thompson se démarquent nettement des études, sur la cohésion notamment (Halliday et Hasan, 1976), qui font de certains procédés ou marqueurs – anaphore, conjonction, etc. – une condition de la cohérence. Ils affirment clairement que les relations propositionnelles sont indépendantes de toute signalisation spécifique. Leur attaque de la notion de signal est radicale :

> It is our view that what we have been calling "signals" do not actually "signal" relational propositions in any direct way. A more appropriate description of their function would be that they constrain the interpretation of relational propositions. (…). Our point is that it is the implicit relations which are important, with the conjunctions acting occasionally to constrain the range of possible relational propositions which can arise at a given point in a text. (1986:71)

C'est un argument important, qui met au centre de l'interprétation le principe de cohérence, et critique implicitement les études centrées sur les marqueurs. Toutefois, il me semble lui-même se prêter à la critique : l'attaque de la notion de *signal*, que je disais radicale, ne l'est sans doute pas assez. C'est peut-être parce que W. Mann et S. Thompson sont eux-mêmes prisonniers d'une conception principalement lexicale de la signalisation textuelle qu'ils en diminuent ainsi l'importance. A l'écrit, communication distanciée, le texte constitue indubitablement la majeure partie des indices sur la base desquels le lecteur va construire un modèle interprétatif, modèle changeant qui informe à tout moment l'interprétation du segment en cours de lecture, et qui en est lui-même transformé. Les linguistes ont eu tendance à se focaliser sur certains de ces indices, en particulier les "connecteurs". Il me semble important de remettre la surface textuelle au centre du processus d'interprétation, mais avec une conception élargie de la notion de signalisation. Ce sera là un des objectifs majeurs de la Partie II, et le thème de la synthèse proposée dans la Partie III au chapitre 7. J'insisterai par ailleurs sur la prise en compte du genre discursif et du domaine dans l'identification de la signalisation textuelle, ces paramètres me paraissant essentiels pour deux raisons : d'abord parce que certaines situations de communication, certaines fonctions sociales des textes exigent un guidage précis du lecteur pour limiter au maximum

[24] Après plusieurs années d'analyses dans le cadre de la RST avec un nombre croissant de collaborateurs, les auteurs font toutefois état d'une forte convergence entre les analyses produites par des analystes entraînés (Mann et Thompson, 1992).

sa liberté d'interprétation (consignes par exemple), d'autre part parce qu'on peut s'attendre à ce que les signaux recherchés, qui sont des marques discursives, varient en fonction de ces paramètres (cf. chap. 8).

• Analyse montante, analyse descendante

Les auteurs du modèle indiquent que l'analyse peut être entreprise de façon montante ou de façon descendante. La seule unité définie, et encore de façon très imprécise, est toutefois l'unité de base, "typically the clause". L'analyse montante procède par le regroupement progressif de propositions en schémas, puis de schémas en schémas de niveau supérieur, jusqu'à ce que le texte entier soit représenté par une relation unique. Pratiquement, ce type d'analyse s'avère difficile pour un texte long, et peu intuitif : on aborde le texte de façon descendante, en faisant intervenir des hypothèses d'interprétation liée à l'inscription du texte dans une situation, et en déterminant des macro-segments. Or ces macro-segments ne reçoivent aucune définition dans la RST, que ce soit sur le plan formel ou théorique. Je proposerai au chapitre 5 un modèle permettant de théoriser les intuitions qui conduisent à la délimitation de macro-segments. Là encore, il s'agit de la question centrale de la signalisation, signalisation non seulement des relations entre segments mais aussi des bornes des segments concernés.

3.2 Analyses en corpus de la structure rhétorique

Mon recours à la RST, dans ma recherche de "prises" sur la cohérence des textes du corpus Étudiants, était motivé par la conscience de l'incapacité des modèles axés sur la structure thématique à rendre compte de la structuration du texte à un niveau autre que celui de la phrase ou du groupe de phrases. En dépit de l'avertissement des auteurs de la RST concernant la signalisation, je continuais à vouloir explorer cette question à travers mes analyses RST réalisées sur le corpus Étudiants (Péry-Woodley 1989; 1993a). Je cherchais en effet alors à situer les résultats des analyses des marques de la structure thématique dans un modèle englobant et suffisamment puissant pour rendre compte de la structure enchâssée des constituants du texte. Le modèle RST était capable de rendre compte de façon intuitivement satisfaisante de la cohérence de segments sans continuité thématique. Par ailleurs il était capable, par le biais de la seule notion de relation propositionnelle et de ses vingt trois réalisations, de rendre compte à la fois de relations d'adjacence et de relations de dépendance. L'étude du corpus était motivée et orientée par un ensemble de questions :

– les macro-segments signalés par les circonstants antéposés ("thèmes marqués") sont-ils mis en évidence par la représentation RST ? De même la saillance discursive attribuée aux phrases clivées se retrouve-t-elle sous la forme de noyaux de haut niveau ?

– dans quelle mesure la corrélation entre noyau et proposition principale d'une part, satellite et proposition subordonnée d'autre part est-elle effective dans mon corpus ? Que se passe-t-il si elle ne l'est pas ?

– peut-on faire le lien entre genre discursif et fréquence relative de certains types de relation ?

Dans une perspective plus globale, je cherchais à voir si la représentation RST de textes d'apprenants perçus comme difficiles à interpéter de façon cohérente rendait compte de cette "déviance". Cette idée d'utiliser la représentation de la structure rhétorique dans une perspective "diagnostique" pour des textes "novices" en langue maternelle ou étrangère était alors nouvelle, et a été reprise par d'autres chercheurs, en particulier T. O'Brien (1995).

Les observations résumées ci-dessous répondent à certaines de ces questions, et orientent les recherches qui permettront de mieux répondre aux autres. Elles portent sur un sous-ensemble de mon corpus composé de trois textes en anglais langue maternelle (ALM) et des trois textes en français langue étrangère (FLE) produits par les mêmes sujets. J'ai procédé de manière descendante, sans chercher systématiquement à parvenir à une analyse

complète jusqu'au niveau propositionnel, et à partir d'une délimitation intuitive des macro-segments[25]. Deux textes sont reproduits ci-dessous pour faciliter la consultation de l'analyse, les autres se trouvent dans l'annexe 1.

3.2.1 Analysabilité des textes

Je reprendrai d'abord la dernière question évoquée, liée au processus d'analyse, c'est-à-dire celle du degré de facilité d'obtention d'une représentation. Certains textes s'analysent aisément en macro-segments reliés par des relations immédiatement identifiables, tandis que d'autres, même s'ils se segmentent facilement, ne peuvent être représentés que par des îlots textuels unis dans un schéma multinucléaire sans relation, le schéma *Jonction*. C'est le cas d'un des textes en français langue étrangère, reproduit ci-dessous avec les unités numérotées pour l'analyse, qui est présentée dans la figure 3.5 (cf. Péry-Woodley, 1993a)[26].

Texte FLE-15 :

[1] En Grande-Bretagne nous avons un système d'enseignement supérieur qui donne l'occasion d'étudier à presque tous les jeunes qui le veulent. [2] Après les études au lycée ou dans n'importe quelle école secondaire, les étudiants peuvent continuer à apprendre par l'enseignement supérieur, selon leur habilité. [3] Il y a des universités – [4] on peut dire qu'elles sont les établissements les plus avancés, [5] parce qu'ils exigent plus de leurs candidats; [6] il faut réussir à trois examens avancés dans la dernière année à l'école, par exemple. [7] Il existe aussi les collèges, qu'on appelle "polytechnics", qui n'exigent pas autant que les universités. [8] Et enfin nous avons les "technical colleges" [9] dont les étudiants, avec un esprit plus pratique et plus technique, peuvent profiter. [10] Donc tous les étudiants ont l'occasion de continuer leurs études à force des niveaux différents qui existent dans l'enseignement supérieur. [11] Naturellement pas pour les étudiants qui ne font pas des efforts à l'école et qui n'ont pas l'intention de continuer, [12] malheureusement les établissements supérieurs ne les considèrent pas.[13] Après avoir dit cela, [14] c'est évident qu'on produira une lacune [15] en réduisant des crédits et des effectifs dans ce champs-ci.

[16] Comme j'ai dit, [17] en ce moment-ci, les étudiants anglais peuvent entrer dans l'enseignement supérieur s'ils l'ont envie. [18] Mais en diminuant les occasions qui existent pour les étudiants, [19] beaucoup d'entre eux seraient dépourvus de cette occasion importante. [20] On ne peut point penser à une situation, où des étudiants assez doués auraient gaspillés de temps à l'école en faisant des efforts en vain. [21] Ils ont besoin d'un but à la fin de leurs études, [22] et sans les occasions qui existent actuellement, ils n'auront pas ce but. [23] Alors, il faut maintenir le système des occasions égales, [24] et le cas échéant les étudiants peuvent se rendre compte qu'il y a quelque chose à laquelle leurs efforts aboutissent.

[25] Il faut aussi examiner l'importance de l'enseignement supérieur pour tout le pays, et non pas seulement pour les individus. [26] De nos jours tous les pays ont besoin des esprits intelligents qui peuvent mener les affaires et les activités du pays d'une manière efficace. [27] L'éducation nous assure qu'il existe de telles personnes [28] et une réduction qui limite les produits humains qui viennent des universités et des autres établissements pédagogiques n'est que plus désavantageuse.

[29] D'ailleurs le problème le plus dangereux qui existe en Grande-Bretagne, c'est le chômage. [30] En entrant dans l'enseignement supérieur [31] les étudiants se

[25] Les analyses RST présentées dans ma thèse ont été menées dans l'urgence et avec très peu d'expérience puisque j'ai découvert ce modèle quelque mois avant la soumission de ma thèse. J'y ai, à l'occasion de la rédaction du présent mémoire, relevé plusieurs erreurs.

[26]La figure 3.5 comporte certaines corrections par rapport au schéma présenté dans (Péry-Woodley 1993a), qui la rendent plus conforme à la représentation RST.

donnent l'occasion de trouver du travail, [32] en devenant plus adultes et plus sensibles d'un monde difficile. [33] Ils évitent les difficultés étouffantes qu'impose le chômage.

[34] À conclure, je dois admettre que je suis très déçu à découvrir que le taux moyen d'accroissement annuel des dépenses publiques relatives à l'enseignement supérieur public est si bas en Grande-Bretagne. [35] Assurément, le gouvernement reconnaît la nécessité d'avoir des citoyens bien enseignés dans le pays et que c'est une matière très importante à considérer.

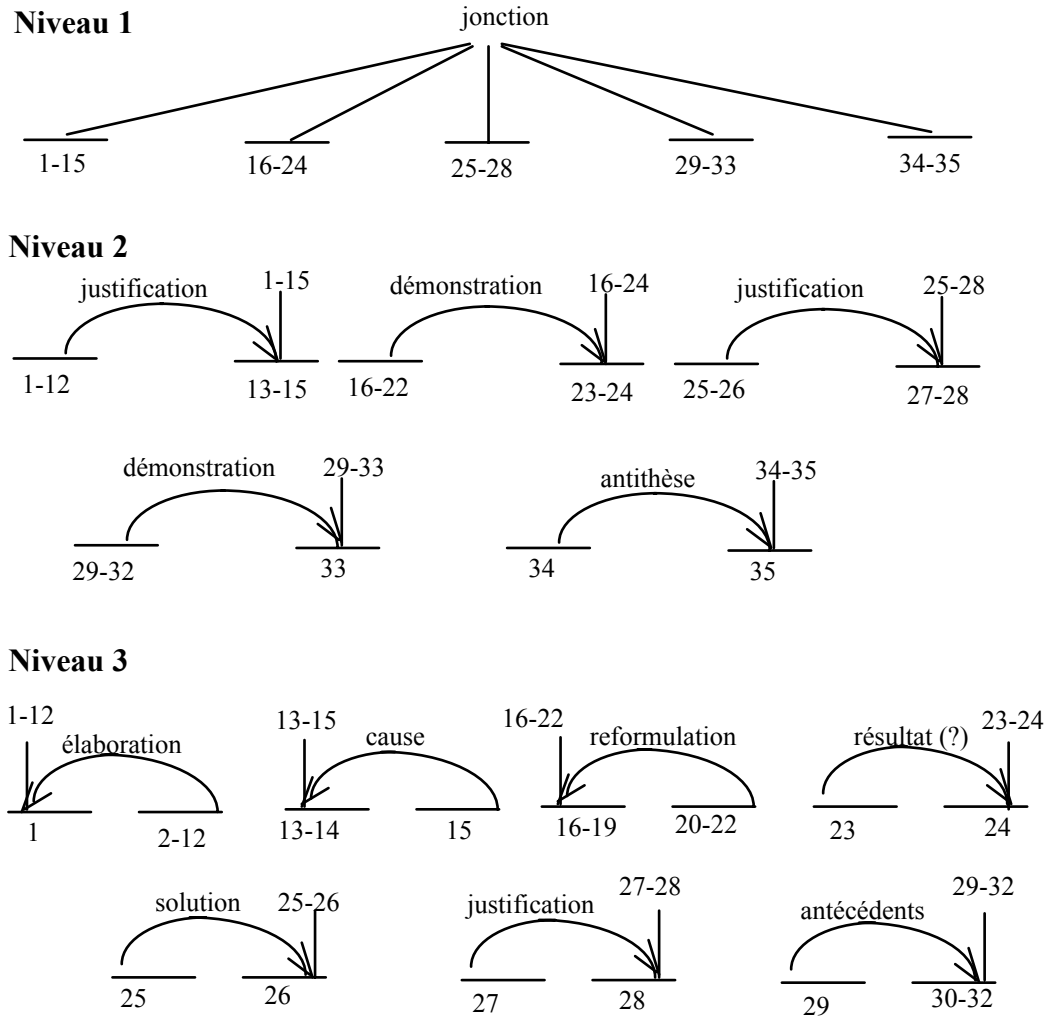


Figure 3.5 : Structure rhétorique du texte FLE-15

La représentation comporte trois niveaux, qui ne fournissent pas une analyse exhaustive puisque le troisième niveau présente des blocs non analysés (2-12, 13-14, 16-19, etc.). J'avais choisi d'effectuer l'analyse de manière descendante pour faire apparaître les grands segments et leurs relations : le premier niveau englobe l'ensemble du texte (unités 1 à 35), le niveau 2 "déplie" chacun des grands segments délimités dans le niveau 1, le niveau 3 "déplie" à son tour les sous-segments, dont certains sont encore des segments complexes. On constate tout de suite le contraste entre la diversité des relations aux niveaux 2 et 3 et le schéma sous-spécifié du niveau 1. En effet, alors que les segments étaient faciles à délimiter et à analyser, aucune hiérarchisation et aucune relation rhétorique claire n'apparaissaient

pour les relier entre eux. Cette difficulté d'analyse reflète bien la difficulté d'interprétation ressentie à la lecture de ce texte.

3.2.2 Circonstants thématiques et délimitation de segments

Là où les circonstants initiaux ont une portée dépassant la phrase, ils délimitent des segments textuels dont on peut se demander s'ils correspondent à des segments RST. Les circonstants en position initiale sont utilisés de façon apparemment semblable dans les sous-corpus ALM et FLE : ils marquent assez systématiquement la borne initiale d'un segment. A y regarder de plus près cependant, on observe une différence de fonctionnement liée sans doute à la nature différente de ces circonstants dans les deux sous-corpus. Il s'agit principalement de propositions conditionnelles ou concessives encadrant une argumentation dans ALM, alors que les circonstants de FLE sont souvent des circonstants de temps ou de lieu qui situent chronologiquement ou géographiquement des données tirées des tableaux statistiques (cf. 2.2.1). Une certaine régularité semble se dégager dans ALM, où ces circonstancielles ouvrent souvent des macro-segments noyaux de relations "argumentatives" telles qu'antithèse, démonstration ou justification. On va examiner ce fonctionnement plus en détail dans le texte ALM-15, texte en anglais langue maternelle rédigé par le même sujet que le texte FLE-15 examiné dans la section précédente :

Texte ALM-15

[1] The issue of whether higher education should be expanded or contracted in Great Britain seems to be purely a question of the financial and economic implications of such changes. [2] Its expansion would doubtless mean that a large amount of money would have to be distributed in addition to that already being used to enable such a move. [3] Conversely, one can assume that the contraction of the higher education system in Great Britain would result in the saving of large sums of money and the country would benefit accordingly. [4] This appears to be dependent upon logic [5] and it is thought that surely the most logical step to take would be a decrease in the amount of financial support given to higher education and the establishments which represent it.

[6] In developing the question, however, [7] one begins to become aware of the distinct advantages of a thriving educational system which overshadow the harshnesses of such economically-minded destruction. [8] Firstly, the system of higher education used in this country can be of great benefit to the individual. [9] On entering into the daily existence which higher education promotes, [10] he or she is being afforded the chance of self-fulfilment and a further development of their personality, [11] which only they can individually express. [12] In most cases, higher education is enjoyed by the student on a basis which is independent of the influence of parents, other than financially. [13] This evidently results in a greater responsability being placed on the student's shoulders [14] and in such a position he can but mature and adopt a greater sense of social awareness. [15] Unfortunately, if the system of higher education was dipped in such a way that the number of students eligible for it was lowered, [16] then a percentage of those leaving school with the intention of studying at a higher level would be denied the benefits of the system [17] and maturity for them would have to come in another way. [18] In this case, it must be asked if a young person leaving secondary school is capable of accepting the pressures and responsibilities which a job would bring to bear. [19] Presumably there are those who have had to mature more quickly because of a greater responsibility imposed at home or in some other way, [20] but I would say that, on the whole, the average school-leaver is not and has not been prepared for these demanding positions, [21a] and it is only through higher education, [22] whether it is in the form of training or academic study, [21b] that he will develop this sense of maturity and fulfilment as an individual. [23] Secondly, it is important for Great Britain as a whole to have a system of higher education which is allowed to flourish

and not a system which is questioned and curtailed by doubt. [24] Higher education is a means of reinforcing the quota of the nation's intellect [25] and the country should have a system, whereby the quest for knowledge is encouraged and not discouraged simply because of the financial situation which prevails. [26] It should be given a certain priority as a result [27] and financial constraints should not be imposed on it.

[28] It can be seen from the statistics that the percentage of people within the 19-24 age group who are studying is lower in Great Britain. [29] Whilst this may say something about the length of courses in other countries, [30] the fact still remains that the lower the percentage, the smaller the number of well-educated, mature people there are being trained for life after school. [31] It is on that note that I wish to conclude [32] by saying that the system of higher education should be promoted and encouraged and not stifled by financial restraint, simply because the government has not taken personal situations into account.

La figure 3.6 représente le premier niveau d'analyse du texte. En contraste avec l'absence de structure du texte FLE-15, on trouve ici un important segment noyau ([6]-[27]) qui entre en relation d'antithèse avec un segment initial ([1]-[5]) et de résumé avec le segment conclusif ([28]-[32]) :

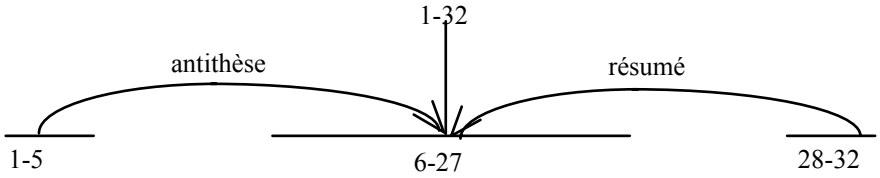


Figure 3.6 : Niveau 1 de la structure rhétorique du texte ALM-15.

Ce segment central s'ouvre bien par une proposition circonstancielle, suivie d'un connecteur :

(43) [6] In developing the question, *however, [7] one begins to become aware of the distinct advantages of a thriving educational system…*

On ne peut vraiment parler ici de circonstant de portée étendue, mais plutôt du marquage d'une borne de segment de haut niveau. On trouve dans ce même texte des exemples de circonstants initiaux de portée étendue, portée qui effectivement constitue un segment dans l'analyse RST, mais ces cas concernent des segments d'assez bas niveau. Ainsi l'exemple (44) donne le début du segment [15]-[17] :

(44) [15] Unfortunately, if the system of higher education was dipped in such a way that the number of students eligible for it was lowered, …

Autant que par leur portée, ces circonstancielles ont un rôle structurant par l'effet de rupture qu'elles créent avec ce qui précède. Elles semblent donc jouer un rôle de signalisation des macro-segments de premier niveau dans ces textes argumentatifs. Cette analyse éclaire d'un nouveau jour la question du paradoxe entre statut de satellite et importance dans la structuration thématique : on voit en effet comment ces propositions, satellites au niveau de l'analyse phrastique, ouvrent des segments majeurs au niveau du texte. Cette fonction n'apparaît pas dans les textes en FLE, inévitablement peut-être puisqu'ils ne sont pas structurés de façon argumentative et qu'il est difficile de relier des macro-segments par une relation englobante.

3.2.3 Syntaxe et structure rhétorique

Des observations un peu plus précises peuvent être rapportées sur la relation entre le statut grammatical d'une proposition et son statut de noyau ou satellite ("nucléarité"). Comme on l'a vu, la relation de nucléarité s'applique à tous les niveaux d'analyse de la

structure rhétorique, mais au niveau propositionnel, elle serait grammaticalisée, et donc "codée" par la relation proposition principale-subordonnée (2.3.3). Deux problèmes semblent se faire jour dans les textes en français langue étrangère : soit le texte consiste en une succession de phrases simples , et il n'y a donc pas de structuration ni de signalisation par la syntaxe ; ou les constructions hypotactiques sont utilisées à mauvais escient et envoient des signaux trompeurs. Les exemples (45) et (46) illustrent le premier cas de figure, (47) et (48) le second :

(45) [25] Il faut aussi examiner l'importance de l'enseignement supérieur pour tout le pays, et non pas seulement pour les individus. [26] De nos jours tous les pays ont besoin des esprits intelligents qui peuvent mener les affaires et les activités du pays d'une manière efficace. [27] L'éducation nous assure qu'il existe de telles personnes [28] et une réduction qui limite les produits humains qui viennent des universités et des autres établissements pédagogiques n'est que plus désavantageuse. (FLE-15)

(46) [19] il est intéressant de voir que [20] de 1965 à 1974, il y a eu un grand changement aux Etats-Unis en ce qui concerne le pourcentage du PNB – de 0,90 à 1,62% – [21] mais en Grande-Bretagne il a eu peu de changement par rapport aux autres pays, seulement de 0,64 à 0,89%. (FLE-2)

En (45), les élaborations de la phrase initiale, les propositions [26] et [27], sont syntaxiquement sur le même plan que la conclusion qui en est tirée en [28]. Au lecteur de structurer cette suite de propositions non hiérarchisées. En transformant (46) en (46'), on peut, par le biais d'une hiérarchisation syntaxique, aider le lecteur à envisager les chiffres concernant la Grande-Bretagne comme centraux, ceux concernant les Etats-Unis comme auxiliaires :

(46') [19] il est intéressant de voir que [20] de 1965 à 1974, alors que les Etats-Unis ont augmenté le pourcentage du PNB consacré à l'enseignement supérieur de 0,90 à 1,62%, [21] en Grande-Bretagne il y a eu peu de changement par rapport aux autres pays, seulement de 0,64 à 0,89%. (FLE-2)

Dans le texte FLE-15 encore, la politique de réduction des crédits dans l'enseignement supérieur, qui est au cœur du problème soulevé dans le sujet de la dissertation, n'apparaît que dans des propositions hypotactiques réduites, donc satellites, et sans sujet exprimé :

(47) [14] c'est évident qu'on produira une lacune [15] ***en réduisant des crédits et des effectifs dans ce champs-ci****.*

(48) Mais ***en diminuant les occasions qui existent pour les étudiants****, beaucoup d'entre eux seraient dépourvus de cette occasion importante.*

Pour clore cet examen de la relation entre syntaxe et structure rhétorique, un bref regard sur les phrases clivées du texte ALM-2, envisagées sur le plan de la saillance dans la section 2.2.2:

(49) [4] Table one shows the rate of scholarization (...). [5] Compared with the US and Italy, [6] (...), [7] Great Britain's percentage (...). [8] In fact, of all the countries mentioned, ***it is Great Britain which*** *has the lowest percentage (...) in higher education.*

[9] The average age (...) is represented in table two. (...) [10] The United States' figures [11] (...). [12] However, Great Britain's figures (...). [13] So ***it is again the British education system which*** *seems to be contracting the most compared with the other countries.*

[14] The average rate of annual increase of public expenditure on higher education is shown in table three. [15] The figures clearly reveal [16] (...). [17]Again ***it is the British education system which*** *is contracting the most in view of the amount of money consecrated to it. (ALM-2)*

Ainsi que j'en avais fait l'hypothèse, ces clivées fonctionnent de façon régulière, au deuxième niveau de l'analyse de ce texte, comme noyaux de segments formés par une relation de démonstration.

J'ai adopté ici une approche exploratoire et illustrative pour aborder le lien fonctionnel entre certains phénomènes syntaxiques et la structure rhétorique. Les exemples suggèrent que des études plus systématiques de l'occurrence de certaines structures syntaxiques par rapport à une représentation de l'organisation textuelle pourraient être révélatrices. De telles mises en relation sont actuellement rares dans la littérature.

3.2.4 Types de relation et genre discursif

La méthode d'analyse et de représentation des textes fournie par la RST suggère une approche nouvelle de la typologie des textes : on pourrait en effet avancer l'hypothèse d'une prédominance de certaines relations en corrélation avec le genre discursif. Dans l'établissement d'une telle corrélation, il me semble qu'il sera important de distinguer les stades de l'analyse. Pour la caractérisation d'un texte par rapport à un genre discursif, les relations de premier niveau – celles qui unissent les macro-segments constitutifs de l'ensemble du texte – me paraissent intuitivement plus significatives que les relations de dernier niveau, entre propositions. Il s'agit là toutefois de simples intuitions qui nécessiteraient d'être validées par des analyses extensives. Je reviendrai sur le classement des textes en types et genres dans le chapitre 8. Ici, cette approche de la typologie présente une pertinence spécifique pour l'étude du corpus Étudiants. J'ai suggéré à plusieurs reprises que les textes en langue étrangère semblaient moins bien se conformer à la consigne de rédaction que les textes en langue maternelle. Cette consigne exigeait un texte argumentatif, et la plupart des textes FLE tiennent davantage du commentaire. J'ai donc cherché à voir dans quelle mesure la représentation RST faisait apparaître cet écart en termes d'usage de relations rhétoriques différentes. L'analyse RST du corpus Étudiants est extrêmement limitée : trois textes ALM et trois FLE. Par ailleurs, la sous-spécification du schéma de premier niveau pour certains textes FLE (cf. 3.2.1) rend la comparaison difficile. Il en ressort tout de même une forte prédominance de relations de type présentationnel dans les textes ALM : démonstration, justification, antécédents, antithèse. Ces relations si elles sont très présentes également en FLE, sont en compétition dans ce corpus avec des relations référentielles, en particulier condition, contraste, cause, résultat. Ces observations tout à fait préliminaires vont dans le même sens que celles issues de l'analyse des circonstants thématiques, cadre argumentatif pour ALM, temps et lieu pour FLE.

À travers l'analyse d'un corpus de textes d'étudiants en langue maternelle et en langue étrangère, j'ai tenté d'identifier au plus près de la surface textuelle certains traits qui peuvent être associés à la cohérence globale. La cohérence n'étant pas une propriété d'un texte, mais le résultat d'une mise en discours, il s'agit de choix, principalement d'ordre syntaxique, susceptibles de guider le lecteur dans l'identification des thèmes et dans l'interprétation des liens entre les prédications sur ces thèmes. Ce qui peut être retenu de ce travail, plus que les résultats de l'analyse de corpus, c'est l'élaboration pas à pas d'un cadre permettant d'aborder la relation entre des agencements de constituants dans la proposition, de propositions dans la phrase, de phrases dans les segments. Ce cadre repose principalement sur deux grandes approches de la modélisation de l'organisation textuelle : la relation de thème – et la structuration qui en découle –, et la structure rhétorique. Dans les deux cas, il s'agit de théorisations en cours d'élaboration, mouvantes, et souvent difficiles à traduire en observations précises. La signalisation permettant l'identification du thème d'un énoncé, d'une relation rhétorique, ne tient pas du codage, mais plutôt de l'exploitation de cooccurrences de traits. Cette conception de la signalisation de l'organisation textuelle se précisera dans la deuxième partie dans le cadre d'un travail plus pointu, centré sur la définition comme objet textuel. Elle fera l'objet d'une réflexion synthétique dans la

troisième partie (chap. 7). Les deux grandes approches explorées dans cette première partie seront reprises, retravaillées par le contact avec d'autres théories, par l'analyse d'autres corpus. La nécessité d'articuler les différents niveaux d'organisation textuelle, que je continuerai à envisager dans les termes des trois métafonctions hallidayennes, m'amènera à confronter la RST au modèle de l'Architecture Textuelle (chap. 5 et 6). La notion de thème, renommée *topique*, sera reprise à la lumière de travaux récents au chapitre 6.

Partie II

Une structure dans des textes : la définition

L'objectif initial de mon travail sur le texte était de me donner des "prises" sur la cohérence textuelle, de manière à pouvoir mettre en relation des choix de formulation à la surface du texte avec cette "qualité" globale. Il est apparu clairement cependant que la cohérence ne pouvait être vue comme une propriété du texte, mais plutôt de sa mise en œuvre dans le discours ; en conséquence, il ne peut être question dans le texte que d'un potentiel de cohérence, potentiel qu'on peut aborder en termes de signalisation de l'organisation textuelle. La première Partie de ce mémoire a été consacrée à l'élaboration d'une approche théorique et méthodologique de différents aspects de cette organisation. Chaque étape de la mise en place de cette approche s'est accompagnée d'analyses dont

l'objectif peut se résumer comme étant d'identifier les procédés de marquage du thème et des relations rhétoriques dans des textes. Il s'agissait donc d'une démarche ouverte, de type exploratoire, fondamentalement orientée par le choix d'aller de la fonction aux marqueurs plutôt que de partir de marqueurs pré-identifiés. Cette démarche a fourni un ensemble de pistes dont certaines vont maintenant être reprises dans des cadres plus contraints, à la fois sur le plan des textes envisagés, et des aspects de l'organisation textuelle pris en compte. Cette deuxième Partie du mémoire va se focaliser sur un objet textuel : la définition. Dans un premier temps, le fonctionnement textuel de la définition sera examiné dans un corpus de textes courts dont chacun constitue une définition. J'envisagerai ensuite ce fonctionnement en contexte dans une étude de définitions intégrées dans des textes dont les visées discursives sont diverses.

Cette deuxième phase s'inscrit dans la continuité des travaux résumés dans la première partie, tout en y ajoutant de nouvelles dimensions, tant sur les plans théorique et méthodologique que sur celui des visées applicatives :

– sur le plan théorique, aux modèles de l'organisation textuelle axés sur la relation thème-rhème et sur les propositions relationnelles (RST), va venir s'ajouter le modèle de la représentation de l'architecture textuelle, qui sera présenté dans le chapitre 5. Ce modèle est d'un intérêt tout particulier puisqu'il est centré sur la notion même de signalisation de l'organisation textuelle ;

– sur le plan méthodologique, et parce que je travaille sur corpus et par conséquent en discours, la question de l'indépendance des marqueurs identifiés par rapport au domaine et au genre discursif sera plus systématiquement examinée ;

– enfin, en ce qui concerne les visées applicatives, elles s'inscrivent pour cette deuxième phase dans le cadre de l'intelligence artificielle et du traitement automatique des langues. Les implications de cette orientation sur la nature de la modélisation des structures textuelles étudiées seront développées au fur et à mesure.

Chapitre 4

Caractérisation linguistique pour une modélisation cognitive

4.1 Le projet MIEL (Modélisation Inductive de l'Elève selon son Langage)

• Présentation générale

Le projet MIEL se situe à une période de transition dans mon travail : j'entreprends en 1989 une formation en sciences cognitives, qui me donne l'occasion de passer cinq mois au sein de l'équipe *Langage et Cognition* du LIMSI, pour y prendre part à une étude sur la modélisation de l'utilisateur pour un système tuteur intelligent[27] (Daniel *et al.*, 1992). Le système, appelé TEDDI (Tuteur d'Enseignement De Définitions Individualisé), a pour but de fournir un contrôle des connaissances, mais surtout une aide dans la formulation de définitions de concepts présentés préalablement dans le cadre d'enseignements formels. Les définitions doivent répondre à des critères non seulement d'exactitude, mais de niveau d'abstraction et de généralisation. L'approche classique, focalisée sur la mise en relation de représentations des connaissances du domaine et des croyances de l'utilisateur par rapport à ces connaissances (Sleeman, 1982; Kobsa & Wahlster, 1988; Kass, 1989) est ici insuffisante. Au-delà de la représentation de ce qui est *dans* la définition, le module de modélisation de l'apprenant dans TEDDI est conçu pour exploiter *comment* cette définition est formulée. Il doit extraire de la formulation linguistique des définitions des informations sur le "style cognitif" de l'étudiant. L'intuition de départ des initiatrices du projet (Nicaud & Prince, 1990), est en effet que le "style cognitif" est inférable à partir du "style linguistique".

[27] Je remercie le directeur de l'équipe, Gérard Sabah, et les initiatrices du projet, Violaine Prince et Lydia Nicaud, auprès de qui j'ai beaucoup appris.

Mon rôle dans ce projet, pour lequel j'avais élaboré dans mes travaux précédents des outils pertinents, est de définir le "style linguistique" en termes de marqueurs identifiables à la surface des textes.

L. Nicaud et V. Prince étaient guidées sur la piste du "style linguistique" par l'observation préliminaire de marqueurs récurrents, qu'elles interprétaient en termes cognitifs. Sur la base des travaux décrits dans la Partie I, j'envisage le "style linguistique" comme l'ensemble des marques résultant d'opérations de mise en texte, et constituant la signalisation de l'organisation textuelle. Je reprends en particulier les deux processus fondamentaux dans la construction d'un texte, d'une part l'établissement et la continuation des thèmes des propositions formant le texte, d'autre part la hiérarchisation de ces propositions et la création de liens, eux-mêmes de nature propositionnelle, entre segments (propositions ou groupes de propositions). Comme dans la partie I, et en réponse aux besoins de l'application visée, je vais partir des fonctions textuelles pour rechercher les marqueurs qui, dans le corpus, leur sont régulièrement associés.

Si le cadre d'analyse reste proche de celui élaboré dans les travaux précédents, les objectifs et la nature du corpus entraînent une orientation différente. Il ne s'agit plus de rechercher dans des textes longs et discursivement complexes tout ce qui peut constituer une marque de structuration thématique ou rhétorique, mais de caractériser sur le plan linguistique des textes courts, et beaucoup plus simples dans la mesure où chacun est uniquement censé formuler une définition. Il s'agit en fait de répertorier les réalisations d'un acte de parole (cf. Flowerdew, 1992), avec l'intérêt méthodologique supplémentaire que cet acte a été réalisé dans des conditions suffisamment contraignantes pour en rendre les diverses manifestations comparables entre elles. Les travaux résumés ci-après sont exposés dans (Daniel *et al.,* 1992; Péry-Woodley, 1993b, 1994).

• Le corpus

corpus	**domaine**	**termes à définir**	**sujets**	**taille**
PSYCHO	psychologie cognitive	*récupération spontanée* *conditionnement* *extinction* *réapprentissage* *constance perceptive* *amorçage sémantique*	12	72 déf. 1 887 occ.
INFO1	informatique	*itération* *module objet*)	53	106 déf. 2 615 occ.
INFO2	informatique	*tri*	121	121 déf. 2 933 occ.
MANA	gestion	*tableau de financement*	39	39 déf. 1 157 occ.

Tableau 4.1 : Corpus pour le projet MIEL[28].

Le corpus de référence (tableau 4.1) est constitué de définitions rédigées par des étudiants, à la demande des enseignants et durant les cours, en réponse à la question : *Qu'est-ce que SN ?* ou *Indiquer la définition de SN*. La consigne était de produire des définitions de type "définition de dictionnaire". Les SN en question sont des termes ayant fait l'objet de définitions explicites en cours. Étant donné la nature fortement contrainte de la situation de production des textes et de leur visée discursive, je considère le corpus comme

[28] Des extraits du corpus sont fournis dans l'annexe 3.

homogène sur le plan du genre discursif[29]. Le projet initial concerne des définitions dans le domaine de la psychologie cognitive. De manière à tester l'impact du domaine sur la formulation des définitions, trois autres recueils sont effectués, auprès d'étudiants en informatique et en gestion.

4.2 Identifier des marqueurs pertinents et repérables de façon automatique

Le travail entrepris dans le cadre du projet MIEL se distingue des études précédentes par deux points importants :

– il s'agit de caractériser linguistiquement des définitions pour une modélisation cognitive ;

– dans le cadre d'un système tuteur intelligent, les marqueurs identifiés doivent permettre un repérage automatique le moins coûteux possible.

Etant donné que l'interprétation cognitive des marqueurs allait être confiée à la psychologue du groupe, ma tâche était d'identifier et de classer les procédés linguistiques utilisés par les sujets pour introduire le thème de leur définition d'une part, pour la structurer d'autre part. La caractérisation des définitions est donc abordée en deux temps : d'abord l'identification de ce qui est présenté comme thème de la première phrase du texte, appelée l'*attaque* de la définition (4.2.1), ensuite l'analyse de la *structure* de la définition, c'est-à-dire des relations liant les propositions successives, et des marqueurs de ces relations (4.2.2).

4.2.1 L'attaque des définitions

Dans une première approche, me fondant sur l'association privilégiée entre thème phrastique et fonction grammaticale sujet d'une part, position initiale d'autre part, j'ai commencé par examiner le point de départ des définitions, l'*attaque*. Deux stratégies opposées se font jour selon que le point de départ de la définition est ou n'est pas le concept à définir. Sont considérées comme prenant pour point de départ le concept à définir :

– les réponses qui en font le sujet grammatical, ex. *Le conditionnement est l'étude des modifications d'un comportement ...* ;

– les réponses elliptiques – nominales –, que le concept à définir apparaisse en titre ou non, ex. *Apprentissage d'un type de comportement ...*

– les réponses commençant par *C'est*, où le concept à définir peut apparaître en titre, sous la forme d'un SN disloqué à gauche, ou pas du tout , ex. *C'est l'acquisition d'un comportement ...*

Ainsi, (1) et (2) s'opposent à (3)-(5) sur le plan de l'attaque : le concept à définir n'est le point de départ de la définition que pour (3)-(5). Les cinq exemples sont les *attaques* de réponses à la question : *Qu'est-ce que la récupération spontanée ?* :

(1) L'animal *est placé dans sa cage. Au bout d'un certain temps on lui présente le stimulus son il va se mettre à baver.*

(2) Après extinction et une période de repos, on *présente de nouveau le stimulus conditionnel à l'animal, on constate de nouveau la réaction conditionnelle.*

(3) Réobtenir *une réponse conditionnée, après extinction, sans qu'il y ait présentation du renforçateur.*

(4) Phénomène *correspondant à la levée naturelle de l'inhibition inhérente à l'extinction d'un conditionnement.*

(5) Réapparition *d'une réponse conditionnelle, ayant subi une extinction, sans renforcement.*

[29] La notion de genre discursif fera l'objet d'une réflexion dans le chapitre 8.

En (1) et (2), le sujet grammatical est un "personnage" intervenant dans le "récit" d'un processus ; les exemples (3)-(5) établissent une relation d'équivalence entre le concept à définir – qui fournit le thème, implicite ici, de la définition – et un segment de texte. Une première classification distingue donc des *définitions-processus,* et des *définitions-concept*. Les définitions-concept font ensuite l'objet d'un classement selon leur expression plus ou moins verbale (tournée vers l'action) ou nominale (tournée vers la conceptualisation). Ces différences peuvent se placer le long d'un continuum, mais la caractérisation des définitions, associée à des marques formelles précises, est nécessairement discrète (tableau 4.2) :

PROCESSUS <----------> CONCEPT

1. histoire L'animal est placé dans sa cage.	**2. action** *Réobtenir* une réponse conditionnée…	**3. générique (processus)** *Phénomène* correspondant à la levée naturelle de l'inhibition …	**4. synonyme (concept)** *Réapparition* d'une réponse conditionnelle …
	(V+Prép+)Inf N+V≠cop	N + Rel N + Prép + Inf N + V-ant	N + SPrép N + SAdj [30]

Tableau 4. 2 : Marqueurs de l'attaque des définitions.

Le type "histoire" ne présente aucune des caractéristiques formelles typiques de la définition et ne peut être identifié que par défaut. On verra que des définitions de ce type peuvent "se rattraper" dans certains cas par la présence de certaines relations dans la suite. Les trois autres types sont caractérisés en fonction de traits syntaxiques qui sont perçus comme reflétant des approches différentes de la tâche. Force est de constater que les différents types se répartissent inégalement dans les trois sous-corpus, faisant intervenir la nature sémantique des termes à définir, sinon le domaine, dans les choix de formulation. Ainsi la question *Qu'est-ce qu'un tri ?* reçoit plus de réponses de type 2 que la question *Qu'est-ce que la récupération spontanée ?*. Il est cependant intéressant de constater que même le tri (INFO2) est défini par plusieurs sujets selon les types 3 et 4, ce qui exclut l'hypothèse d'une détermination absolue du choix de formulation par les caractéristiques sémantiques du terme :

(6) Tri : Ordonner un ensemble d'éléments selon un ordre précis (type 2).

(7) Un tri est un mécanisme qui permet d'ordonner un ensemble d'éléments de même type suivant un critère donné (type 3).

(8) Un tri est l'ordonnancement d'éléments dans un ordre voulu (type 4).

Les définitions de type 3 et 4 sont caractérisées par la prédominance d'éléments nominaux : le terme à définir est mis en relation soit avec un hyperonyme (*genus*) soit avec un synonyme, toujours suivi d'un modifieur. Dans ce corpus, les SN à modifieur de type adjectival ou prépositionnel semblent être toujours des synonymes, alors que les SN à modifieur verbal ou propositionnel (relatives, etc.) sont des hyperonymes dont le modifieur exprime les *differentiae*. La consultation de dictionnaires de référence indique que seuls ces types 3 et 4 sont représentés dans les définitions lexicographiques. Ces formulations signalent chez le scripteur une capacité à conceptualiser le terme à définir de façon à pouvoir ensuite l'utiliser comme thème d'un énoncé. C'est sans doute dans ce sens que le tuteur devra guider les étudiants. Mais ces décisions sont du ressort des psychologues du groupe de recherche, et mon examen linguistique des définitions, qui s'est limité jusqu'à présent à leur *attaque*, n'est pas terminé.

[30] V= verbe; Inf = infinitif; N = nom; cop = copule; Rel = proposition relative; Prép = préposition; V-ant = participe présent; SPrép = syntagme prépositionnel; SAdj = syntagme adjectival.

4.2.2 La structure des définitions

Dans cette deuxième partie de l'analyse, je tente de rendre compte de l'agencement des propositions constitutives de la définition et des relations qui les relient. La structure des définitions est abordée à travers deux modèles de l'organisation textuelle : celui de W. Mann et S. Thompson, la Rhetorical Structure Theory (cf. chap. 3), et celui de K. McKeown (1985). Ce dernier, élaboré spécifiquement pour la génération automatique de définitions, est plus restrictif dans la mesure où, au lieu de poser une structure récursive souple comme dans le modèle de la RST, il organise les relations en schémas textuels associant relations obligatoires et facultatives. Ces schémas fournissent un cadre pour l'intégration des relations rhétoriques en structures "canoniques". K. McKeown précise bien que ses schémas ne sont pas des prescriptions, mais une modélisation pour la génération de "common patterns of text structure" (1985: 53). Il m'a paru intéressant, dans mon optique de caractérisation de définitions d'apprenants, de confronter mes données à son modèle.

Reprenant la notion élargie de marqueur textuel élaborée dans la Partie I, et à partir d'une analyse interprétative de type RST, j'ai identifié des séries de marqueurs associés à trois types de relations : *identification, explicitation-illustration et situation-explication*. Les deux dernières sont des sous-classes d'élaboration (Mann & Thompson, 1988; Hovy, 1990). La première s'inspire de la relation d'identification définie par K. McKeown comme "the identification of an item as a member of some generic class" (1985:21). Cette définition me semblait problématique parce qu'elle traitait non pas d'une proposition relationnelle entre deux segments mais bien du contenu propositionnel lui-même. Ma tentative de redéfinition (Péry-Woodley, 1993b, Pascual & Péry-Woodley, 1995) me semble toutefois, *a posteriori*, peu convaincante, et je vais dans cette présentation corriger ce qui m'apparaît maintenant comme erroné : l'identification constitue le noyau de la définition (dans le sens pris par ce terme dans la RST), auquel les autres propositions se rattachent par des relations d'explicitation-illustration ou de situation-explication. L'identification ne constitue pas en soi une relation, si ce n'est avec la question[31]. L'identification (en gras dans les exemples) peut ouvrir la définition, comme en (9) et (10), ou la clore, comme en (11) et (12). Pour cette analyse de la structure des définitions, je considère que le texte est constitué par l'ensemble question-réponse. Les propositions sont numérotées pour l'analyse RST ([Q] = question ; [R1]-[Rn] = réponse) :

> *(9) [Q] Qu'est-ce que le conditionnement? [R1]* ***C'est l'acquisition d'un comportement nouveau en réponse à un stimulus neutre*** *(ne provoquant pas de réponse initialement). [R2] Il s'obtient à l'aide d'expériences répétées et par des renforcements du stimulus neutre (appelé stimulus conditionnel).*
>
> *(10) [Q] Qu'est-ce qu'un module objet? [R1] Module objet:* ***c'est le résultat de la compilation d'un programme source,*** *[R2] le module objet peut être lié pour obtenir du code exécutable.*
>
> *(11) [Q] Qu'est-ce que la récupération spontanée? [R1] Après avoir observé une extinction et après une période de repos si on recommence une série de tests, on observe une RC à la présentation de SC[32] : [R2]* ***c'est la récupération spontanée de l'apprentissage.***
>
> *(12) [Q] Qu'est-ce qu'un module objet? [R1] Module objet : Un programme, tel qu'il est écrit dans un langage évolué, n'est pas compréhensible par l'ordinateur. [R2] Il doit pour cela être traduit en une séquence d'instructions plus élémentaires que celui-ci pourra comprendre. [R3]* ***Un module objet est une telle séquence d'instructions.***

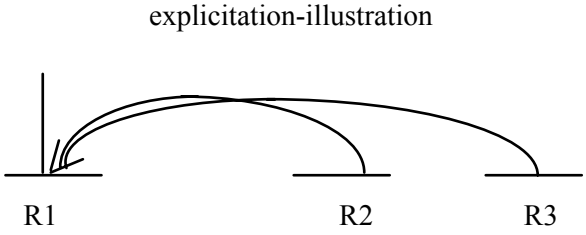
[31] L'identification comme relation entre la question et la reformulation définitoire à l'aide d'un hyperonyme m'est apparue un temps comme une solution à retenir. Mais c'est l'ensemble de la définition, et pas seulement la reformulation du terme, qui est en relation avec la question.

[32] RC = réaction conditionnelle ; SC = stimulus conditionnel.

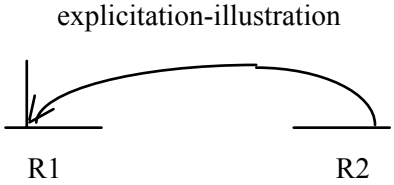
Il convient donc de distinguer les marqueurs d'identification initiale, qui se confondent avec ceux des attaques 3 et 4, et les marqueurs d'identification finale : *c'est + terme à définir* ou *terme à définir + copule*. C'est à cette identification finale que je faisais référence dans la section précédente en disant que des définitions "mal attaquées" (type 1) pouvaient "se rattraper" par la suite, comme le font (11) et (12).

La relation d'**explicitation-illustration** lie deux segments dont l'un, le noyau, est une assertion (souvent, mais pas nécessairement, la reformulation définitoire jouant le rôle d'identification), et dont l'autre, le satellite, présente des exemples illustrant cette assertion :

(13) [Q] Qu'est-ce qu'un tri? [R1] Un tri permet de classer des éléments suivant un certain ordre (croissant ou décroissant). (R2] Il peut se faire sur des éléments numériques, alpha-numériques ou alphabétiques. (R3] On peut trier des éléments suivant plusieurs méthodes: quicksort, dichotomie, insertion, heapsort, sélection, méthode bulle.



(14) [Q] Qu'est-ce qu'un tri?. [R1] Un tri permet d'ordonner des données suivant une relation d'ordre préalablement établie. [R2] D'un point de vue algorithmique nous avons plusieurs principes de tri qui sont plus ou moins performants, par exemple le tri "bulle" qui est très lent, ou le heapsort qui lui est très performant, c'est à dire très rapide.

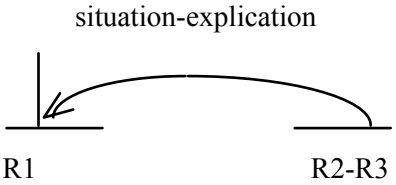


La composante la plus caractéristique de cette relation est l'énumération, qui peut être introduite par *nous avons*, *il existe*, *par exemple.* Une remarque mérite d'être faite à ce stade au sujet de la nature des marques de relations. Avec les énumérations, qui sont repérables formellement (cf. chap. 7), les traits qui marquent les bornes d'un segment de texte servent également de marques de relation. En effet, dans les définitions, il semble bien que les énumérations jouent systématiquement le rôle d'illustration. En l'absence de marques lexicales classiques de la relation d'illustration, tel *par exemple*, l'identification d'un segment illustratif grâce aux marques formelles de l'énumération peut ainsi permettre l'identification de la relation[33].

La relation de **situation-explication** est une autre forme d'élaboration : elle relie une assertion-noyau, souvent la reformulation définitoire jouant le rôle d'identification, et un segment-satellite (une ou plusieurs propositions) qui développe cette assertion de manière à la situer ou à l'expliquer. Les exemples (11) et (12) ci-dessus illustrent cette relation dans le cas d'une identification finale ; on la retrouve dans l'exemple (15) avec une identification initiale :

[33] Je reviendrai à plusieurs reprises (5.2, 6.1, 7.1) sur cette intrication entre structurer et segmenter, et le double rôle des marqueurs qui en découle.

(15) [Q] Qu'est-ce que l'extinction? [R1] C'est le désapprentissage. [R2] Si l'on ne présente plus que le stimulus neutre il n'y aura plus de réponse. [R3] Pour Pavlov après le conditionnement on ne présente plus que le son le chien ne salivera plus.



Le tableau 4.3 rassemble les marqueurs qui, dans le corpus, permettent de repérer les identifications et les relations entre les propositions. Certains de ces marqueurs sont sans doute insuffisamment discriminants et auraient besoin d'être envisagés dans le cadre de configurations et non de façon isolée. Cette question sera reprise dans la section sur l'implémentation (4.3.2).

identification	explicitation-illustration	situation-explication
• Initiale *:* attaque 3 ou 4 • Finale : *{., ;} c'est* +dét+terme	(par) exemple N1,N2, ..., Nn (etc.) (...) il existe / nous avons plusieurs/(de) nombreux SAdj1 ou SAdj2 ou SAdjn	si P1 , (alors) P2 lorsque P1, alors P2 quand P1, alors P2 P1, alors P2 P1, P2 temps futur P1, on observe/constate P2

Tableau 4.3 :Marques d'identification et de relations entre propositions[34].

4.3 Remarques conclusives

4.3.1 Analyse des définitions

Pour clore cette analyse, je voudrais revenir sur trois aspects importants de l'approche de l'organisation des textes poursuivie dans le cadre du projet MIEL : la notion de schéma canonique de définition, la position initiale ou finale de l'identification, et la variation liée au domaine.

• Schémas canoniques

La notion de schémas textuels canoniques est incontournable en génération automatique de textes. Elle peut sembler moins légitime, et même potentiellement pernicieuse en analyse, lorsqu'il s'agit d'identifier toutes les définitions, en tenant compte de toutes les variations dans leur formulation. Etant donné les objectifs de mon étude, je laisserai de côté dans le modèle de K. McKeown (1985) les règles concernant les relations obligatoires et facultatives ; je reprendrai en revanche l'idée que l'identification est une composante nécessaire d'une définition. L'identification est en effet le noyau auquel se rattachent les différentes élaborations. Un texte fait d'illustrations ou d'explications sans identification est donc un texte sans noyau, qui ne me semble pas pouvoir être reconnu comme une définition. Le contraste entre (16) et (17), deux définitions de la récupération spontanée, illustre et justifie ce jugement :

[34] Le tableau 4.3 fait appel aux abréviations suivantes : dét = déterminant; N = nom; SAdj = syntagme adjectival; P = proposition ; les parenthèses signalent des traits optionnels dans des configurations, les slashs des alternatives.

(16) Après une extinction et un temps de repos, si on représente le stimulus conditionnel, on observe le comportement initialement induit par le conditionnement : c'est le phénomène de la récupération spontanée.

(17) Après extinction et une période de repos, on présente de nouveau le stimulus conditionnel à l'animal, on constate de nouveau la réaction conditionnelle.

Ces deux réponses sont très proches dans leur formulation : elles comportent toutes deux une explication dans un format hypothético-déductif classique. La différence est cependant importante : la première comporte une identification qui est absente de la seconde. Cette différence est bien sûr très pertinente dans le cadre de l'étude de la variation dans la formulation des définitions pour la modélisation de l'utilisateur d'un tuteur intelligent. Elle a sa pertinence également pour le repérage de contextes définitoires pour l'extraction d'information, qui va devenir un objectif central dans la suite des travaux sur la définition.

• Identification initiale ou finale

On a vu que l'identification peut ouvrir la définition ou la clore. En ce qui concerne la modélisation de l'utilisateur, ce paramètre n'a jusqu'à présent pas reçu d'interprétation et n'a donc pas été exploité. J'y reviendrai dans le chapitre suivant, où j'examinerai non plus des définitions isolées, constituant à elles seules des micro-textes, mais des définitions intégrées dans des textes. Je m'intéresserai en particulier, à la lumière des travaux présentés dans la première partie, aux différents rôles de la définition dans le texte selon que le terme à définir est l'élément thématique ou rhématique (5.2.3).

• Variation liée au domaine

On se souviendra que le corpus avait été constitué de façon à permettre d'évaluer l'impact du domaine sur la formulation des définitions. Vu son caractère marginal par rapport aux objectifs du travail, cette étude n'a pas été approfondie et n'autorise que des remarques indicatives, qui vont dans le sens d'une variation liée au domaine :

– en ce qui concerne l'attaque des définitions, le corpus PSYCHO privilégie les attaques de type 1 (histoire), qui sont peut représentées dans le corpus MANA, et absentes des corpus INFO1 et 2. Ce dernier est caractérisé par la fréquence des attaques de type 2 (action), dont il n'y a qu'une occurrence dans le corpus PSYCHO. Les types 3 et 4, en revanche, associés aux définitions "canoniques", sont distribués de façon régulière dans les sous-corpus.

– en ce qui concerne la structure, on constate une préférence pour la relation de situation-explication dans le corpus PSYCHO, en contraste avec l'explicitation-illustration typique des corpus INFO1 et 2.

4.3.2 Mise en œuvre du repérage automatique

Plutôt que de présenter le détail de l'implémentation – effective, mais problématique et très partielle – à laquelle les analyses ci-dessus ont donné lieu (cf. Péry-Woodley, 1993b, ch. 3), je voudrais proposer quelques éléments de réflexion concernant la mise en œuvre de tels repérages de marqueurs. Si j'ai décidé d'intégrer ce compte-rendu de ma contribution au projet MIEL dans le présent mémoire, malgré son caractère peu abouti, c'est en effet parce qu'il tâtonne vers l'approche de la notion de marqueur qui va se préciser dans le chapitre suivant, toujours dans le cadre de la définition.

Tout d'abord, l'approche développée ici part de l'hypothèse qu'il existe à la surface des textes, tout au moins dans certains genres discursifs, des marques formelles d'opérations discursives qui peuvent donc, une fois adéquatement modélisées, faire l'objet d'un repérage automatique. Cette manière d'envisager le détail des réalisations textuelles se prête à des traitements de surface ("shallow processing"), qui ont l'avantage d'être peu gourmands en connaissances linguistiques. Le projet MIEL s'intègre dans une architecture complexe

combinant analyses syntaxique et sémantique. Les configurations de marqueurs qui m'intéressent pour le repérage des définitions ne nécessitent toutefois ni une analyse syntaxique complète ni une analyse sémantique. De nombreuses informations me semblent pouvoir être dérivées des agencements de formes et de catégories grammaticales. Pour les repérer et les extraire, il suffit de pouvoir étiqueter le texte et de disposer d'un langage d'expressions régulières permettant de composer des filtres à base de formes et de catégories.

La deuxième partie de ces réflexions concerne la notion même de marqueur textuel. Le projet MIEL était parti d'intuitions déclenchées par des récurrences lexicales. Ainsi postulait-on un style illustratif à partir d'expressions telles que *par exemple*. J'ai cherché à resituer ces intuitions dans un modèle de la structuration des textes, et à identifier pour certaines fonctions ou relations textuelles les marqueurs qui leur sont régulièrement associés. Cette démarche qui prend comme point de départ la fonction ou la relation pour découvrir les marqueurs qui lui sont associés conduit à une ouverture de la notion de marqueur : on se rend compte qu'aux marqueurs lexicaux classiques peuvent s'associer ou se substituer d'autres marqueurs, lexico-syntaxiques, typographiques, ponctuationnels. On s'aperçoit aussi que dans de nombreux cas il ne s'agit pas d'une marque, mais d'une configuration de marques dont il faudrait peut-être faire une analyse plus fine pour en pondérer les différents éléments. Le repérage d'une énumération, par exemple, peut impliquer la disposition et la ponctuation, la détermination des SN énumérés, ainsi que, dans ce corpus, des marqueurs lexicaux tels *par exemple*. Je reprendrai ces questions dans le prochain chapitre à propos des définitions, et d'une façon plus générale dans le chapitre 7.

Chapitre 5

Articuler les niveaux d'organisation textuelle : le cas de la définition[35]

Les travaux présentés dans la Partie 1 de ce mémoire s'attachaient à préciser et à rendre opérationnelles deux optiques sur l'organisation textuelle : l'étude de la structure thématique et celle de la structure rhétorique. Focalisés sur l'identification de marqueurs de ces deux structures, ces travaux ne se donnaient pas pour but de les intégrer ou même de les articuler entre elles. Le chapitre 4, consacré à la modélisation des définitions pour un tuteur intelligent, reprenait ces deux optiques, avec une tentative d'intégration limitée au cadre de l'application : l'identification, nécessaire pour qu'il y ait définition, peut se repérer d'emblée par l'examen de la structuration thématique de l'attaque de la définition, ou bien faire l'objet d'une signalisation spécifique dans la seconde partie du texte définitoire.

Le problème de l'articulation des différents niveaux d'organisation textuelle est à la fois plus général et plus complexe, et va constituer un axe majeur de ce chapitre 5, l'autre axe restant la signalisation de cette organisation. C'est encore autour de la définition que vont se constituer les différents modes d'investigation de ces problèmes, dont le compte-rendu s'organise en trois étapes. La première définit les fondements théoriques et méthodologiques de l'étude et présente le corpus d'analyse (5.1). Les deux suivantes en résument les résultats, l'analyse interne de la définition d'abord (5.2), puis l'analyse de son fonctionnement dans le texte où elle s'inscrit (5.3).

[35] Pour sa majeure partie, l'étude de la définition qui fait l'objet de ce chapitre a été menée dans le cadre du projet *Cognition, Discours procédural, Action*, financé par le GIS Sciences Cognitives (1996-98), et coordonné par J. Virbel (Institut de recherche en Informatique de Toulouse), J-M. Cellier (Laboratoire Travail et Cognition) et J-L. Nespoulous (Laboratoire Jacques Lordat). Ma collaboratrice principale a été Elsa Pascual, de l'IRIT, jusqu'à sa disparition prématurée et tragique en août 1997.

La difficulté de l'étude de l'organisation textuelle vient du fait que les principes d'organisation, qui correspondent à différentes fonctions des textes, sont multiples. Il est donc nécessaire d'envisager plusieurs niveaux d'organisation textuelle, et leurs modes d'articulation. Le courant de recherche le plus directement concerné par la complexité de l'organisation textuelle est celui qui prend pour objet les relations de discours et leur marqueurs. On y rencontre diverses tentatives d'organisation de la taxinomie des relations et des marqueurs en méta-classes fonctionnelles. Une distinction récurrente, bien que problématique, est souvent tentée entre relations et marqueurs *sémantiques* d'une part, et *pragmatiques* d'autre part (van Dijk, 1981; Redeker, 1990). W. Mann & S. Thompson distinguaient ainsi dans les premières versions de leur Rhetorical Structure Theory des relations dites "*subject-matter*" d'autres dites "*presentational*" (Mann & Thompson, 1986; 1988). Au delà de ces distinctions binaires, d'autres auteurs ont, pour rendre compte de l'hétérogénéité des relations et des marqueurs, de l'enchevêtrement des structures sous-jacentes, fait appel à la tripartition hallidayenne des métafonctions linguistiques – *idéationnelle*, *interpersonnelle* et *textuelle* – (Maier & Hovy, 1993; Bateman & Rondhuis, 1997). Je vais reprendre cette tripartition pour tenter de mieux cerner, à la lumière du modèle de l'Architecture Textuelle, les ressources linguistiques mises en œuvre par la composante textuelle.

5.1 Le niveau textuel

5.1.1 Position du problème

Pour rendre perceptible l'existence de ces différents niveaux de fonctionnement et d'organisation des textes, je prendrai le cas des définitions dans les manuels de logiciels qui vont constituer le corpus d'analyse de ce chapitre. La plupart du temps, les définitions sont dans ces textes une façon privilégiée de formuler des consignes. Plutôt qu'une formulation classique, du style : *Pour faire X, utiliser la commande Y*, on trouve régulièrement : *Y est une commande qui permet de faire X*. La formulation du "dire comment faire" passe donc par la définition. On peut envisager ce fonctionnement comme une superposition de niveaux : consigne sur un niveau, définition sur un second, sans parler, sur un troisième, de l'expression de relations entre objets du monde du logiciel.

La tripartition des métafonctions linguistiques dans le modèle de M. Halliday (Halliday, 1985; Halliday & Hasan, 1976; cf. 1.3 *supra*) semble pouvoir fournir un début de modélisation de ces fonctionnements superposés. Les métafonctions idéationnelle, interpersonnelle et textuelle suggèrent une interprétation de la triple fonction des définitions dans les manuels de logiciels : elles semblent fonctionner comme des consignes au niveau interpersonnel, comme des définitions au niveau textuel, et enfin, au niveau idéationnel, comme l'expression d'une relation d'hyperonymie et d'une description fonctionnelle. Je vais m'attacher tout particulièrement à préciser le niveau textuel, et à identifier les ressources dont disposent les langues pour la création de textes. Dans les ouvrages cités, M. Halliday examine ces ressources à trois niveaux :

– le niveau de la proposition : ordre des mots pour signaler le thème, proéminence phonologique pour signaler l'information nouvelle ;

– le niveau du groupe de propositions : utilisation de la syntaxe pour signaler les relations entre propositions, de la ponctuation pour marquer les frontières de phrases ;

– le niveau du texte, où il se focalise sur la description des procédés de cohésion : référence, substitution, ellipse, conjonction, cohésion lexicale.

Je me propose ici d'étendre l'examen des ressources pour créer du texte au delà de la notion de cohésion, à partir d'un modèle pragmatique, – le modèle de représentation de l'architecture textuelle –, qui s'intéresse à l'organisation explicite des textes (propriétés de mise en forme, y compris visuelle) comme reflet des intentions du scripteur.

Le choix de l'écrit, la focalisation sur la signalisation, une approche ouverte de celle-ci, centrée sur les fonctions plutôt que sur des marqueurs lexicaux prédéfinis, ces caractéristiques de mes travaux antérieurs à 1994 m'ont rendue particulièrement réceptive au modèle de représentation de l'architecture textuelle (Virbel, 1985; 1989; Pascual, 1991), que j'ai rencontré à mon arrivée à Toulouse, et dont je fais ci-dessous un exposé sélectif, en lien avec mes préoccupations. Conviée à me joindre à l'atelier *Texte et Communication* du *Pôle de Recherche en Sciences Cognitives de Toulouse*, j'ai en effet eu la chance d'entamer une proche collaboration avec Elsa Pascual, qui avait beaucoup contribué à expliciter et à développer le modèle proposé par Jacques Virbel. Les sections 5.2 et 5.3 sont le reflet de cette collaboration.

5.1.2 Le modèle de représentation de l'architecture textuelle

Ce modèle appréhende le texte à partir d'un de ses niveaux d'organisation, appelé ici son architecture, considéré comme nécessairement, bien que variablement, signalisé à la surface du texte. L'ensemble des procédés qui rendent perceptible l'architecture d'un texte est appelé sa *mise en forme matérielle*. La mise en forme matérielle recouvre les marques lexicales et syntaxiques, mais aussi les marques visuelles : typographie, disposition et ponctuation. La première question que l'on peut se poser est celle du statut de ces marques visuelles dans une étude des ressources linguistiques pour la création de texte. J. Virbel (1985) et E. Pascual (1991) montrent la relation d'équivalence fonctionnelle qui existe (si on laisse de côté pour l'instant les considérations d'adéquation au genre discursif ou au moment du texte) entre des formulations principalement visuelles et des formulations qu'on appellera principalement "discursives". Les images de texte de la figure 5.1 illustrent cet argument :

Image de texte 1.

1. _____ 1.1 _____ _____ 1.2 _____ _____ 1.3_____ _____	**1. _____** Dans cette section, je vais présenter trois axes d'approche de _____. Premièrement, _____ La deuxième approche _____ _____. Troisièmement, _____

Image de texte 2.

Définitions A: _____ B: _____ C: _____	A est_____. B peut se définir comme_____ _____ . On appelle C _____.

Figure 5.1: Formulations discursives et visuelles.

Dans l'image de texte 1, la même structuration du texte en trois parties est réalisée – dans l'intention qu'elle soit reconnue par le lecteur – à droite comme à gauche. De même dans l'image de texte 2, trois définitions sont formulées à droite comme à gauche. Les formulations de gauche se fondent principalement sur les procédés visuels : disposition, ponctuation et typographie, tandis que celles de droite sont principalement "discursives". Ces exemples, fabriqués pour être maximalement distants, suggèrent que les procédés de signalisation de l'architecture peuvent être perçus comme se plaçant sur un continuum allant du complètement discursif au complètement visuel. Il n'y a bien sûr pas de codage d'une propriété architecturale par un procédé typographique ou dispositionnel, mais un principe

général de contraste[36]. Si ces formulations sont perçues comme équivalentes, c'est parce qu'elles réalisent le même "acte textuel". Pour que de tels actes textuels soient réussis, il faut qu'ils soient reconnus, et que, dans la figure 5.1, les arguments des performatifs soient compris comme des sections du chapitre 1 (image de texte 1), ou comme trois définitions (image de texte 2).

Maintenant que l'orientation générale du modèle a été posée par ce rapide parcours illustré, j'en exposerai les fondements et les principes de façon un peu plus systématique, bien que très limitée et sélective, pour ensuite faire le lien avec mon approche de l'organisation des textes.

Acte textuel

Pour expliciter la notion d'"acte textuel", je citerai E. Pascual (1991:48) :

> L'examen des formes discursives mises en regard des phénomènes de mise en forme matérielle ainsi que celui de leurs conditions énonciatives suggèrent qu'elles possèdent une valeur performative et qu'on peut les caractériser comme des actes de discours particuliers à vocation spécifiquement textuelle. Lorsqu'une personne écrit 'Je définis A comme B' ou 'DEFINITION de A : B', elle demande, dans les deux cas, que le segment textuel B soit considéré comme la définition de A.

L'acte textuel est donc un cas particulier d'acte de discours, avec la spécificité que les performatifs impliqués sont des performatifs métalinguistiques, dont la performativité est dirigée vers le texte lui-même. Ces actes textuels peuvent être réalisés par la présence dans le texte des performatifs, c'est ce qui se passe dans les formulations pleinement discursives (ex. *J'organise le chapitre 1 en 3 parties*), ou ils peuvent être inférables à partir des traces de l'effacement du performatif, comme dans les formulations principalement visuelles (typographie, disposition).

Métalangage textuel

On retrouve ici la conception harrissienne (Harris, 1968) de la relation langue-métalangage : Harris montre qu'à toute phrase correspond une phrase du métalangage, et que les propriétés syntaxiques, morphologiques, ponctuationnelles, etc., de la première phrase sont des traces sur la langue de l'effacement du métalangage. Ainsi, à la phrase *Vas-tu venir ?* correspond la phrase *Je demande si tu vas venir.* ; le point d'interrogation, l'inversion sujet/verbe et le trait d'union étant les traces de l'effacement du performatif. Ces principes théoriques peuvent être généralisés au texte (Virbel, 1985). C'est ainsi que le modèle de l'architecture textuelle est basé sur la reconstitution d'un certain niveau de métalangage, qui permet de rendre compte des phénomènes architecturaux. Ce métalangage textuel est également dans la langue, et peut être décrit en termes de relations opérateurs-arguments. Les opérateurs sont des prédicats architecturaux (*définir, diviser en chapitres, commenter, énumérer*,…) ; leurs arguments sont des segments textuels appelés *objets textuels*. Un objet textuel est donc un segment correspondant à une formulation métalinguistique spécifique et rendu perceptible par la mise en forme matérielle.

Métalangage et mise en forme matérielle

Se pose alors le problème de la nature plus ou moins conventionnelle des procédés couverts par la notion de mise en forme matérielle. Cette question est tout à fait fondamentale à la fois pour la théorie, pour laquelle la signalisation est centrale, et pour les applications concernées soit par le repérage d'objets textuels dans des textes soit par leur

[36] On ne peut pas non plus parler de codage dans le cas des marqueurs lexico-syntaxiques, mais plutôt d'une exploitation pragmatique d'éléments qui deviennent ainsi multi-fonctionnels. Notre hypothèse est que des régularités existent, à condition de les envisager dans des configurations genre/domaine plutôt qu'en "langue générale" (cf. chap. 8).

génération automatique. Les auteurs du modèle ont une position très claire : il n'y a pas de convention absolue régissant les procédés de mise en forme matérielle[37], mais un principe général de contraste en tant que système d'identités et de différences. Quelle que soient la manière de réaliser un acte textuel donné, par exemple la segmentation en chapitres, les procédés utilisés pour le chapitre 1 devront se retrouver pour tous les autres chapitres. Pour expliquer cette possibilité qu'a la langue d'exprimer des illocutoires d'une manière non entièrement et non conventionnellement explicite, les auteurs (cf. Pascual 1991: 49) font appel à la réflexion de Strawson sur la relation entre intentions et conventions : un auteur aurait non seulement l'intention de produire un effet sur le lecteur, mais celle de produire cet effet par le fait que le lecteur reconnaisse l'intention de l'auteur (Strawson, 1971). Des intentions illocutoires peuvent donc être reconnues sans que leur expression soit systématiquement conventionnelle : le principe de contraste est dans ce cadre vu comme un marqueur d'intentionnalité, le contenu illocutoire précis dérivant d'autres éléments de l'interprétation.

5.1.3 Une méthodologie pour aborder le niveau textuel en corpus : métalangages et sous-langages.

L'approche de la définition qui va être exposée a ceci de nouveau par rapport aux travaux existants dans le cadre du modèle de l'architecture textuelle qu'elle s'attache à caractériser des réalisations de mise en forme matérielle en corpus et pour un objet textuel particulier. Cette orientation méthodologique nécessite d'être située sur le plan théorique, en particulier par rapport aux travaux de Harris, qui informent à la fois le modèle de la représentation de l'architecture textuelle, et l'approche des textes spécialisés qui s'élabore au sein de l'opération *Traitement Automatique des Langues : terminologie et organisation conceptuelle* de l'ERSS, dans laquelle ces travaux s'inscrivent.

Rappelons un aspect important de la notion harrissienne de sous-langage (Harris, 1968; Harris *et al.*, 1989). Alors qu'en langue générale il est pratiquement impossible de formuler les restrictions de sélection s'appliquant à un opérateur donné, il existe des systèmes linguistiques caractérisés par une syntaxe et un lexique restreints. Dans ces systèmes, les schémas de phrases sont des combinaisons particulières de sous-classes de mots, c'est-à-dire que les restrictions de sélection y sont intégrées dans la grammaire. Ces systèmes au fonctionnement spécifique, qui recouvrent métalangages et langages de disciplines scientifiques et techniques, reçoivent le nom de *sublanguage*[38]. Cette intuition théorique s'accompagne chez Harris et ses collaborateurs de la mise au point d'une méthode de description des sous-langages, c'est-à-dire de mise en évidence des classes de mots et des patrons syntaxiques qui les constituent. C'est surtout dans le cadre de sous-langages scientifiques, en médecine et en pharmacie, que cette méthode est développée, à partir de corpus de textes (cf. en particulier Sager *et al.*, 1987).

C'est bien dans l'optique des sous-langages qu'est entreprise l'étude du métalangage textuel. Mais celui-ci, contrairement au métalangage grammatical par exemple, n'a pas fait l'objet de formulations et de reformulations au cours des siècles. S'il existe bien un métalangage issu de la réflexion sur le formatage dans les domaines de la typographie et de la mise en page, cette réflexion, comme le note J. Virbel, a porté sur les aspects

[37] Cette restriction va dans le sens de nos intuitions de scripteurs : on connait la difficulté que présente l'adoption de conventions cohérentes pour marquer nouveaux termes, exemples ou citations dans le cadre d'un article ou d'un livre.

[38] La traduction française de ce terme par "sous-langage" pose au moins deux problèmes : on peut se demander s'il s'agit de sous-langage ou plutôt de sous-langue (cf. Dachelet, 1994:93); par ailleurs, le terme implique qu'il s'agit de sous-ensembles de la langue générale, alors qu'ils comportent souvent des traits propres (cf. Habert *et al.* 1997:149).

morphologique et exécutoire, et peu sur la syntaxe, encore moins sur la sémantique, des phénomènes concernés :

> il est possible d'exprimer des ordres tels que : souligner, éditer en italique, centrer, éditer en colonnes, etc…, mais non de caractériser les phénomènes textuels sous-jacents, tels que : "mise en évidence", "titrer", "citer", etc… (Virbel, 1985:12).

Ce n'est donc pas à partir de corpus de textes, mais à partir d'une approche lexico-grammaticale systématique que J. Virbel (Virbel, 1985; 1989) et ensuite M Landelle (Landelle, 1988) en abordent la description. Il s'agit de déterminer des opérateurs (verbes tels que *organiser, résumer, paraphraser, titrer,* …) et les arguments (noms dénotant des segments de textes : *texte, partie, paragraphe, titre*…) pouvant réaliser des schémas dont on a spécifié la syntaxe et les transformations. Le travail de M. Landelle (1988) concerne ainsi le schéma

$$N_0\ V\ N_1\ en\ N_2$$

Tom *découpe* son *texte* en huit *chapitres*

Il s'agit donc de décrire le métalangage textuel hors discours, "en langue".

L'approche adoptée ici se distingue de ces premiers travaux sur le métalangage textuel de deux façons :

– d'abord elle est davantage préoccupée par les réalisations effectives, les traces du métalangage textuel à la surface des textes, que par la reconstitution du sous-langage spécialisé relatif à l'architecture textuelle. En cela, elle se rapproche des travaux sur les sous-langages scientifiques et techniques ;

– ensuite, il s'agit d'une approche "en discours" doublement inscrite dans la ligne des travaux de Harris. Influencée par les linguistiques de corpus, et la conscience de la variation dans les productions langagières, elle s'attache à mettre en lumière la mise en forme matérielle d'un objet textuel spécifique dans une configuration genre discursif/domaine. En d'autres termes, elle cherche à mettre au jour les réalisations du métalangage textuel dans le cadre d'un sous-langage de domaine, en tenant compte également du genre discursif, autre facteur de variation[39]. On peut, en fonction de la théorie, envisager deux conséquences de ce choix méthodologique : d'une part on peut s'attendre à ce que le sous-langage textuel correspondant à un sous-langage de domaine dans une réalisation correspondant à un genre discursif particulier soit d'une description plus aisée que le sous-langage textuel "général" recouvrant tous les performatifs textuels possibles : nombre d'opérateurs restreint, nombre d'arguments restreint pour chaque opérateur, restrictions sur les combinaisons d'objets ; d'autre part, on peut s'attendre, dans ces configurations, à des régularités dans les réalisations (mise en forme matérielle) d'un acte textuel particulier.

Les travaux sur la définition présentés ci-après cherchent à explorer systématiquement en corpus régularités et variations dans la mise en forme matérielle de la définition. Avant d'en entamer la présentation, je vais préciser et justifier les choix méthodologiques qui se sont opérés au fur et à mesure.

Constitution de corpus : configurations domaine-genre

Comme je le notais dans (Péry-Woodley, 1995:226; cf. 8.2.2 *infra*), la notion de sous-langage n'épuise pas les possibilités de variations dans les textes : C. Montgomery et B. Glover (1986) font ainsi état des ressemblances entre les sous-langages des livres de cuisine et ceux d'autres manuels techniques, qui partagent, selon eux, de nombreux traits grammaticaux et typographiques du fait qu'ils impliquent tous la spécification de procédures

[39] La prise en compte du genre discursif constitue un ajout par rapport à l'approche harrissienne, qui envisage la variation uniquement en termes de domaine (cf. chapitre 8).

pour l'exécution de tâches. À l'inverse, plusieurs auteurs (Grishman & Kittredge, 1986; Adam, 1991; Biber & Finegan, 1994) s'intéressent à la question de l'hétérogénéité interne des textes : ainsi les manuels comportent des descriptions et des consignes, les textes narratifs des descriptions, les articles scientifiques des sections (méthode, résultats, discussion) qui varient de façon systématique quant à l'usage de certains traits linguistiques. Je reprends le terme courant de *genre discursif* pour regrouper les facteurs externes liés au canal, à la relation entre les participants, à la visée discursive, qui peuvent ainsi rassembler des textes appartenant à des domaines distincts, ou au contraire, dans le cas du dernier facteur, provoquer des variations à l'intérieur des textes. Je proposerai au chapitre 8 la notion de *document de travail* pour la classification situationnelle par rapport à laquelle on veut pouvoir identifier les variations linguistiques. De façon informelle, les documents seront présentés ici en fonction de leur rôle dans le monde professionnel auquel ils appartiennent.

De la variation aux invariants

Dans cette conception de la variation où interviennent de façon croisée des facteurs liés au domaine et des facteurs liés au genre, seules des configurations genre/domaine spécifiques semblent pouvoir être décrites adéquatement (cf. 8.1.1 *infra*). On peut d'ailleurs s'interroger sur la possibilité de travailler sur corpus en "langue générale". Les critiques formulées par Chomsky et les rationalistes à l'égard des approches empiriques restent dans ce cas tout à fait pertinentes. Sur quel principe constituer un corpus représentatif de la langue générale ? Pourtant, que ce soit au plan de la description linguistique ou des applications, on ne peut se contenter de descriptions *ad hoc* : il est nécessaire au contraire de pouvoir généraliser les observations résultant de l'étude de corpus spécifiques. On a besoin de savoir, par exemple, si les schémas qui vont permettre d'extraire automatiquement les définitions de manuels de logiciels peuvent ou non être utilisés pour d'autres domaines/genres, si certains aspects de ces schémas sont invariants, si certaines parties des définitions sont plus sujettes aux variations que d'autres. Mon approche, en collaboration avec J. Rebeyrolle, est de partir d'une hypothèse de variation, et de rechercher des invariants dans des analyses de corpus représentant différentes configurations genre/domaine (Péry-Woodley & Rebeyrolle, 1998; Péry-Woodley, 1998; Rebeyrolle & Péry-Woodley, 1998). Cette approche permet d'obtenir une caractérisation "générale" sans perdre le détail des variations.

Après ce long préambule théorico-méthodologique, j'en arrive à la présentation des objectifs de l'étude et du corpus (5.1.4). Les résultats occuperont ensuite deux sections. La première (5.2) se focalisera sur les marqueurs de définition. La deuxième (5.3) sera consacrée à la définition *dans le texte* : son intégration dans la structure du texte, qui introduira la confrontation de deux modèles d'analyse de cette structure, la RST et le modèle de l'architecture textuelle.

5.1.4 Les définitions dans des textes : présentation et corpus.

Mon intérêt pour la définition comme objet de recherche, enclenché par l'étude entreprise au LIMSI en 1990 (chap. 4) s'est trouvé relancé par mon arrivée en 1994 dans une équipe active en recherche sur l'acquisition de connaissances terminologiques. La définition représentant une sorte de concentré d'informations sémantiques, sa modélisation dans le but d'un repérage automatique s'intégrait bien dans les objectifs de l'équipe, d'autant plus que la recherche de marqueurs pour cette modélisation rejoignait la recherche de marqueurs de relations sémantiques. Même s'il y a convergence avec la recherche en acquisitionde connaissances, ce travail s'en distingue cependant, dans la mesure où Elsa Pascual et moi-même cherchions à modéliser la définition en tant qu'objet textuel : caractériser les traces du métalangage qui permettent de la reconnaître – marqueurs de relations sémantiques et autres marques –, dégager les régularités de sa structure interne, et ses modes d'intégration dans le texte qui l'entoure.

Les sections qui précèdent (5.1.1 à 5.1.3) ont permis de définir un cadre théorique et méthodologique pour l'étude des définitions dans des textes. Celle-ci vise donc une modélisation fine des réalisations linguistiques des définitions, c'est-à-dire des réalisations de mise en forme matérielle grâce auxquelles ces objets textuels sont rendus perceptibles en discours. Les configurations de marqueurs qui doivent permettre de construire des filtres pour le repérage des définitions dans les textes incluent des marqueurs visuels : ponctuation, typographie, disposition. Nous faisons par ailleurs l'hypothèse que ces marqueurs risquent de varier en fonction de paramètres liés au domaine et/ou au genre discursif, et nous abordons par conséquent leur étude dans des corpus homogènes par rapport à ces paramètres. Finalement, cette étude ne se contente pas d'élaborer des schémas représentant les différentes mises en forme matérielle des définitions, elle en examine la distribution au long du texte, dans le but de faire le lien – important en génération de texte – entre moment du texte et choix de formulation.

Au delà de ces objectifs descriptifs, l'étude est motivée par un objectif d'ordre théorique : il s'agit de mettre en relation différents modèles de structuration des textes, s'intéressant à différents niveaux d'organisation textuelle. L'approfondissement du niveau textuel de la tripartition hallidayenne, grâce au modèle de l'architecture, me paraît en effet indispensable à une modélisation plus adéquate de l'interaction entre ces niveaux d'organisation.

Outre l'application en terminologie/terminographie mentionnée plus haut, la modélisation des définitions constitue un enjeu dans deux grandes familles d'applications :

– la génération automatique, qui intéressait tout particulièrement Elsa Pascual ;

– l'extraction d'informations à partir de documents textuels. Au delà de l'approche fondée sur le repérage de marqueurs, qui s'apparente aux travaux en extraction de connaissances, cette étude cherche à tester la faisabilité d'un repérage d'objets textuels fonctionnels – définitions, conclusions, exemples – dans des documents textuels. Un tel repérage, associé à un balisage de type HTML ou SGML, permettrait en effet des recherches plus ciblées dans les textes, de même qu'on peut déjà, avec certains moteurs de recherche sur Internet, limiter la recherche au titre ou au résumé des entrées.

En fonction de ces objectifs, un corpus limité mais cohérent en termes de genre discursif et de domaine a été constitué, comportant trois manuel de logiciels :

1) une section du manuel d'un logiciel d'analyse de texte[40] (66 294 occurrences) : LOG1. Ce texte sert de texte de référence pour la première phase de l'étude.

2) le manuel d'un logiciel de compression de fichiers (19 357 occ.) : LOG2.

3) le manuel d'un système de gestion de bases de données (57 158 occ.) : LOG3.

Seront aussi évoqués dans la partie comparative de l'étude :

– un extrait d'un manuel de géomorphologie[41] (32 397 occ.) : GEO.

– un guide pour le développement de projets en Génie Logiciel[42] (54 403 occ.) : GELOG.

Dans un premier temps, l'étude s'est focalisée sur LOG1, à partir de l'idée que les schémas résultant de l'étude descriptive, et résumant la grammaire de la définition dans ce premier corpus, permettraient d'élaborer des filtres pour le repérage dans les autres corpus. Le manuel dont est extrait LOG1 comporte, outre des objets "périphériques" comme un avertissement, des remerciements ou un index, sept parties numérotées de 1 à 7. L'une présente le logiciel, l'autre en explique l'installation, et ainsi de suite. La sixième, LOG1, concerne l'utilisation même du logiciel.

---

[40] Daoust, F. (1996). SATO (Système d'Analyse de Textes par Ordinateur) version 4.0, Manuel de référence. Centre ATO Université du Québec à Montréal. Voir annexe 4.

[41] Derruau, M. (1988) Précis de géomorphologie, Masson, 7ème Edition.

[42] Il s'agit d'un guide de Génie Logiciel mis au point et utilisé par une entreprise française.

5.2 Les marqueurs de définition

En fonction du modèle de représentation de l'architecture textuelle, nous partons de l'hypothèse que les définitions dans les textes portent les traces du métalangage associé à l'acte textuel qu'elles réalisent. Ce sont ces traces qui permettent au lecteur d'interpréter à partir du texte l'intention de l'auteur, ici celle de définir, et de lire le segment de texte en question comme une définition (si tout se passe bien...). Il s'agit donc ici d'identifier les marques formelles qui signalent la présence d'une définition. Le premier stade est de constituer un corpus de définitions qui permettront d'isoler des régularités formelles, sans partir de listes de marques *a priori.* Nous avons bien sûr fait appel aux travaux antérieurs sur la définition (Martin, 1990; Riegel & Tamba, 1987; Riegel, 1990; Candel, 1994; Hathout, 1996 *inter alia*), mais surtout initialement à notre compétence de lecteur, à notre aptitude à reconnaître dans des textes dont le lectorat visé nous inclut les segments conçus pour être interprétés comme des définitions[43]. S'enclenche alors un processus itératif qui consiste en la description des microstructures textuelles qui signalent les définitions, accompagnée de retours aux textes (au moyen de filtres élaborés avec un logiciel d'analyse de textes) et d'affinement de cette première description, jusqu'à l'obtention d'une structure stable qui prend la forme d'une configuration faisant intervenir des éléments lexicaux, syntaxiques, typographiques et dispositionnels.

Cette méthode nous a amenées à identifier ce que nous avons appelé des schémas de définitions. J'en présente ci-dessous le dernier état descriptif (5.2.1), suivi d'un exercice de formalisation de la relation entre schémas syntaxiques de base et variantes syntaxiques (5.2.2). C'est la variation à l'intérieur d'un même texte qui est examinée ici, c'est-à-dire les différentes réalisations de l'objet définition, dont on verra en 5.3 qu'elles correspondent à des fonctions différentes dans la structure globale du texte. Dans la section 5.2.3, c'est un autre aspect de la variation qui est abordé, celui lié au domaine et au genre discursif, à partir d'une étude faisant appel pour la comparaison aux corpus GEO et GELOG.

5.2.1 Les schémas de définition

Ce que je vais présenter ici diffère inévitablement des résultats exposés dans (Pascual & Péry-Woodley, 1995; 1997a; 1997b). Les schémas ci-dessous ont en effet été affinés au cours des analyses subséquentes, en particulier grâce à la collaboration avec Josette Rebeyrolle.

La première approche de la description est une approche distributionnelle qui met en jeu des classes auxquelles ont été attribuées des étiquettes descriptives :

Nc : un nom classifieur (hyperonyme)

Nn : un nom propre au domaine (qui fait l'objet de la définition, autrement dit, le *definiendum*)

V= : soit la copule, soit un verbe appartenant à une classe restreinte

Vp : {*permettre, servir à, avoir pour effet, être utilisé pour, ...*}

Le tableau 5.1 présente les combinaisons attestées dans le corpus LOG1. On trouve d'abord le schémas étendu (SE), dans lequel toutes les "cases" distributionnelles sont remplies, puis les schémas réduits (SR).

	Nc1	**Nn**	**V= Nc2**	**Vp**	**SV**

[43] Resterait à tester, par le biais d'expérimentations psycholinguistiques, le degré d'accord entre différents lecteurs quant à cette identification.

SE 0 réduction	§ La commande	**Distance**	est un analyseur lexico-statistique.	Elle permet de	comparer statisti-quement les lexiques de deux sous-textes quelconques d'un corpus
SR1 réduction N*c1*		§ Le **filtre**	est un patron de fouille	qui permet de	définir ....
SR1' réduction V= N*c2*	§ L'analyseur	**COMPARAISON**		permet de	marquer ...
SR2 réduction N*c1*, V= N*c2*		§ **CARACTERISER**		permet de	préciser le fonctionn-ement du journal
SR1'' réduction V= N*c2* V*p*	§ L'analyseur	**SEGMENTATION**			découpe ...
SR2' réduc. N*c1*,V= N*c2*, V*p*		§ **APPLIQUER**			lance ...

Tableau 5.1 : Les schémas de définition.

Quelques précisions sur la représentation de la mise en forme matérielle dans ce tableau :

- au plan dispositionnel, on note que la structure se trouve toujours en début de paragraphe (symbolisé par §). cette contrainte ne semble pas pouvoir être enfreinte sans porter préjudice à l'expression d'une relation générique et donc définitoire. Les schémas présentés ici peuvent être considérés comme les bornes initiales d'une microstructure textuelle de type définitoire.

- Nn, le terme à définir, est toujours marqué typographiquement que ce soit par le gras, par des lettres majuscules, des guillemets ou de l'italique ;

- V= appartient à une classe qui peut être définie en extension (dans notre corpus : {*être, désigner*}), et est employé régulièrement au présent ;

- le groupe Nc1 Nn, ou Nn s'il n'y a pas de Nc1, est toujours précédé d'un article défini (ou sans article si Nn peut être traité comme un nom propre) ;

- Nc2 est toujours un SN indéfini.

Dans une optique fonctionnelle plutôt que descriptive, la définition est caractérisée de façon classique par l'expression du *genus* et des *differentiae*, réalisée dans ce corpus par un ou deux classifieurs[44] pour le *genus*, et un modifieur ((Vp) SV) pour les *differentiae.* Dans ce manuel de logiciel, l'expression du *genus* situe l'objet dénoté par le terme à définir dans l'univers des objets du logiciel, les *differentiae* en donnent une description fonctionnelle :

genus Nc1 Nn V= Nc2	*differentiae* Mod : (Vp) SV
La commande DISTANCE est un analyseur lexico-statistique	Elle permet de comparer statistiquement les lexiques de deux sous-textes quelconques d'un corpus

L'étiquetage des schémas réduits dans le tableau 5.1 reflète le désir de ne pas perdre de vue cette structure : les schémas SR1 (SR1, SR1' et SR1'') sont ainsi rapprochés parce

[44] Lorsqu'il y a deux classifieurs, comme dans l'exemple pour SE, le premier est présupposé, le second posé (cf. 5.2.3). Il apparaît que le premier est toujours un hyperonyme du second : un *analyseur* est une sorte de *commande*.

qu'ils forment une définition complète, avec expression du *genus* et des *differentiae*, mais où le *genus* n'est exprimé qu'une fois, par une assertion (V= Nc2) pour SR1, par un classifieur antéposé (Nc1) pour SR1' et SR1''. SR2 et SR2' en revanche sont des définitions incomplètes dans la mesure où leur manque l'expression du *genus*. Il peut même sembler paradoxal de considérer encore une telle structure comme une définition. Je reviendrai sur ce paradoxe apparent dans la section 5.3. SR1'' et SR2' sont en outre caractérisés par des réductions dans la formulation des *differentiae*.

On peut se demander, à l'examen du tableau 5.1, si des constituants présents dans les exemples attestés pourraient éventuellement disparaître, et si au contraire de nouveaux schémas possibles pourraient être créés par la présence de constituants absents. Je présente l'examen systématique de la combinatoire dans la prochaine section, suivi de l'analyse des liens syntaxiques entre les différentes réalisations des schémas.

5.2.2 L'obtention des variantes syntaxiques[45]

Le tableau 5.2 est le résultat de l'examen systématique des combinaisons des différents constituants, avec indication dans la colonne de droite des combinaisons effectivement attestées (référence de schéma), des combinaisons envisageables (A créer), et des combinaisons impossibles (~).

Nc1	**Nn**	**V=**	**Nc2**	**Vp**	**SV**	**Schéma**
+	+	+	+	+	+	SE
+	+	+	+	-	+	À créer
+	+	+	-	+	+	~
+	+	+	-	-	+	~
+	+	-	+	+	+	~
+	+	-	+	-	+	~
+	+	-	-	+	+	SR1'
+	+	-	-	-	+	SR1''
-	+	+	+	+	+	SR1
-	+	+	+	-	+	À créer
-	+	+	-	+	+	~
-	+	+	-	-	+	~
-	+	-	+	+	+	~
-	+	-	+	-	+	~
-	+	-	-	+	+	SR2
-	+	-	-	-	+	SR2'

Tableau 5.2 : Combinatoire des catégories.

Quelques explications :

a) Nn et SV étant les éléments de base des schémas ne peuvent être soumis à cet examen systématique. Sans terme à définir, sans *differentiae*, pas de définition…

b) V= et Nc2 vont obligatoirement de pair, puisque le rôle de V= est précisément d'établir la relation d'équivalence entre le *definiendum* qui le précède et le *definiens* qui le suit. C'est là la base même de la définition classificatoire : *L'objet EXTRACTION* (Nc1

[45] Cette section reproduit presque textuellement Pascual et Péry-Woodley 1997c. Bien que j'aie pris une part active à l'élaboration du travail présenté dans cet article, et que j'aie malheureusement assumé seule sa rédaction après la disparition d'Elsa Pascual, l'idée de départ avait été largement impulsée par elle, et m'a conduite dans des territoires qui m'étaient quelque peu étrangers. D'où une difficulté d'intégration dans le reste de mes travaux, et par conséquent de reformulation.

Nn) *désigne* (V=) *un fichier ASCII* (Nc2). Comme on va le voir, cet aspect classificatoire peut être absent de la définition (V= et Nc2 tous deux absents), mais il est inconcevable de rencontrer V= ou Nc2 sans son compagnon. Le tableau 5.2 fait donc état de combinatoires fonctionnellement impossibles (notées ~), qu'on ne cherchera pas à générer[46].

c) les catégories sur lesquelles on peut jouer sont par conséquent Nc1 et Vp. Le tableau 5.2 indique deux schémas non-attestés mais possibles (A créer), qui ne diffèrent de schémas attestés que par l'absence de Vp :

Nc1 Nn V= Nc2 SV
Nn V= Nc2 SV

Les textes suivants, qui bien que non-attestés semblent tout à fait acceptables, ont été produits selon ces deux schémas possibles :

*(18') Le paramètre **entier** est un filtre numérique qui sélectionne les contextes à afficher ou à exporter*

à partir de l'exemple attesté :

*(18) Le paramètre **entier** est un filtre numérique qui permet de sélectionner les contextes à afficher ou à exporter*

De même :

*(19') Le **filtre** est un patron de fouille qui définit les entrées du dictionnaire que l'on veut sauvegarder.*

à partir de :

*(19) Le **filtre** est un patron de fouille qui permet de définir les entrées du dictionnaire que l'on veut sauvegarder.*

Apparaît donc clairement ici un méta-schéma constitué par la présence de toutes les catégories, dont les variantes peuvent être décrites en termes de réduction à zéro d'une ou plusieurs catégories. Pour vérifier l'hypothèse d'un méta-schéma unique, nous en avons examiné de plus près l'obtention selon la méthode harrissienne (Harris, 1968; 1982; 1991).

Des schémas observés aux phrases élémentaires : la méthode harrissienne

Dans le modèle mathématique élaboré par Harris, la langue apparaît comme un système d'ensembles définis par des relations entre éléments. Les phrases ont ainsi la propriété d'être décomposables en phrases primaires ou élémentaires; inversement, il est possible d'engendrer de manière récursive toutes les phrases possibles à partir d'un sous-ensemble de phrases élémentaires. Ceci est faisable parce que le modèle définit les relations entre phrases de la langue et phrases élémentaires en termes d'un nombre fini de transformations et d'opérateurs. La formulation d'une théorie du langage repose pour Harris sur cette abstraction d'éléments primitifs à partir des objets directement observables. Nous montrons comment la formulation des liens formels qui unissent nos schémas de définition nous a amenées à modifier notre description initiale.

Si deux schémas de phrase sont reliés par une transformation, la différence dans les séquences de mots ou de classes de mots constituant ces schémas peut être :

– une permutation des classes de mots ou des constantes,
– l'addition ou l'omission d'une constante,
– l'addition ou l'omission d'une classe de mots.

[46] Notons que cette optique "génération" n'implique pas exclusivement une visée applicative en génération de texte, puisqu'on peut aussi souhaiter générer systematiquement toutes les variantes possibles pour le repérage automatique.

Les transformations sont obtenues par application successive d'un nombre restreint d'opérateurs :

– les opérateurs de phrase : $\varphi_{s.}$ Les opérandes de ces opérateurs sont des phrases, qui se comportent alors comme des noms, devenant un sujet ou un objet de l'opérateur.

– les opérateurs de connexion : φ_c : S_1, S_2 ∅ S_1 C S_2. Il s'agit d'opérateurs portant sur des couples de phrases, réalisés par exemple en français par les conjonctions de coordination et de subordination.

– l'opérateur de permutation sur les symboles : $\varphi_{p..}$

– l'opérateur de réduction à zéro : $\varphi_{z.}$ Les réductions à zéro interviennent après l'application de φ_S ou de φc, rendant possible dans certaines conditions l'effacement d'éléments reconstituables à partir de la phrase réduite. C'est le cas dans nos métaschémas pour les sujets de propositions infinitives ou les antécédents répétés de pronoms relatifs. On verra qu'elles s'appliquent aussi sur les métaschémas eux-mêmes pour produire des variantes.

– l'opérateur de changement morphophonématique : $\varphi_{m.}$ Il agit sur la forme phonématique d'un morphème sous l'effet d'un opérateur.

Il faut ajouter à cette liste l'opérateur φk, opérateur de bonne formation, qui produit les phrases élémentaires. L'application successive de ces opérateurs est à la base de presque toutes les transformations. Il convient de distinguer dans cette liste les opérateurs de bonne formation et de connexion, qui sont des opérateurs dits "incrémentiels" parce qu'instrumentaux dans la construction de la signification de la phrase, des trois derniers, qui sont qualifiés d'opérateurs "déformationnels", leur application ne modifiant pas la signification.

Un exemple : l'obtention d'un méta-schéma particulier

Dans un premier temps, nous partons d'un exemple attesté, qui va nous permettre d'exposer pas à pas et de façon illustrée l'obtention d'un méta-schéma. Après avoir analysé cet exemple en phrases élémentaires, nous décrivons dans le détail le processus de sa reconstruction par l'application d'opérateurs. Nous présentons ensuite le processus général, sous la forme d'un arbre tenant compte des différentes réalisations.

Notre analyse va prendre comme point de départ un schéma complet, exemplifié par la définition attestée suivante :

*(20) L'objet **message** désigne un écran de messages qu'on peut superposer sur l'écran du journal*

Cette définition est analysable en trois phrases élémentaires :

– *Message est un objet*

– *Cet objet désigne un écran de messages*

– *X superpose cet écran de messages sur l'écran du journal*

Les transformations qui s'appliquent à ces phrases élémentaires sont détaillées ci-dessous (les éléments soulignés sont ceux sur lesquels porte l'opérateur) :

1. Nn être un Nc1 φ_k

Message <u>est un objet</u>

2. Ce Nc1 V= un Nc2 φ_k

<u>Cet objet désigne un écran de messages</u>

3. Le Nc1 Nn V= un Nc2 $\varphi_c(\varphi_m(\varphi_p(\varphi_z(1))), \varphi_z(2))$

*L'objet **message** désigne un écran de messages*

4. X V= Ωi φ_k

X superpose cet écran de messages sur l'écran du journal

5b. X Vp que 4 $\varphi_s(4)$

X peut que X superpose cet écran de messages sur l'écran du journal

6bγ. X Vp 5b $\varphi_m(\varphi_z(5))$

X peut superposer cet écran de messages sur l'écran du journal

7bγ. 3 [qu] 6bγ $\varphi_c(3,6)$

*L'objet **message** désigne un écran de messages [qu] on peut superposer cet écran de messages sur l'écran du journal*

8bγ. Le Nc1 Nn V= un Nc2 qu' on Vp V-inf (SN) $\varphi_m(\varphi_z(7))$

*L'objet **message** désigne un écran de messages qu'on peut superposer sur l'écran du journal*

Commentaire :

– les phrases 1, 2 et 4 sont les phrases élémentaires de départ (φ_k) ;

– la phrase 3 est obtenue par la connexion (φ_c) des phrases 1 et 2, connexion qui s'accompagne de la réduction à zéro de V= ($\varphi_z(1)$), de la permutation de Nn et de Nc1 (φ_p), de la transformation du déterminant (φ_m), et de la réduction à zéro de *Ce Nc1* ($\varphi_z(2)$) ;

– la phrase 5b est le résultat de l'application d'un opérateur de phrase (φ_s) qui lui intègre la phrase 4 ;

– la phrase 6bγ résulte d'une réduction à zéro de *que X* ($\varphi_z(5)$) et du passage du verbe à l'infinitif (φ_m) ;

– l'application d'un opérateur de connexion sur 3 et 6 donne la phrase 7bγ ;

– 7bγ, par l'application d'une réduction à zéro ($\varphi_z(7)$) et la réalisation de l'opérateur [qu] par *qu* (φ_m), devient 8bγ, notre phrase de départ.

Les quatre réalisations du méta-schéma

Si de 1 à 4, le processus d'obtention des méta-schémas s'applique de la même façon à toutes les définitions, les indices ajoutés à la numérotation à partir de 5 indiquent en revanche qu'on a alors plusieurs options, l'exemple développé ci-dessus appartenant à la branche *bγ*. La figure 5.2 représente sous forme arborescente les divers embranchements du processus général d'obtention des méta-schémas. La première fourche, en 4, correspond à des constructions relatives différentes, objet du verbe comme ci-dessus, ou sujet comme dans

*(21) Le paramètre **entier** est un filtre numérique qui permet de sélectionner les contextes à afficher ou à exporter*

L'exemple (21) correspond à la connexion des deux phrases :

3. Le Nc1 Nn V= un Nc2

*Le paramètre **entier** est un filtre numérique*

6a Ce Nc2d Vp 5a

Ce filtre numérique permet de sélectionner les contextes à afficher ou à exporter

La réduction à zéro s'opère de la même façon que pour la branche *a*, mais l'opérateur [qu] sera dans ce cas réalisé par *qui* pour donner :

8a Le Nc1 Nn V= un Nc2 qui Vp Vi-inf (Sn)

Outre cette distinction entre relative sujet et objet, représentée par les branches *a* et *b* respectivement, trois autres différences syntaxiques entraînent une complexification de l'arbre d'obtention des méta-schémas. La branche *a* autorise une forme de connexion entre les phrases constituantes autre que la transformation relative de la branche $a\alpha$, comme le montre l'exemple ci-dessous :

(22) La commande ***Distance*** *est un analyseur lexico-statistique. Elle permet de comparer statistiquement les lexiques de deux sous-textes quelconques d'un corpus.*

Ici, on garde deux phrases et la connexion se fait par le biais de la pronominalisation. C'est la branche $a\beta$. Pour la branche *b*, on a vu dans l'exposé initial un exemple où l'antécédent de la relative était complément direct du verbe de la phrase 6. Il s'agit de la branche $b\gamma$. L'antécédent peut aussi être un complément prépositionnel, et ainsi appartenir à la branche $b\delta$ comme dans :

(23) L'objet ***EXPORTATION*** *désigne un fichier de listage standard en ASCII sur lequel pourront être exportés divers résultats affichables*

On arrive donc à l'arbre d'obtention des méta-schémas représenté dans la figure 5.2.

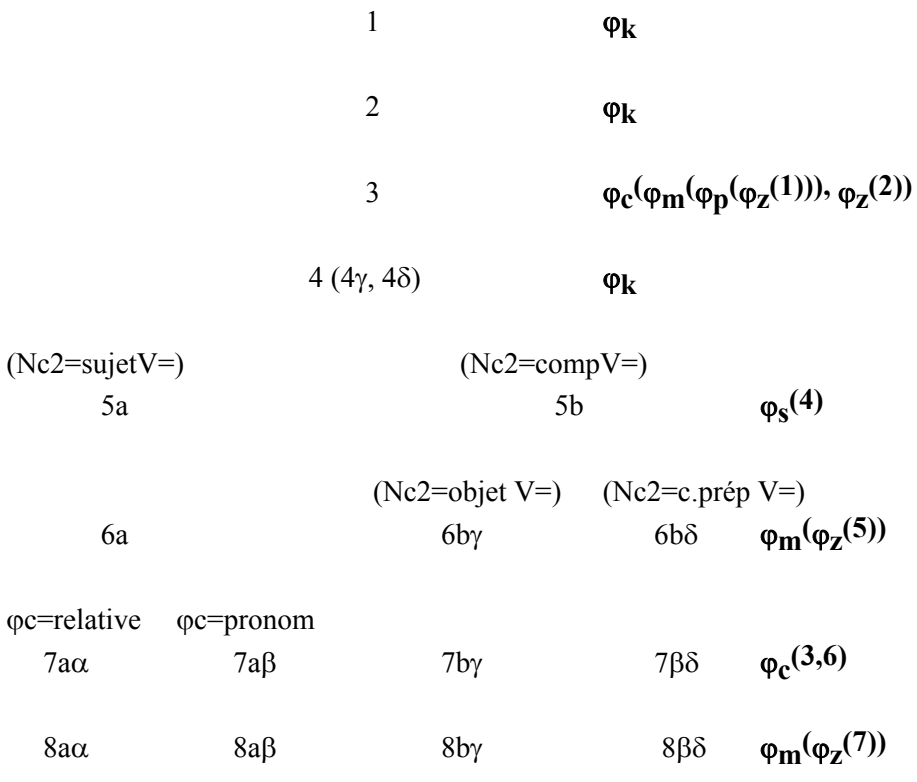


Figure 5.2 : Obtention des méta-schémas.

La figure 5.2 résume l'obtention des quatre méta-schémas. Elle montre ce qui les rassemble, – un processus commun faisant intervenir les mêmes opérateurs –, et ce qui les distingue, – des relations syntaxiques différentes au sein des phrases élémentaires de départ, et une réalisation différente de l'opérateur de connexion en ce qui concerne la branche $a\beta$.

Transformations sur les méta-schémas

Pour être complète, cette modélisation doit maintenant envisager les transformations qui peuvent s'appliquer aux quatre méta-schémas, examiner les contraintes auxquelles elles

sont soumises, et en projeter les résultats. Les attestations de transformations dans le corpus sont de deux types : réductions à zéro et permutations.

a) Réductions à zéro

On a déjà observé, lors de l'examen de la combinatoire des catégories, certaines des contraintes qui s'appliquent aux réductions à zéro : Nn et SV ne peuvent y être soumis, V= et Nc2 sont inséparables et ne peuvent s'effacer que simultanément. On peut maintenant y ajouter les observations suivantes, présentées dans l'ordre syntagmatique des éléments[47], à partir d'une variante d'un schéma du tableau 5.1 :

Nc1	Nn	V=	Nc2	Vp	SV
(24) La commande	**Distance**	est	un analyseur lexico-statistique	qui permet de	comparer statistiquement les lexiques de deux sous-textes quelconques d'un corpus

1) **Nc1** : la réduction à zéro de Nc1 semble n'être soumise à aucune condition. A partir de tout méta-schéma on doit donc pouvoir systématiquement dériver un schéma réduit. Pour la phrase (24), cela donne :

*(24') **Distance** est un analyseur lexico-statistique qui permet de comparer statistiquement les lexiques de deux sous-textes quelconques d'un corpus.*

2) **[V= Nc2]** : la réduction à zéro de [V= Nc2] a pour condition que Nc2 soit sujet de Vp, en d'autres termes elle ne s'applique qu'à la branche *a* de l'arbre d'obtention des méta-schémas. Rappelons que dans le méta-schéma, V= met en relation d'équivalence Nc2 et le couple Nc1 Nn. Une fois Nc2 réduit à zéro, c'est ce couple Nc1 Nn qui devient sujet de Vp. L'application de cette transformation à notre exemple produira :

*(24'') La commande **Distance** permet de comparer statistiquement les lexiques de deux sous-textes quelconques d'un corpus.*

3) **VP** : la réduction à zéro de Vp nécessite également que Nc2 soit sujet de Vp (branche *a*). Elle est toutefois indépendante de la réduction à zéro de [V= Nc2], comme le montrent les phrases suivantes, dérivées de deux états différents de notre exemple :

*(24''') La commande **Distance** est un analyseur lexico-statistique qui compare statistiquement les lexiques de deux sous-textes quelconques d'un corpus*

*(24'''') La commande **Distance** compare statistiquement les lexiques de deux sous-textes quelconques d'un corpus*

b) Permutation par passivation

Nous faisons état dans (Pascual & Péry-Woodley, 1997c) de deux types de permutation : passivation et permutation de Nc1. La seconde nécessiterait une étude plus approfondie, comme elle met en cause la relation entre les deux expressions possibles du *genus*, par Nc1 sans assertion, et par Nc2 avec assertion. Je ne la traiterai donc pas ici. Je reviendrai en revanche sur la question de l'assertion du *genus* en 5.2.3.

La passivation requiert que Vp ait un sujet autre que Nc2 (branche *b* de l'arbre d'obtention des méta-schémas), toujours réalisé par *on* dans les attestations, comme dans l'exemple ci-dessous :

Nc1	Nn	V=	Nc2	Vp	SV

[47] Pour la clarté de la présentation, j'illustre toutes les réductions à partir du même exemple. Les attestations en corpus sont bien sûr différentes. L'utilisation d'un seul exemple permet également de mettre en valeur la faculté génératrice de la méthode.

(25) L'objet **MESSAGE** désigne un écran de messages qu'on peut superposer sur l'écran du journal

L'application de la passivation à cette phrase, avec effacement de *on*, donne :

*(25') L'objet **MESSAGE** désigne un écran de messages qui peut être superposé sur l'écran du journal.*

Ce travail doit être envisagé comme une étape dans la modélisation de la structure interne de l'objet textuel définition et de son fonctionnement dans les textes à consignes. Il permet de mettre au jour les liens formels unissant les diverses formulations de définitions par une analyse qui est en même temps une méthode de génération systématique. Nous avons en effet montré que des métaschémas sont obtenus de façon régulière par l'application d'un petit nombre d'opérateurs. Ces métaschémas donnent ensuite lieu à des transformations qui produisent autant de variantes. Les diverses configurations sur lesquelles s'appliquent les opérateurs ont été décrites, et on a pu montrer que les types d'opérations et l'enchaînement de leur application étaient communs à toute une famille de schémas. Ceci équivaut à la mise au point d'une méthode systématique d'obtention de définitions.

5.2.3 Stabilité et variation

Les analyses résumées dans les sections précédentes visent à constituer une grammaire précise des définitions d'un manuel de logiciel, en tenant compte des aspects visuels de leurs réalisations. Elles ne prétendent aucunement à la généralité, et pourtant ne se contentent pas non plus d'être des analyses purement *ad hoc*, liées à une application bien particulière. Il s'agit du premier temps d'une étude qui doit être étendue dans deux directions : à d'autres manuels de logiciels, et à d'autres configurations genre/domaine. Dans ce qui suit, je présente quelques résultats permettant de commencer à cerner des zones de variation et de stabilité dans les définitions d'un corpus à l'autre.

a) Variations à l'intérieur de la configuration genre/domaine représentée par les manuels de logiciels

J'évoque ici, en raison du lien avec les travaux sur l'énumération qui seront présentés au chapitre 7, quelques éléments de réflexion sur les marqueurs de définitions. On se souviendra que le *genus* peut faire l'objet d'une réalisation double (schéma étendu SE), simple (schémas réduits SR1, SR1' et SR1'') ou nulle (schéma réduit SR2 et SR2'). Le tableau 5.3 ci dessous reprend le tableau 5.1 en le simplifiant pour résumer les modalités de l'expression du *genus*[48] :

	genus				***differentiae***
	N*c1*	**N*n***	**V=**	**N*c2***	**Mod**
SE 0 réduction	§ La commande	**Distance**	est	un analyseur lexico-statistique.	Elle permet de comparer statistiquement les lexiques de deux sous-textes quelconques d'un corpus
SR1 réduction N*c1*		§ Le **filtre**	est	un patron de fouille	qui permet de définir ....
SR1' réduction V= N*c2*	§ L'analyseur	**COMPARAISON**			permet de marquer ...
SR2 réduction N*c1*, V= N*c2*		§ **CARACTERISER**			permet de préciser le fonctionnement du journal.

Tableau 5.3 : Schémas de définition : réalisations du genus.

[48] Les constituants Vp SV étant ici regroupés sous la catégorie Modifieur, les schémas SR1'' et SR2' se trouvent ramenés aux schémas "parents" SR1' et SR2 respectivement.

J'ai fait état plus haut de l'apparent paradoxe qui consiste à interpréter comme une définition un énoncé auquel manque un aspect constitutif, – même définitoire!, – de la définition : l'expression du *genus*. Les réalisations du schéma SR2 peuvent-elles légitimement être acceptées comme des définitions ? Cette question m'a amenée à regarder de plus près mes données initiales. L'extension de l'analyse à d'autres manuels de logiciels (LOG2 et LOG3) permet également quelques observations.

D'abord, l'examen des énoncés de type SR2 en contexte a révélé que la réduction totale du *genus* n'avait lieu que dans certaines configurations dispositionnelles. Ce schéma réduit est en effet intimement lié à l'énumération, qui à elle seule peut exprimer une mise en relation d'hyperonymie, comme on le voit en élargissant le contexte de l'exemple de schéma SR2 (tableau 5.3) :

(26) Cinq actions s'appliquent à cet objet : AFFICHER, CARACTERISER…:

- AFFICHER permet de…

- CARACTERISER permet de préciser le fonctionnement du journal

- ...

Le *genus* est en fait bien présent dans la phrase introductive (*actions*), et la relation d'hyperonymie est exprimée sous la forme de l'énumération : *AFFICHER, CARACTERISER*… sont des *actions*.

Par ailleurs, les résultats préliminaires de l'étude d'autres manuels de logiciels suggèrent que la formulation privilégiée par le premier texte étudié est particulièrement redondante, puisque le *genus* y est généralement exprimé à la fois dans la définition et par la structuration du texte. Ainsi dans l'exemple (27), la première définition fait appel à un classifieur de type Nc1, *L'analyseur*, pour préciser le *genus* de *COMPARAISON*, alors que cette relation hyperonymique est également établie par la structuration des titres ; la troisième définition reprend même toute la taxinomie, avec une catégorisation qui pourrait être paraphrasée comme : *Distance* est-un *analyseur* est-une *commande*. Les autres manuels étudiés privilégient au contraire des formulations plus laconiques, du style de (27') et (27'')[49], où les relations hyperonymiques sont indiquées uniquement par la numérotation et/ou la typo-disposition :

(27)

6. Commandes du programme d'interrogation

6.1. Analyseurs

6.1.1 Comparaison

L'analyseur ***COMPARAISON*** *permet de marquer…*

6.1.2 Comptage

L'analyseur ***COMPTAGE*** *permet de compter…*

6.1.3 Distance

La commande ***DISTANCE*** *est un analyseur …*

(27')

6. Commandes du programme d'interrogation

6.1. Analyseurs

6.1.1 Comparaison *: permet de marquer…*

6.1.2 Comptage *: permet de compter…*

6.1.3 Distance *: analyseur lexico-statistique…*

(27'')

- ***Commandes du programme d'interrogation***

[49] L'exemple (10) est une représentation partielle et simplifiée de la structure du manuel SATO, (10') et (10'') sont fabriqués à partir de (10) pour permettre la comparaison.

– Analyseurs

- ***Comparaison****: permet de marquer…*
- ***Comptage****: permet de compter…*

- ***Distance****: analyseur lexico-statistique…*

La mise en parallèle de ces formulations fait ressortir des similarités fonctionnelles, et étend la typologie des marqueurs de la relation d'hyperonymie. Cette relation peut en effet être inférée sur la seule base de la hiérarchie des titres. Les énumérations peuvent fonctionner comme des définitions à deux niveaux : chaque élément peut être une définition (comme dans l'exemple (9)), avec un *genus* commun exprimé dans la phrase introductive ; par ailleurs, l'énumération dans son ensemble peut également constituer une définition, dont le *genus* se trouve dans l'introduction et les *differentiae* dans les éléments énumérés. C'est le cas pour la partie 6.1 (ex. (27)), dont la structure sera analysée dans la section 5.3.

b) Variation entre configurations genre/domaine

L'objectif ici est de confronter les "grammaires" de définitions élaborées à partir de textes appartenant à des configurations genre/domaine différentes. Au premier texte du corpus décrit en 5.2 (LOG1) s'ajoutent pour cette étude les sous-corpus GEO et GELOG[50] :

Outre qu'ils diffèrent sur le domaine, ces textes n'ont pas la même visée discursive et ont des relations diverses à leur public : GEO est un ouvrage didactique destiné à des étudiants novices dans le domaine de la géomorphologie ; LOG1, manuel de référence d'un logiciel spécialisé, est destiné à des lecteurs déjà familiers avec les manuels de logiciels mais novice dans l'utilisation de ce logiciel, alors que GELOG, guide pour le développement de projets en Génie Logiciel, s'adresse à des experts du domaine.

• *Distribution des quatre schémas*

Nous avons abordé la comparaison de ces trois sous-corpus en termes de présence/absence des quatre principaux schémas, tels qu'ils sont présentés dans le tableau 5.3. Le tableau 5.4 donne le rappel des schémas et la distribution de chacun dans les trois sous-corpus :

	Schémas				**Corpus**		
	Nc1	**Nn**	**V= Nc2**	**Mod**	**GEO**	**LOG1**	**GELOG**
SE	+	+	+	+		présent	présent
SR1	–	+	+	+	présent	présent	présent
SR1'	+	+	–	+		présent	présent
SR2	–	+	–	+		présent	

Tableau 5.4 : Distribution des quatre schémas de définition.

Deux des schémas ont un comportement particulier : SR1 parce qu'il est le seul à apparaître dans chacun des trois sous-corpus, et qu'il est le seul schéma réalisé dans le corpus GEO ; SR2 parce qu'il n'apparaît que dans le corpus LOG1. On a d'ailleurs vu, pour ce dernier, que le *genus* y est tout de même exprimé, bien que de façon un peu indirecte, dans l'introduction de l'énumération dans laquelle ce schéma se trouve systématiquement. En ce qui concerne le cas de SR1, il faudrait creuser la différence fonctionnelle entre les définitions où la relation hyperonymique fait l'objet d'une assertion, celles où la "case" V=Nc2 est remplie (SE et SR1), et les définitions où cette relation est réalisée par un

[50] Cette section est le fruit d'une collaboration avec Josette Rebeyrolle, qui prépare une thèse à l'ERSS sous la direction d'A. Borillo.

syntagme nominal complexe (SR1')[51]. Notons que c'est le corpus de géomorphologie (GEO) qui se démarque par l'emploi exclusif de SR1, où la relation d'hyperonymie ne se fait que par l'assertion : dans ce texte didactique, on peut en effet voir l'objectif principal comme étant précisément d'introduire les termes du domaine et de construire la taxinomie des objets.

Ces observations se trouvent renforcées par des occurrences, dans le manuel de géomorphologie uniquement, d'un schéma de définition que je n'ai pas traité ici, du type *La mer qui gèle (...) prend d'abord une contexture huileuse (...) : c'est le* slush. Ce schéma place le *focus* (dans le sens que donne à ce mot la théorie de la structure d'information) sur la dénomination nouvelle (cf. Flowerdew, 1992), en conformité avec un objectif central dans un texte didactique : l'introduction de nouveaux termes. A l'opposé, ce schéma serait inconcevable dans les deux guides du domaine logiciel, où ce sont les *differentiae* et les informations fonctionnelles qu'elles véhiculent qui sont la partie la plus développée des définitions, et constituent régulièrement le segment focalisé. Dans le manuel LOG1, on l'a vu, les consignes prennent régulièrement la forme de définitions : *X est une commande qui permet de faire Y*, ou *la commande X permet de faire Y*, se liront comme *utiliser la commande X pour faire Y*. L'expression des *differentiae* y joue donc un rôle majeur.

Ces variations dans la structuration des définitions semblent avoir trait au genre discursif des textes. On va examiner maintenant les variations dans la formulation des *differentiae*, qui, elles, paraissent liées au domaine.

• *Variation dans la formulation des* differentiae

Le modifieur qui exprime les *differentiae* peut être réalisé par un nombre restreint de constructions (relative, participe présent, syntagme adjectival ou prépositionnel, proposition indépendante). Même si des correspondances syntaxiques peuvent être établies entre ces constructions (par transformation, réduction, etc., cf. 5.2.2), il demeure intéressant d'examiner comment les réalisations sont distribuées dans les différents sous-corpus. Le tableau 5.5 illustre à travers quelques exemples les principaux types de modifieurs recensés dans chacun des sous-corpus.

	Nc1	**Nn**	**V= Nc2**	**Mod**
GEO	–	§ Les *griffures*	sont des sillons	étroits.
LOG1	–	§ Le **filtre**	est un patron de fouille	qui permet de définir les entrées du dictionnaire que l'on veut sauvegarder.
LOG1	§ La commande	**Distance**	est un analyseur lexico-statistique	Elle permet de comparer statistiquement les lexiques de deux sous-textes quelconques d'un corpus
GELOG	–	§ Le guide d'élaboration de la documentation de spécification	est un guide méthodologique	pour la production des documents de spécification des logiciels scientifiques.

Tableau 5.5 : Exemples de variations des differentiae.

[51] Le schéma SR1' rassemble pour ce nouveau corpus deux réalisations différentes du couple Nc1 Nn : "l'analyseur COMPARAISON permet..." et "le Plan de Développement Logiciel Standard répond...". Dans notre approche, elles peuvent, en effet, l'une et l'autre être envisagées comme des réductions du Schéma Etendu : "l'analyseur COMPARAISON est un analyseur qui permet...", "le Plan de Développement Logiciel Standard est un plan qui répond...".

Là encore, il s'agit de résultats préliminaires d'une étude en cours, qui ne comporte pas pour le moment de quantification systématique des occurrences des différents schémas. Il apparaît cependant nettement que le texte GEO, le manuel de géomorphologie, privilégie les modifieurs de type adjectif et participe passé. Au contraire, dans les deux textes appartenant au domaine du logiciel (LOG1, GELOG), les modifieurs sont essentiellement des subordonnées relatives, et parfois des propositions indépendantes liées par pronominalisation, ou encore des subordonnées ou syntagmes prépositionnels introduits par *pour*. Pour expliquer ces distributions spécifiques, force est de considérer les informations sémantiques liées au domaine auquel appartiennent les textes du corpus. Il convient en effet de distinguer, d'une part, les définitions descriptives du domaine de la géomorphologie, dont le but est de préciser les propriétés des objets (adjectifs), leur localisation, leur mode de formation (participes passés) ; et, d'autre part, les définitions fonctionnelles du domaine logiciel, qui sont régulièrement exprimées au moyen de subordonnées relatives dans lesquelles le prédicat est un verbe comme *permettre*, *servir à* ou des subordonnées prépositionnelles introduites par *pour*.

5.2.4 Variations, invariants, et repérage automatique

La section précédente a décrit les variations observées dans les différents sous-corpus au cours de l'élaboration d'une grammaire des définitions pour leur repérage dans des bases de données textuelles à partir d'une analyse de surface. Une question d'importance, à la fois par rapport aux applications visées et au problème de l'identification de fonctionnements discursifs généralisables, concerne l'existence d'invariants dans la structure ou la réalisation des définitions. Plusieurs possibilités se présentent pour la construction d'une grammaire de cet ordre : on peut envisager une grammaire limitée aux invariants, une grammaire englobant toutes les variations identifiées, ou plusieurs grammaires partiellement incompatibles.

Étant donné son inachèvement, l'étude des variations, si elle autorise quelques conclusions partielles, permet surtout de formuler un certain nombre de questions. Sur le plan de la structure, les observations confirment le modèle classique en *genus-differentiae*. A l'intérieur de cette structure, l'expression du *genus* varie cependant d'une façon que l'on peut se représenter comme un continuum allant de l'assertion explicite "X est un Y" (*Le **filtre** est un patron de fouille*), où Y est un hyperonyme de X, à la formulation minimale par le biais de la structuration du texte et du titrage. Entre ces extrêmes se situe la réalisation par un SN complexe de type "le Y X", par exemple *la commande **Distance***. On retrouve avec cette analyse l'hypothèse issue du modèle de représentation de l'architecture textuelle, et développée par Elsa Pascual (Pascual, 1991), selon laquelle la mise en forme matérielle des objets textuels se situe sur un continuum allant d'une formulation entièrement discursive à une formulation entièrement typo-dispositionnelle. La différence est qu'il s'agit ici de l'expression d'une relation sémantique (hyperonymie) et non de la signalisation d'un objet textuel. On peut considérer qu'on est en présence d'un point d'articulation entre les niveaux de fonctionnement du texte, l'expression de l'hyperonymie (niveau *idéationnel* chez Halliday, cf. 1.3) faisant partie intégrante de la formulation d'une définition (niveau *textuel*). Quoi qu'il en soit, la notion de marqueur risque d'être amenée à évoluer : l'équivalence fonctionnelle entre marqueurs "discursifs" et marqueurs typo-dispositionnels démontrée dans le cadre du modèle de représentation de l'architecture textuelle s'appliquerait ainsi tout autant aux marqueurs de relations sémantiques.

Ce qui avait initialement été interprété comme une variation structurelle, présence ou absence du *genus* s'est donc trouvé ramené à une variation dans sa formulation. Dans cette variation, la caractéristique assertion/non assertion semble liée au genre discursif et à l'importance de l'attribution du *genus* par rapport à la visée discursive. L'autre élément de la structure, les *differentiae*, qui sont toujours exprimées, présente un nombre limité de réalisations que l'on peut ramener à un nombre encore plus restreint de combinaisons de phrases élémentaires (cf. 5.2.2). Par ailleurs la fréquence relative de ces différentes

réalisations varie apparemment en fonction du domaine, ainsi que du type d'informations sémantiques en jeu

Le travail d'élaboration d'une grammaire des définitions est loin d'être terminé. Ce qui a été décrit dans cette section 5.2, ce sont les marqueurs de l'énoncé correspondant à la borne initiale de l'objet textuel définition. On verra dans ce qui suit comment cet objet s'inscrit dans le texte qui l'inclut. Parmi les marqueurs identifiés, ceux de l'expression du *genus* constituent l'ensemble le plus réduit et le plus stable, apparemment indépendant du domaine, sinon du genre discursif. Pour pouvoir utiliser ces marqueurs dans la constitution de filtres, il faudrait :

– d'abord déterminer la forme du texte sur laquelle les filtres vont pouvoir être utilisés, sachant qu'un objectif d'une grammaire de surface telle que celle-ci est de limiter les exigences en pré-traitement et en informations auxiliaires nécessaires à l'application des filtres. L'état actuel de la grammaire implique un étiquetage morpho-syntaxique, une analyse syntaxique limitée (groupes nominaux et verbaux) et un balisage de type HTML pour les marques typo-dispositionnelles ;

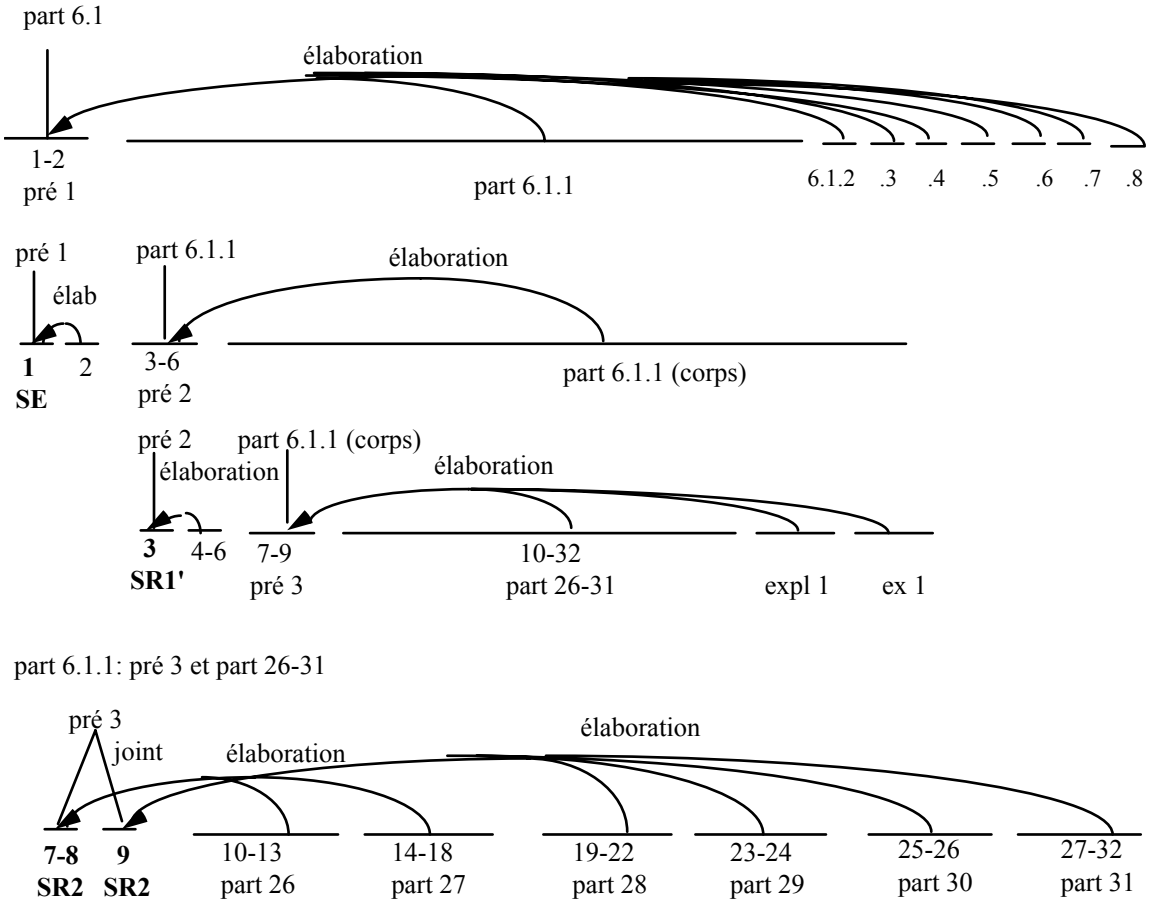
– mieux comprendre les interactions entre les différents marqueurs de manière à définir précisément des configurations minimales, la plupart de ces marqueurs étant en effet insuffisants pris individuellement. A l'intérieur de ces configurations, il serait peut-être utile d'affecter à certains marqueurs une pondération en fonction de leur fiabilité . Ces questions seront développées dans le chapitre 7.

5.3 La définition dans le texte

Interpréter les définitions de ces manuels de logiciels comme des consignes, c'était déjà une façon de les envisager en contexte, mais en termes de leur fonction *interpersonnelle* (cf. 1.3). Ici, en collaboration avec Elsa Pascual, c'est leur rôle sur le plan de l'organisation textuelle qui va être examiné, à travers les deux modèles qui ont été jusqu'à présent traités séparément et qu'on va maintenant tenter de rapprocher : le modèle de représentation de l'architecture textuelle et la Rhetorical Structure Theory (Pascual & Péry-Woodley, 1995; 1997a; Péry-Woodley, 1998). Nous avons cherché à montrer comment les définitions participent à l'architecture et à la structure rhétorique du texte, comment elles sont elles-mêmes architecturées et construites rhétoriquement. On verra que ces deux pans de la question sont intimement liés puisque le texte "de référence" (LOG1), le seul sur lequel cet aspect de l'analyse ait été mené, est constitué par une série de définitions enchâssées.

5.3.1 Les définitions dans la structure du texte

Je rappelle que le texte soumis à l'analyse est le chapitre 6, intitulé *Commandes du programme d'interrogation* du manuel du logiciel d'analyse de texte SATO. A l'intérieur de ce chapitre, nous nous sommes focalisées sur la partie 6.1, intitulée *Analyseurs*, les analyseurs constituant un type de commande. La figure 5.3 donne une représentation hybride de cette partie, hybride dans le sens où elle intègre deux analyses, l'une, descendante, en termes d'objets textuels, l'autre, montante, en termes de structure rhétorique. La première a été exécutée par E. Pascual sur la base de la mise en forme matérielle qui rend les objets textuels perceptibles. J'ai effectué la seconde selon la méthode interprétative de la RST. Les schémas RST résultant des relations rhétoriques sont donc étiquetés à la fois selon ce modèle en termes de propositions (numérotées) et de relations rhétoriques, et selon le modèle de l'architecture en termes d'objets textuels. Le mot *partie* est utilisée dans ce modèle comme un terme générique recouvrant chapitre, section, sous-section, etc. Lorsque ces parties coïncident avec des parties numérotées dans le texte, cette numérotation est préservée (partie 6.1, 6.1.1, etc.). Les parties non-numérotées dans le texte ont fait l'objet d'une numérotation dans le cadre de l'analyse en objets textuels (parties 26 à 31).



part 6.1.1: pré 3 et part 26-31

Figure 5.3 : représentation RST/architecture[52]

• Des définitions enchâssées

Au tout début du texte, le présentatif 1 définit le *terme « analyseur »* comme une *classe d'objets*, qui sont énumérés dans la proposition suivante selon un schéma d'élaboration. Ce présentatif pris dans son ensemble fait lui-même l'objet d'une élaboration de même type, sous la forme des huit parties (6.1.1 à 6.1.8) qui reprennent tour à tour les différents objets énumérés pour les développer. Ce qui est frappant dans cette représentation, c'est que ce schéma *noyau + satellite d'élaboration* se reproduit ainsi à plusieurs niveaux d'analyse : on le retrouve lorsqu'on pénètre dans la partie 6.1.1, et ensuite dans la partie 25, et dans les parties 26 à 31 qui la constituent. En d'autres termes, on peut considérer que la partie 6.1, schéma d'élaboration, est une définition, constituée de définitions elles-mêmes constituées de définitions.

• Objets textuels vs. segments rhétoriques

Dans cette analyse, les objets textuels qui sont identifiés sur la base de la mise en forme matérielle (principalement disposition et ponctuation) correspondent tous à des

[52] Les abréviations de la figure 3 correspondent aux objets textuels suivants :
part : partie
pré : présentatif
expl : explication
ex : exemple.

segments RST. Ce qui découle de cette correspondance, c'est que les marques de mise en forme matérielle peuvent aussi signaler des segments RST. Or les auteurs du modèle de la RST, tout en indiquant que l'analyse peut être abordée de manière descendante aussi bien qu'ascendante, ne donnent aucune précision quant aux marques permettant l'identification de segments de haut niveau. En fait, toute analyse descendante ne peut que se fonder sur une prise en compte – intuitive et non explicite – de la mise en forme matérielle pour identifier ces segments. C'est précisément cet aspect de la production de sens dans l'interprétation des textes que le modèle de représentation de l'architecture textuelle cherche à rendre explicite.

Étant donné que structure rhétorique et architecture modélisent des niveaux distincts de la structuration des textes, on peut toutefois s'interroger sur la généralisabilité de la correspondance, totale dans le texte LOG1, entre segments rhétoriques et objets textuels[53]. On pourrait cependant supposer que c'est là que se rejoignent, ou devraient idéalement se rejoindre pour la lisibilité du texte, les composantes idéationnelle, interpersonnelle et textuelle. C'est peut-être bien ce qui est exprimé par les théories naïves du paragraphe qui l'envisagent comme une partie d'un texte portant sur un thème et réalisant une intention. Certainement, dans notre analyse, les segments RST, qui n'existent dans ce modèle que comme le lieu d'application d'une relation, acquièrent grâce à cette correspondance un statut sur le plan textuel : statut organisationnel – parties, énumérations, présentatifs, …. – , ou statut fonctionnel comme c'est le cas pour les définitions. Il existe par ailleurs des liens forts entre certains objets textuels et certaines relations rhétoriques : ainsi les schémas de définitions présentés en 5.2 sont systématiquement les noyaux de schémas rhétoriques d'élaboration. Mais toute élaboration n'est pas une définition : il s'agit, on l'a vu, de différents niveaux de fonctionnement. La relation rhétorique d'élaboration a bien trait à la composante idéationnelle : on élabore une proposition ou un ensemble de propositions (qui constitue(nt) le noyau de la relation) ; c'est en revanche sur le plan de la composante textuelle que l'objets textuel définition s'inscrit dans le texte et participe à son architecture.

• Signalisation des segments vs. des relations

Les deux modèles du texte qui fondent cette étude s'opposent à première vue assez radicalement en ce qui concerne la signalisation : le modèle de représentation de l'architecture des textes accorde une place centrale à la mise en forme matérielle, trace du métalangage textuel, qui rend perceptibles les objets textuels ; la RST, au contraire, rejette tout lien systématique entre les relations rhétoriques et une quelconque signalisation [54]. La position des auteurs de la RST nous semble légitime dans sa prudence, son souci de bien spécifier la nature sémantico-pragmatique des relations, et d'éviter des associations simplistes entre marqueurs et relations. Par ce dernier point elle n'est d'ailleurs pas si loin du modèle de l'architecture, qui définit le principe de fonctionnement de la mise en forme matérielle en termes de contraste, et non pas de codage régulier. La double originalité du modèle de l'architecture est de poser théoriquement la présence d'une signalisation, et d'ouvrir la notion de signal de manière à inclure les marques visuelles, qui font pleinement partie de la panoplie de moyens linguistiques pour créer du texte. Nous faisons par ailleurs l'hypothèse que des régularités peuvent sans doute être identifiées à condition de découper la "population" – les textes – en termes de critères pertinents. Dans cette étude, les textes sont regroupés en fonction du domaine et du genre discursif. S'il y a isomorphie dans ce corpus entre objets textuels et segments RST, il serait donc possible d'utiliser des marques de mise en forme matérielle pour identifier des segments ou des relations rhétoriques (à l'intérieur d'une même configuration genre/domaine). On a évoqué les liens qui semblent

[53] Des travaux plus récents sur l'énumération (Luc *et al.* 1999) mettent d'ailleurs en cause cette correspondance.

[54] "The applicability of a relation definition never depends directly on the form of the text being analyzed; the definitions do not cite conjunctions, tense, or particular words. RST structures are, therefore, structures of function rather than structures of forms."(Mann & Thompson, 1987:19).

unir la définition et la relation d'élaboration, ou l'énumération et cette même relation. En fait on voit que c'est ici un objet textuel, tel l'énumération, repérable par ses propriétés de mise en forme matérielle, qui signale la relation dont il est le satellite. Inversement, les schémas de définitions résument les propriétés de mise en forme matérielle des noyaux de relations d'élaboration qui constituent des définitions.

C'est donc ici la borne initiale du segment définition qui a été modélisée. Cette borne signale par la même occasion la relation d'élaboration dont elle est le noyau. On se rend compte que la signalisation n'est pas du même ordre fonctionnel pour la structure rhétorique et pour l'architecture : marqueurs de relations pour l'une, de segments pour l'autre. Mais la distinction n'est peut-être pas pertinente : déjà dans un précédent travail sur la définition (Daniel *et al.*, 1992; chap. 4), je montrais que la relation d'élaboration pouvait être réalisée par l'expression *par exemple* suivie d'une énumération, ou bien simplement par une énumération. Le marqueur lexical *par exemple* représente l'explicitation "discursive" de la relation ; un pendant visuel pourrait être les deux points. En l'absence de telles marques, l'énumération suffira, dans certaines conditions, à ce que le segment concerné soit interprété comme une relation d'élaboration. On voit que signalisation de segment et signalisation de relation peuvent être envisagées comme intimement liés, ce qui accroît la pertinence de la mise en relation des modèles.

5.3.2 Distribution des schémas de définition dans le texte

La représentation RST/architecture de la figure 5.3 indique l'apparition de différents schémas de définition dans la structure. Ainsi, le noyau du présentatif (relation d'élaboration) de la partie 6.1 est un SE (Schéma Étendu), où le *genus* est doublement exprimé. Au niveau inférieur, le noyau du présentatif de la partie 6.1.1 est un schéma réduit de type SR1' (une seule expression du *genus*). Si l'on descend encore dans la hiérarchie du texte, on voit que le présentatif du corps de la partie 6.1.1 comprend deux schémas réduits de type SR2 (pas de réalisation du *genus* dans la définition). Existerait-il une corrélation entre la formulation des définitions et la structure du texte? Le résultat de l'analyse de la distribution des schémas de définition dans l'ensemble de la partie 6 du texte LOG1 (135 pages) est donné dans la figure 5.4 : le type de schéma est inscrit en regard de la partie numérotée dans laquelle il apparaît.

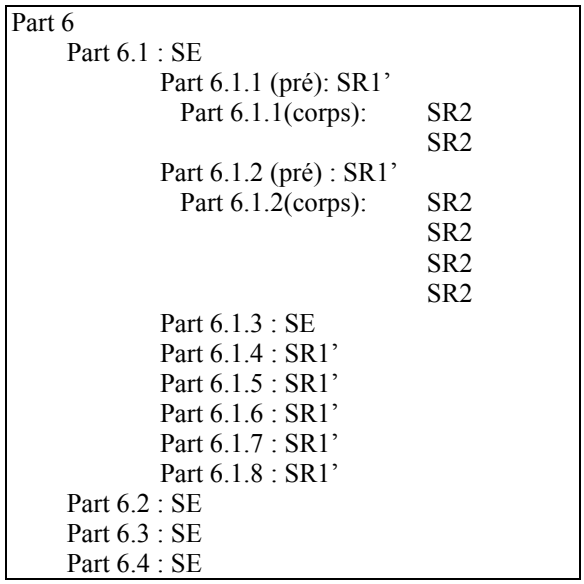


Figure 5.4. Distribution des schémas de définition.

La corrélation suggérée par la figure 5.3 se confirme donc pour l'ensemble du chapitre 6. Les définitions, ou plus précisément les noyaux de définitions, qui ouvrent chaque partie (6.1 à 6.4) sont toutes de type SE. Les sous-parties (6.1.1 à 6.1.8) commencent pour la plupart (7 sur 8) avec des définitions de type SR1'. Les définitions qui apparaissent dans les composantes de ces sous-parties sont toutes, quant à elles, des schémas réduits de type SR2.

Les figures 5.3 et 5.4 mettent en relation l'analyse fine de la mise en forme matérielle des définitions avec la structuration globale d'un segment étendu. Les régularités qui apparaissent dans la distribution des schémas de définition sont révélatrices de l'aspect dynamique de la construction du texte. Les définitions sont envisagées ici comme des objets textuels qui correspondent à des relations d'élaboration dont les noyaux sont caractérisés par une mise en forme matérielle spécifique. On a vu comment les traits de cette mise en forme réalisent de diverses façons la structure *genus* + *differentiae*. Ce que montre la figure 5.4, c'est que l'expression du *genus* gagne ou perd de l'importance selon le moment du texte : elle est systématique – mise en place de la taxinomie des objets constitutifs de l'univers du logiciel – lorsqu'on introduit un nouvel objet ; en revanche lorsque l'on progresse dans la description de cet objet, on rencontre des définitions où l'objet de l'assertion est non plus le *genus* mais les *differentiae*, c'est-à-dire ici la description fonctionnelle, qui guide l'action (puisque ces définitions réalisent en fait des consignes).

Les travaux résumés dans ce chapitre ont débouché sur des résultats descriptifs et sur une réflexion d'ordre théorique, à laquelle j'ai tenté de donner forme, mais qui n'est pas achevée. Au plan de la description, ces travaux ont permis de caractériser des schémas de réalisation des définitions dans des manuels de logiciels, schémas qui sont le premier stade d'une représentation pouvant permettre un repérage automatique. Une comparaison a été amorcée avec des définitions apparaissant dans d'autres configurations genre/domaine, qui a permis d'identifier des zones de variation et des zones de stabilité dans la mise en forme matérielle des définitions dans ces différents corpus. Enfin, l'examen des régularités dans la distribution des différents schémas de mise en forme matérielle en fonction du moment du texte a montré que les définitions faisant office de consignes privilégient nettement les aspects classificatoires en début de grande partie pour s'orienter de plus en plus vers l'expression des caractéristiques fonctionnelles au fur et à mesure que l'on s'enfonce dans la hiérarchie du texte. Ces résultats constituent des avancées descriptives susceptibles d'être utilisées en extraction d'information (cf. Jacquemin & Bush, 2000) et en génération de texte (cf. Luc, 1998a et b).

Au plan théorique, la confrontation de deux modèles de l'organisation textuelle, envisagée dans le cadre des métafonctions du langage dans la théorie de Halliday, a permis de faire un pas vers un modèle mieux à même de rendre compte de la complexité des fonctionnements textuels. En me focalisant sur la définition, j'ai pu commencer à articuler les différents niveaux de fonctionnement correspondant aux trois métafonctions hallidayennes. Cette réflexion se poursuit dans le chapitre 6 avec une étude qui va faire le lien avec la Partie I en réintroduisant la notion de *thème* (ou *topique*). La confrontation des modèles a également permis de faire évoluer la notion de marqueur : j'ai proposé d'élargir cette notion pour inclure les marqueurs visuels, dont je défends la nature proprement linguistique, conception qui sera développée dans le chapitre 7. J'ai cherché ensuite à affiner leur étude en corpus pour tenir compte de variations liées au genre discursif et au domaine. Cette approche variationniste des marqueurs, ainsi que les notions de genre discursif et de domaine, sont l'objet du chapitre 8.

Partie III

Trois fils d'Ariane : niveaux d'organisation textuelle, marqueurs, corpus

Dans mon parcours, trois thèmes jouent le rôle de fils d'Ariane. Je vais les reprendre ici un à un – articulation des niveaux d'organisation textuelle, définition de la notion de

marqueur, variation dans les corpus – tantôt pour présenter des travaux récents ou proposer un bilan, tantôt pour ouvrir des perspectives de travaux futurs.

Chapitre 6

Niveaux d'organisation textuelle

Structure d'information et structure thématique, structure rhétorique, architecture textuelle, ce sont différents niveaux de l'organisation des textes qui ont été envisagés et qui ont motivé et informé les analyses de corpus présentées dans les deux premières parties de ce mémoire. La question de l'articulation de ces niveaux d'organisation, articulation nécessaire à l'élaboration d'un modèle de la cohérence, parcourt l'ensemble de ces travaux, et se trouve spécifiquement soulevée dans le chapitre 5, avec référence au modèle de M. Halliday. Autour de la définition, j'y mets en relation la structure thématique et la structure rhétorique, la structure rhétorique et l'architecture. Je reprends ici ce thème de l'articulation des niveaux d'organisation textuelle, pour y ajouter deux éclairages importants, issus d'abord de l'étude de M. Charolles sur l'encadrement du discours (Charolles, 1997), ensuite de la théorie du centrage (Walker, Joshi & Prince, 1998) et du modèle du discours dans lequel elle s'intègre (Grosz & Sidner, 1986).

Dans le premier, M. Charolles traite de la structuration du discours par le regroupement de propositions dans un *cadre*, et du rôle de certaines expressions dans la signalisation de cette structuration. Quant à la théorie du centrage, qui fournit une méthode précise d'analyse de la cohérence à l'intérieur d'un segment de discours, elle constitue le volet local d'un modèle qui envisage aussi la cohérence intersegmentale. Ces deux éclairages me paraissent d'autant plus pertinents par rapport à mes propres travaux qu'ils s'intéressent de très près aux réalisations linguistiques associées aux modes de structuration qu'ils décrivent. J'en ferai une brève présentation dans ce qui suit en résumant un travail récent (Péry-Woodley, à paraître) dans lequel, à partir de l'examen d'expressions introductrices de cadres, je m'efforce de mettre en relation l'analyse du focus attentionnel selon le centrage avec la structure d'information et des aspects de l'architecture. Ce résumé constitue la première partie de ce chapitre, dont la deuxième partie présente un projet éditorial qui amorce un dialogue entre chercheurs et modèles sur l'articulation des niveaux d'organisation textuelle.

6.1 Encadrement du discours, structuration thématique et centrage.

J'ai exploré, dans le chapitre 2 (2.1), le fonctionnement de certains circonstants selon qu'ils apparaissent en début ou en fin de phrase. Après S. Thompson (1985), j'ai observé que les circonstants (propositions circonstancielles ou adverbiales) placés à l'initiale ont souvent une portée qui dépasse l'énoncé dans lequel ils apparaissent et qui peut s'étendre sur plusieurs énoncés. S. Thompson parle de *thèmes marqués*. Ces expressions, qui ont donc un

rôle dans la structuration du texte par le fait qu'elles délimitent un segment par l'étendue de leur portée, sont en effet de par leur position détachée à gauche à rapprocher de la notion de thème ou de topique. Pourtant elles coexistent la plupart du temps avec une autre expression, le sujet grammatical, qui est lui-même très souvent associé à la relation topique de l'énoncé. Certaines des expressions introductrices d'univers de discours décrites par M. Charolles (*op.cit.*) s'apparentent elles aussi, sur le plan formel, à des topiques disloqués à gauche. La coexistence de deux entités potentiellement topicales peut s'expliquer si elles fonctionnent sur des plans différents de structuration. La théorie du centrage fournit des outils conceptuels et méthodologiques qui permettent d'explorer cette hypothèse. Le modèle du focus global auquel elle se rattache ouvre par ailleurs une autre perspective sur les expressions introductrices d'univers de discours : en tant que signaux de l'ouverture d'un cadre, sont-elles à rapprocher des "cue-phrases" qui marquent l'ouverture d'un nouvel espace de focus et donc le début d'un nouveau segment de discours ?

6.1.1 L'encadrement du discours

• Les expressions introductrices de cadres

On l'a vu dans le premier chapitre, l'approche de M. Charolles se réclame de la pragmatique à travers la notion d'un principe général de cohérence sous-tendant l'interprétation des discours (Charolles, 1983; 1995). Dans son étude récente de l'encadrement du discours (Charolles, 1997), il s'attache toutefois à distinguer ce qui, dans l'examen de l'interprétation des discours, a trait à des calculs mettant en jeu l'intelligence générale des participants de ce qui relève d'une approche linguistique. Cette étude a précisément pour objet les marques de nature linguistique, marques de cohésion, qui codent les instructions relationnelles qui vont guider l'interprétation. M. Charolles y distingue, à un niveau très général, deux grandes classes de marques de cohésion :

– des expressions qui signalent qu'une certaine relation doit être établie entre deux unités adjacentes ou proches (anaphores et connecteurs) ;

– des expressions qui marquent que plusieurs unités doivent être traitées de la même manière relativement à un critère spécifié par ces expressions.

C'est à cette seconde classe qu'appartiennent les *expressions introductrices de cadres de discours* : leur fonction est de signaler que "plusieurs propositions apparaissant dans le fil d'un texte entretiennent un même rapport avec un certain critère, et sont, de ce fait, regroupables à l'intérieur d'unités que nous appellerons **cadres**". (1997:4). Parmi ces cadres, M. Charolles distingue les univers de discours, dont j'étudie ici un sous-ensemble. Il reprend pour les définir la formulation logico-sémantique de R. Martin (1983:37) : "l'ensemble des circonstances, souvent spécifiées sous forme d'adverbes de phrase, dans lesquelles la proposition peut être dite vraie". Voici deux exemples du type d'expressions qui va être examiné :

(1) ***Pour Maurice Volkowitsch****, les axes et les carrefours structurent l'espace.*

(2) ***Selon le rapport de la commission Vérité et Réconciliation, chargée notamment d'enquêter en 1991 sur les crimes commis pendant la dictature****, Marcos Quezada n'a pas survécu a la répétition de chocs électriques infligés par les policiers.*

M. Charolles examine dans le détail, à partir d'exemples, le fonctionnement des cadres dans la construction et l'interprétation des textes ; il pose le problème de l'interaction entre ce mode d'organisation et les relations interpropositionnelles, et s'interroge sur son rôle dans l'interprétation des connecteurs et des expressions anaphoriques. Parmi ses observations, je retiendrai particulièrement celle concernant la *portée* des introducteurs d'univers de discours (dorénavant IU), qui est à mettre en relation avec mes observations inspirées par l'étude de S. Thompson sur les circonstancielles de but évoquée plus haut (Cf. 2.2.1). M. Charolles note qu'"à la différence des modalités dont la portée ne peut facilement s'étendre à d'autres propositions que celles auxquelles elles sont adjointes, les expressions

introductrices d'univers du discours ont une propension à intégrer plus d'une proposition et elles jouent, de ce fait, un rôle bien particulier dans l'interprétation" (*op.cit.*:23). Ce rôle procédural et cognitif est décrit comme étant de servir d'une part "à régler les opérations de mobilisation de connaissances requises pour l'interprétation pas à pas des relations entre propositions", d'autre part "à répartir les contenus propositionnels dans des blocs homogènes relativement à un critère spécifié par le contenu de l'introducteur" (*op.cit.*:24).

Sur le plan formel, M. Charolles décrit les expressions introductrices de cadres comme des "groupes syntaxiques périphériques adjoints à la phrase", apparaissant "très souvent en tête de phrases". Elles appartiennent donc aux constructions détachées, comme les dislocations, dont j'ai examiné certains fonctionnements sur le plan de la structure d'information (2.2.2). C'est cette ressemblance formelle entre certaines expressions introductrices de cadres et les dislocations qui m'a incitée à examiner la relation entre l'encadrement du discours tel qu'il est envisagé par M. Charolles et la structure d'information. Cela m'a amenée dans un deuxième temps à faire appel aux outils méthodologiques de la théorie du centrage pour préciser cette relation.

• Caractérisation syntaxique

Ainsi que le signale B. Combettes (1998), les constructions détachées posent un problème particulier à la syntaxe, dans la mesure où elles échappent précisément aux relations de dépendance et de rection qui constituent son "fonds de commerce". Elles font donc souvent l'objet de définitions négatives peu satisfaisantes. Pour caractériser les expressions qui me concernent, je ferai appel aux critères pris en compte par B. Fradin dans son inventaire des constructions à détachement (Fradin, 1990), et par K. Lambrecht dans son étude typologique de la dislocation (Lambrecht, à paraître[55]). Mon utilisation de ces critères diffère toutefois fondamentalement de la leur dans la mesure où je ne cherche pas à définir une construction mais simplement à caractériser formellement, pour ensuite l'examiner en termes de structure d'information et de centrage, un sous-ensemble parmi les IU définis sur le plan discursif par M. Charolles : ceux qui comportent une expression référentielle.

Le premier critère a trait à la pause ou au dénivelé intonatoire, transcrit indifféremment par une virgule à l'écrit, qui séparerait l'élément détaché du reste de l'énoncé. Au cours de son article, B. Fradin ramène ce critère au rang de simple paramètre à prendre en compte dans la description. Il semble toutefois pertinent pour ce qui me concerne. K. Lambrecht pose pour la dislocation que le constituant détaché se situe hors des bornes de la proposition contenant le prédicat, soit à sa gauche, soit à sa droite. Les IU se conforment à la première de ces conditions, mais sont systématiquement détachés à gauche.

Un autre critère (le deuxième de B. Fradin) distingue deux types de constructions détachées en fonction de la nature argumentale du syntagme détaché, et donc l'existence ou non d'un lien anaphorique avec la proposition qui le suit. Contrairement à B. Fradin qui s'intéresse principalement aux syntagmes détachés réalisant un argument de la prédication, je ne retiendrai pour cette étude que ceux qui ne font pas l'objet d'une reprise dans la proposition. Il va de soi en conséquence que le troisième critère de B. Fradin sera satisfait : le syntagme détaché est "en surplus" des éléments dont l'agencement suffit à constituer un énoncé complet. Ces expressions s'apparentent donc aux cas particuliers de détachement et de dislocation décrits respectivement par B. Fradin comme "constructions détachées sans rappel" et par K. Lambrecht comme "topiques non-liés". C'est précisément la possibilité qu'une entité extra-argumentale puisse être potentiellement topicale, et entrer "en concurrence" avec le sujet grammatical, que cette étude se propose d'examiner. La question du statut topical des IU sera posée dans la prochaine section.

[55] Les références de pages données dans le cours du mémoire pour cette publication à paraître sont celles du manuscrit.

K. Lambrecht (1998; à paraître) insiste sur la distinction à faire entre la notion de topique, qui évoque une fonction pragmatique associée à une position syntaxique, et celle d'adjoint, qui évoque une relation grammaticale ou sémantique entre un référent et une prédication. En ce qui concerne les IU, la position syntaxique – détachée à gauche – est constitutive. Certes, les expressions lexicales qui réalisent des IU peuvent également être utilisées comme des adjoints, et dans ce cas être placées en incise dans la prédication (ex. *Il faisait déjà nuit, **selon la concierge**, lorsqu'ils sont arrivés*). Il paraît clair cependant que si elles apparaissent ailleurs que dans un syntagme détaché à gauche, ces expressions perdent leur spécificité fonctionnelle d'introducteurs de cadre, en particulier leur aptitude à étendre leur portée au-delà de la proposition immédiate. Par ailleurs, contrairement aux adjoints, les IU à l'étude autorisent le lien anaphorique avec un argument à l'intérieur de la proposition. On vient de le voir, les cas que l'on va considérer sont caractérisés par l'absence d'un tel lien anaphorique, mais ce lien n'est pas impossible avec les expressions concernées, comme en témoigne l'exemple attesté ci-dessous :

*(3) … Guy Georges tue. La première fois en 1991. **Selon lui**, **il** a "flashé" en croisant sa victime.*

En (3), le clitique sujet *il* doit être interprété comme référant à la même entité que le pronom accentué *lui* dans l'IU, c'est-à-dire *Guy Georges*. Bien que les cas comme (3) soient exclus de cette étude, ils servent à montrer que les IU fonctionnent plus comme des dislocations, autorisant le lien anaphorique avec un argument, que comme des adjoints, qui ne l'autorisent pas.

Pour résumer, mon étude des IU va se focaliser sur les expressions constituées par des syntagmes :

– détachés, c'est-à-dire situés hors des bornes de la proposition, à sa gauche, séparés d'elle par une virgule ;

– non-argumentaux (contrairement à (3)), c'est-à-dire ne faisant pas l'objet d'une reprise pronominale dans la proposition, qui est syntaxiquement complète, et par rapport à laquelle le syntagme détaché est en surplus.

Sur le plan syntaxique, il s'agit donc d'expressions caractérisées d'une part par leur position, d'autre part par l'absence de relation grammaticale avec le reste de l'énoncé. La relation qui unit l'IU et la prédication qui suit, et éventuellement un certain nombre d'énoncés dans sa "portée", va maintenant être examinée successivement en termes de structure d'information et de centrage.

6.1.2 Introducteurs d'univers de discours et structure d'information

Parmi les expressions introductrices d'univers de discours, M. Charolles signale de nombreuses locutions adverbiales figées (*en général*, *à vrai dire*, *d'habitude*, etc.). Ici, c'est sur les syntagmes non-figés et porteurs d'une expression référentielle, comme les exemples (1) et (2) ci-dessus, que je vais me focaliser. Ce sont en effet ces expressions référentielles qui confèrent aux IU un potentiel topical, ce sont elles qui peuvent faire l'objet d'une reprise anaphorique dans l'énoncé suivant (ex. *Pour Maurice Volkowitsch, les axes et les carrefours structurent l'espace. **Il** considère…*).

Mon exploration du fonctionnement des IU prend donc comme point de départ un rapprochement formel avec les dislocations, dont l'interprétation fonctionnelle fait l'objet d'un large consensus dans la littérature : la fonction discursive des dislocations serait en effet de marquer un constituant comme topique par rapport à une prédication qui en constitue le commentaire. Le topique, on l'a vu dans le chapitre 2, est généralement défini en termes d'"aboutness", ou "à propos" (cf. Berthoud, 1996), et de "givenness". La notion de topique implique une relation d'"aboutness" entre une entité et une prédication dans un contexte de discours. Par ailleurs, pour qu'une entité puisse être interprétée comme topique d'un énoncé dans ce sens, deux conditions pragmatiques doivent être réunies : l'entité

dénotée par l'expression "topicale" doit être identifiable par l'interlocuteur, et elle doit être saillante dans le discours. La dislocation à droite signale la continuation d'une relation déjà établie, contrairement à la dislocation à gauche, qui signale l'annonce ou l'établissement d'une nouvelle relation de type topique entre un référent et une prédication. L'exemple (4), emprunté à K. Lambrecht (cf. ex. (13), 2.2.2), illustre cette distinction :

(4) A et B à table ; A regarde son assiette :

A : Ça n'a pas de goût, ***ce poulet****.*

B : ***Le veau****, c'est pire.*

Leurs caractéristiques formelles – détachement à gauche, présence d'une expression référentielle, absence de lien anaphorique avec la prédication – m'amènent donc à envisager le sous-ensemble d'IU étudiés ici comme un cas particulier de dislocation à gauche, les topiques non-liés, caractérisés précisément par le fait qu'ils ne présentent aucun lien anaphorique avec un argument (explicite ou implicite) ou un adjoint à l'intérieur de la proposition, et qu'ils ne possèdent pas d'équivalents disloqués à droite (Lambrecht, à paraître:11-12). Je formule donc sur cette base l'hypothèse de travail selon laquelle les entités réalisées par ce que j'appelle IU auraient le statut de topique. Cette hypothèse se trouve également confortée par la possibilité de rapprocher certains introducteurs de cadres d'un autre type d'expression topicale : il s'agit de ce que K. Lambrecht appelle "scene-setting topic" (Lambrecht, 1994:118), S. Dik "thème" (topique non-argumental à gauche de la proposition, cf. Dik, 1997, vol.2), et W. Chafe topique "à la chinoise" ("Chinese-style topic", cf. Chafe, 1976). La définition qu'en donne W. Chafe est d'ailleurs remarquablement proche de la définition des expressions introductrices de cadres donnée en 6.1 : un élément qui établit "a spatial, temporal or individual framework within which the main predication holds" (Chafe, *op.cit.*).

Je vais maintenant examiner le rôle qu'ont dans le centrage ces expressions référentielles non argumentales et supposées topicales.

6.1.3 Introducteurs d'univers de discours et centrage

La théorie du centrage est décrite par B. Grosz *et al* (1995:204) comme "a theory that relates focus of attention, choice of referring expression, and perceived coherence of utterances, within a discourse segment"[56]. Cette théorie adopte, similairement à M. Charolles (1997), une perspective axée sur le traitement cognitif des discours, mais d'une façon plus formalisée et algorithmique liée au fait que, dans son cas, la modélisation du traitement ("processing") vise des applications informatiques. Aussi bien qu'un modèle explicatif et prédictif, la théorie du centrage se présente comme une méthode qui donne la possibilité de "formally test claims associated with topichood and coherence by using centering as a tool to track the flow of salience (or focus of attention) within discourse" (Hurewitz, 1998:277). Un des objectifs principaux de l'étude qui fait l'objet de cette section est d'examiner la contribution des IU à la construction du modèle de discours.

Le point de contact entre le centrage et ce qui précède est la notion de topique, bien que cette notion n'appartienne pas en tant que telle à la panoplie conceptuelle du modèle[57].

---

[56] Je traduis par *énoncé* le terme *utterance* par lequel les auteurs du modèle dénotent l'unité de base. A l'écrit, il pourrait s'agir soit d'une unité syntaxique (cf. ma définition d'unité syntaxique, section 2.1 et de T-unit, section 2.3.1) soit de la phrase ponctuationnelle, qui n'y correspond pas toujours (voir Kameyama, 1998 pour une étude du Centrage dans les phrases complexes). Dans le travail fondateur de Grosz (1977), les segments de discours étaient définis en termes de tâches dans des dialogues finalisés. Grosz et Sidner (1986) les définissent comme des agrégats d'énoncés correspondant à une intention qui s'inscrit dans une structure de buts discursifs.

[57] Je ne fais ici qu'esquisser les aspects du modèle nécessaires à la compréhension de mon étude des IU. Pour un exposé complet, se reporter à Walker, Joshi & Prince (1998), et à Cornish (2000).

Selon la théorie du centrage, pour tout énoncé E_i d'un segment de discours D, les entités évoquées par E_i, appelées *centres anticipateurs*, constituent un ensemble, Ca(E_i, D). À l'intérieur de cet ensemble, la théorie identifie une entité particulière qui est le *centre rétroactif,* Cr(E_i, D), de l'énoncé E_i. Cette entité jouit d'un statut spécial qui lui est conféré par deux propriétés : c'est l'entité la plus centralement concernée par E_i, et elle fait le lien entre E_i et ce qui précède. On retrouve là les deux caractéristiques principales associées au topique, et déclinées avec quelques différences selon les théories : "à propos" ("*aboutness*") et "identifiabilité" ("*givenness*"). La notion de centre rétroactif, comme son nom l'indique, traite du lien entre E_i et E_{i-1}. Par ailleurs, le modèle du centrage aborde le lien avec l'énoncé suivant en faisant appel à un classement des Ca en termes de saillance, classement fondé principalement pour l'anglais sur le rôle grammatical des expressions référentielles. Ce classement identifie le *centre préféré,* Cp, classé en tête, et permet ainsi d'anticiper le Cr de l'énoncé suivant. Le modèle prévoit quatre types de transition d'un énoncé au suivant, dont la plus naturelle, ou cohérente, est la transition par *continuation*, dans laquelle l'entité réalisée par le Cr de E_{i-1} l'est également par le Cr de E_i, et où elle est le Cp de E_i . Je reproduis ci-dessous le tableau récapitulatif de ces transitions, suivi d'exemples utilisés dans la présentation initiale du modèle (Grosz *et al.*, 1995), dans la traduction française de F. Cornish (à paraître) :

	$Cr(E_i) = Cr(E_{i-1})$ 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 6.1 : Transitions dans le modèle du centrage.

– Continuation du Cr antérieur :

a Susan a offert un hamster à Betsy.

b ***Elle*** *lui a rappelé que les hamsters étaient assez sauvages.*

– Rétention du Cr antérieur :

a Susan a offert un hamster à Betsy.

b ***Elle*** *lui a rappelé que les hamsters étaient assez sauvages.*

c ***Betsy*** *lui a dit qu'****elle*** *aimait beaucoup ce cadeau.*

– Déplacement "en douceur" :

a Susan a offert un hamster à Betsy.

b ***Elle*** *lui a rappelé que les hamsters étaient assez sauvages.*

c ***Betsy*** *lui a dit qu'****elle*** *aimait beaucoup ce cadeau.*

a ***Elle*** *a dit que c'était tout à fait ce qu'elle voulait.*

– Déplacement brutal :

a Jean est un type sympa.

b Il a rencontré Marie hier.

c ***Lucie*** *était avec elle.*

La préférence pour la transition par *continuation* est fonction d'une règle énoncée comme suit :

> Règle 2 : En ce qui concerne les relations entre énoncés, des séquences continuant le Cr de l'énoncé précédent (*continuation*) sont préférées par rapport à des séquences qui ne font que le retenir (*rétention*), et des séquences qui retiennent le Cr précédent sont préférées par rapport à des séquences qui le déplacent (*déplacement*). La transition de

type *déplacement en douceur* est préférée par rapport à la transition de type *déplacement brutal*. (traduction de Cornish, à paraître).

Ce bref exposé de la théorie du centrage me permet de formuler à présent plus précisément une des questions qui motivent cette étude. Elle concerne fondamentalement les rôles respectifs, sur le plan de la construction du modèle de discours, de l'entité réalisée par l'expression détachée et de celle réalisée par le sujet de la prédication. Je la décompose en plusieurs sous-questions :

– en ce qui concerne le lien avec les énoncés qui précèdent celui où apparaît l'IU (regard vers l'arrière), quel est le statut de l'entité réalisée par l'IU ? Peut-elle être interprétée comme le Cr de l'énoncé ?

– en ce qui concerne l'anticipation de ce qui suit (regard vers l'avant), comment cette entité doit-elle être classée dans l'échelle de classement des Centres anticipateurs ("*Cf ranking*") ? Les échelles proposées pour diverses langues prennent-elles en compte les constructions détachées ?

– le modèle peut-il fournir une manière de représenter et d'exploiter, dans le calcul du focus attentionnel, l'observation selon laquelle les IU peuvent étendre leur portée au-delà de la proposition qui suit ?

– si cette observation revient à assigner aux topiques détachés une fonction différente de celle des topiques argumentaux, comment ce double fonctionnement peut-il s'intégrer dans la théorie du centrage ?

6.1.4 Examen en corpus des expressions introductrices d'univers de discours

L'approche que je choisis pour me constituer des observables pertinents pour aborder ces questions est bien sûr l'analyse de corpus. Ici il s'agira d'un corpus d'extraits de textes où apparaît un sous-ensemble d'expressions référentielles réalisant des IU conformes à la caractérisation formelle établie en 6.1.1, celles introduites par *pour* et *selon*.

Le choix du corpus de départ, dont sera extrait le corpus d'expressions en contexte, n'est pas sans importance. Une étude de l'impact du domaine et du genre discursif sur les modalités d'encadrement du discours et les types d'expressions introductrices sera évoquée dans la section 7.3.1. Le corpus utilisé ici est celui qui a été construit pour cette étude, il permettrait donc un examen des variations liées à ces paramètres. Pour ce travail toutefois, restreint comme il l'est dans cette phase exploratoire à un étroit sous-ensemble d'expressions, l'aspect variationniste sera laissé de côté. Le corpus d'expressions est constitué de 36 textes, dont certains contiennent plus d'un IU. Ceux-ci sont au nombre de 59.

La première étape de cette recherche d'observables pertinents sera d'examiner la relation entre l'énoncé comportant un IU (E_i) et les énoncés qui le précèdent, de manière à préciser le statut des deux entités réalisées respectivement par l'IU et le sujet grammatical de E_i. Ensuite, l'examen se portera sur le Cr de l'énoncé suivant celui où apparaît l'IU (E_{i+1}). Y a-t-il reprise de l'entité réalisée par l'IU, de celle réalisée par le sujet grammatical, ou d'une autre entité ? Par quels types d'expression ces reprises se font-elles ? Les différentes transitions entre énoncés ont-elles des degrés différents d'acceptabilité en termes de cohérence discursive, ainsi que le prévoit la théorie du centrage (Règle 2) ? Dans les cas d'IU à portée étendue, il faudra enfin examiner l'interaction entre leur fonction dans le centrage (niveau local de la composante attentionnelle) et dans la définition du "focus global" du discours.

• Regard vers l'arrière : statut de l'entité réalisée par l'IU en relation avec les énoncés précédents

a) Propriétés du SN de l'IU

Dans la section 2.3, les IU ont été rapprochés d'un type de construction disloquée, les topiques non-liés. Ce rapprochement suggère que, comme les dislocations, ils auraient pour effet de marquer un constituant comme dénotant le topique à propos duquel la prédication formulerait un commentaire. Le topique est ainsi défini en termes d'"à propos", on l'a vu, mais aussi en termes d'identifiabilité cognitive. Ceci conduit K. Lambrecht (à paraître) à stipuler que les constituants disloqués doivent être des expressions définies, à moins qu'elles ne puissent susciter une interprétation générique. On a vu par ailleurs dans la section 6.1.3 que le Cr de la théorie du centrage est très proche du topique : l'entité la plus centralement concernée par l'énoncé, et celle qui fait le lien avec ce qui précède. Les contraintes quant aux types d'expression capables de tenir ce rôle sont exprimées de manière scalaire ("hiérarchie du donné", cf. Gundel *et al.,*1993; Gundel, 1998; à paraître), et vont dans le même sens que celles formulées par K. Lambrecht :

pronom inaccentué > démonstratif > article défini > article indéfini.

L'examen des données révèle quelques cas qui ne paraissent pas se conformer à ces principes. Les IU en *pour* et *selon* que j'ai relevées tendent à avoir une fonction bien particulière qui est d'introduire, en nommant leur origine, des propos rapportés ou des prises de position[58]. Dans la plupart des cas, le syntagme nominal dénote un individu (ex. (1)) ou un groupe, quelquefois un texte (ex. (2)). Dans plusieurs cas, comme dans (1), il s'agit d'un nom propre dont c'est la première apparition dans le texte. Le fonctionnement particulier des noms propres peut cependant expliquer ces occurrences. Plus surprenant, on trouve dans un même texte plusieurs IU dont le SN est indéfini :

> *(5) Le secteur communautaire pare aux manques, aux effets "pervers" du libéralisme (...). Certains mouvements ou lieux communautaires jouent le rôle d'organisations de services.* ***Pour un acteur universitaire,*** *"le socio-communautaire est plus lié aujourd'hui à l'idée de service. Ceci d'autant plus que l'on assiste dans le même temps à un désengagement de l'État. Cela s'adresse à certaines clientèles spécifiques, notamment le milieu du travail social." Il y voit une certaine source d'inquiétude.* ***Pour une géographe, spécialisée dans le développement local,*** *" il n'y a plus d'intermédiaires ...*

Il y a donc lieu de s'interroger sur la présence d'un SN indéfini, ne se prêtant pas à une interprétation générique, pour dénoter une entité qui, si elle peut constituer le topique ou le Cr de l'énoncé, devrait être identifiable. Ces cas peuvent s'apparenter à l'analyse que donne M. Charolles de l'exemple suivant :

> *(6) À Paris, Luc a été acclamé par la foule.*

M. Charolles écrit qu'en entendant cet énoncé à la radio, aux informations, avec certaines propriétés intonatives, il comprendrait que le journaliste ouvre ainsi une "rubrique" Paris, dont il s'attendrait qu'elle soit suivie soit d'autres faits ayant trait à cette ville, soit de rubriques parallèles *à Lyon* ou *à Rome* (1997:30). En (5), on a bien cet effet de liste des différents interlocuteurs qui ont participé à la série d'interviews rapportées par le texte : "*Pour un chercheur, ... Pour un acteur de ce milieu, ... Pour un acteur universitaire, ... Pour une géographe, ...*". Dans le cas des noms propres, il s'agit aussi d'une série d'auteurs dont les propos sont rapportés. Cet effet peut s'apparenter à celui de "frame inference" invoqué par K. Lambrecht pour expliquer qu'un interlocuteur puisse soudain prédiquer sur "le veau", en l'introduisant par une dislocation à gauche, alors qu'il était jusque-là question du poulet dans les assiettes (exemple (4)).

[58] C'est le cas pour 56 de mes 59 exemples.

Plusieurs questions ressortent de ces exemples qui contredisent le modèle sans toutefois choquer l'intuition : on peut se demander s'ils mettent en cause le critère d'identifiabilité de l'entité dénotée par le topique, manifesté par la nécessité pour le SN d'être défini (ou pronominal). De façon plus radicale, dans ces exemples, il semble impossible d'envisager les blocs de discours direct ou rapporté comme étant "à propos" de leur énonciateur... C'est donc plutôt le statut de topique des IU (et de Cr potentiel pour l'entité qu'ils réalisent) qui est en question. Deuxièmement, dans la mesure où ils font attendre une série, il est envisageable qu'ils aient un rôle particulier dans la segmentation du discours et dans la relation entre focus local et focus global.

b) Introducteur et établissement d'un nouveau topique

Les deux aspects de la définition du topique, "à propos" et "identifiabilité", se trouvent donc un peu bousculés par ces données. Celles-ci semblent aussi quelque peu rebelles à un autre aspect de la description des dislocations. De nombreux auteurs, on l'a vu, distinguent fonctionnellement la dislocation à gauche de la dislocation à droite : là où la seconde signale la continuation d'une relation de topique déjà établie, la première, dont je rapproche les IU, signale l'établissement d'une nouvelle relation de topique entre un référent et une prédication. Dans les termes de la théorie du centrage, cette opération constituerait une transition autre que par continuation ou par rétention. Or dans mon corpus, à plusieurs reprises, l'entité réalisée par l'IU en E_i est la même que celle réalisée par le Cr de E_{i-1}. Si cette entité constitue également le Cr de E_i, il s'agirait d'une transition d'un de ces deux types. En voici un exemple :

> *(7) Ce sociologue connu du Saguenay est assez réticent lorsqu'on évoque les réseaux de communication.* ***E_{i-2}*** *Il affirme avoir du mal à discerner à quel type de société s'adresse le consortium….* ***E_{i-1}*** *Et de s'inquiéter des silences des promoteurs sur cette question.* ***E_i*** ***Pour lui****, il y a une certaine antinomie entre réseau et territoire.*

En (7), on observe un pronom inaccentué en E_{i-2}, un sujet non-exprimé en E_{i-1}, il y a donc bien continuation d'énoncé en énoncé, et avec l'expression pronominale de la même entité dans l'IU, qui donc peut difficilement être interprété comme signalant l'introduction d'une nouvelle relation de topique.

Cette observation va dans le même sens que les précédentes et suggère qu'on est ici, en ce qui concerne le regard vers l'arrière, en présence d'un fonctionnement déviant par rapport à la dislocation "classique". Reste à voir dans quelle mesure cette déviance met en cause la possibilité pour l'entité réalisée par l'IU d'être le Cr de l'énoncé.

• Regard vers l'avant : transitions après un IU

On l'a vu, un aspect essentiel de la caractérisation des expressions qui constituent l'objet de cette étude est l'absence de reprise de l'entité réalisée par l'IU dans la prédication qui le suit, ce qui exclut que la même entité soit réalisée dans l'IU et par le sujet grammatical. On a donc dans de nombreux cas[59] au moins deux candidats distincts au rôle de Cp de l'énoncé : l'entité réalisée par l'IU et celle qui occupe la fonction de sujet grammatical. Deux axes d'examen du corpus se présentent en lien avec le type de transition de l'énoncé E_i à l'énoncé E_{i+1} :

– quelle entité constitue dans le corpus le Cr de E_{i+1} ? Trois cas de figure semblent envisageables : l'entité réalisée par l'IU, l'entité réalisée par le sujet grammatical de E_i, ou une autre entité.

– quelles expressions référentielles rencontre-t-on pour ces différents cas ?

[59] Il est également assez fréquent que l'IU soit suivi d'une construction présentationnelle (ex. *il y a…*, cf. 2.2.2) comme dans (7) et dans la deuxième occurrence de (5).

Les observations effectuées sur ces deux axes permettront d'envisager des questions qui concernent directement la théorie du centrage, tout particulièrement le classement des entités réalisées par des IU dans l'échelle de classement des Ca.

a) Cr(E_{i+1}) = entité réalisée par l'IU en E_i

L'entité à laquelle réfère l'IU en E_i peut sous certaines conditions être également celle dénotée par le Cr de E_{i+1} :

> *(8)* ***Pour lui*** *[Pierre W. Boudreault], il y a une certaine antinomie entre réseau et territoire. Le territoire, lieu chaud d'expression de la communauté et le réseau, lien froid vers l'extérieur.* ***Il*** *opère une importante distinction entre interrelations et interactions, d'un côté des signes qui circulent, de l'autre des être humains qui communient.*

L'acceptabilité de cet exemple dans cette section dépend du statut qu'on accorde à l'unité ponctuationnelle "*Le territoire … l'extérieur*", dont le statut syntaxique n'est pas clair. Si l'on y voit une extension de l'énoncé E_i, et non pas un énoncé à part entière, la phrase commençant par "*Il*" est l'énoncé E_{i+1}. C'est tout le problème, évoqué plus haut, de la délimitation des segments et des unités. En tout état de cause, les occurrences de reprises pronominales semblent toutes soumises à des conditions particulières, dont la plus notable est la présence en E_i d'une construction présentationnelle, comme c'est le cas en (8), et donc l'absence d'entité réalisée par le sujet grammatical. L'exemple (5), partiellement repris en (9), représente un cas de figure un peu différent : le premier IU est suivi d'un pan de discours direct, entre guillemets, que je propose de traiter comme un bloc[60]. L'entité réalisée par l'IU et qui constitue le Cr de E_{i+1}, là aussi, n'est donc pas en compétition avec celle que réaliserait un sujet grammatical de même niveau.

> *(9) Pour un acteur universitaire, " le socio-communautaire est plus lié aujourd'hui à l'idée de service. Ceci d'autant plus que l'on assiste dans le même temps à un désengagement de l'État. Cela s'adresse à certaines clientèles spécifiques, notamment le milieu du travail social."* ***Il*** *y voit une certaine source d'inquiétude.*

Dès qu'il existe une entité réalisée par le sujet grammatical, la reprise par un pronom clitique sujet semble plus difficile. Le corpus n'en fournit pas d'exemple. On trouve en revanche des exemples de reprises de l'entité réalisée dans l'IU par une expression démonstrative, comme (10). Cet exemple n'est cependant pas facilement interprétable, dans la mesure où la reprise en E_{i+1} concerne une expression, *la commission*, enchâssée dans *le rapport de la commission* :

> *(10)* ***Selon le rapport de la commission Vérité et Réconciliation, chargée notamment d'enquêter en 1991 sur les crimes commis pendant la dictature,*** *Marcos Quezada n'a pas survécu a la répétition de chocs électriques infligés par les policiers.* ***Cette commission [?Elle]*** *avait alors rejeté la version officielle des autorités selon lesquelles le jeune militant se serait suicidé, l'institut médico-légal de Temuco ayant confirmé que sa mort était due à de violentes décharges électriques.*

La transformation de (10) en (11) par extraction de *la commission* rend un peu plus acceptable l'utilisation du clitique, qui reste cependant problématique :

> *(11)* ***Selon la commission Vérité et Réconciliation, chargée notamment d'enquêter en 1991 sur les crimes commis pendant la dictature****, Marcos Quezada n'****a pas survécu a la répétition de chocs électriques infligés par les policiers. Cette***

[60] Je suis en cela Kameyama, qui propose d'analyser le discours direct comme un segment de centrage enchâssé qui est inaccessible au segment de centrage de niveau supérieur (Kameyama, 1998:107)

commission {?Elle] *avait alors rejeté la version officielle des autorités selon lesquelles le jeune militant se serait suicidé….*

Ces observations sur des exemples attestés en français vont dans le même sens que les expérimentations effectuées par P. Gordon *et al.* (1993, expérimentations 2 et 3) sur des exemples fabriqués (en anglais) dans le but de déterminer le statut respectif de l'entité réalisée par le sujet grammatical et de celle réalisée par un syntagme prépositionnel détaché à gauche. Ces expérimentations font intervenir la notion de pénalité du nom répété comme marqueur du Cr d'un énoncé : le temps de lecture augmente si le Cr d'un énoncé est réalisé par la répétition du nom plutôt que par l'anaphore pronominale attendue. Cette pénalité s'attache nettement à la reprise par nom répété du sujet grammatical de l'énoncé précédent, alors qu'elle n'est pas observée dans le cas de reprise de l'entité réalisée dans le syntagme prépositionnel. Les auteurs en concluent que, si cette pénalité est bien un marqueur du Cr, le sujet grammatical est le Cr dans tous les cas où cela est possible, même s'il n'est pas en position initiale. La mise en relation des conclusions de cette étude avec mes observations n'est pas immédiate : il s'agit d'une expérimentation psycholinguistique visant à tester pour l'anglais l'impact de l'utilisation de noms répétés là où un pronom serait attendu, alors que je cherche à examiner, sur corpus et pour le français, les conditions dans lesquelles l'entité réalisée par un syntagme prépositionnel peut faire l'objet d'une reprise. Ces travaux mettent néanmoins tous deux en lumière la difficulté pour l'entité réalisée par un IU à être le Cr de l'énoncé en présence d'un sujet grammatical.

b) $Cr(E_{i+1})$ = entité réalisée par le sujet grammatical en E_i

Le corpus offre quelques exemples où le Cr de E_{i+1}, reprend l'entité réalisée par le sujet grammatical de E_i. Les expressions concernées peuvent être des pronoms comme dans les exemples (12) et (13), des expressions définies, ou démonstratives comme en (14) :

(12) ***Pour Maurice Volkowitsch,*** *les axes et les carrefours structurent l'espace.* ***Ils*** *introduisent une différenciation entre les lieux en fonction de la distance de ces lieux aux éléments d'un réseau.*

(13) ***Pour Louis Kahn,*** *la lumière appartient à cette catégorie ;* ***elle*** *n'est pas seulement ce qui nous rend visibles les choses, mais* ***elle*** *est la substance même, contenant les lois naturelles, connues ou ignorées.*

(14) ***Selon Bourdos lui-même,*** *ce roman, réputé inadaptable, est "d'une incroyable densité".* ***Cet ouvrage complexe*** *mêle en effet plusieurs histoires.*

On a donc ici la situation déjà évoquée de compétition pour le statut de topique entre les entités réalisées respectivement par l'IU et le sujet grammatical. La configuration où le $Cr(E_{i+1})$ correspond au sujet de E_i paraît plus naturelle que celle où le $Cr(E_{i+1})$ correspond à l'entité réalisée par l'IU de E_i (cf. 1)) dans la mesure où elle ne semble pas s'assortir de conditions particulières. Je n'ai pas à ce stade du travail cherché à identifier les paramètres qui interviennent dans la détermination du type de reprise.

c) Classement des entités et transitions

Le classement des Centres anticipateurs (Ca) pour l'anglais est au départ fondé sur l'échelle de S. Brennan *et al* (1987), axée sur les fonctions grammaticales :

Sujet > Objet(s) > Autres

Selon cette échelle, les entités réalisées par le sujet ont plus de chances d'être le Cr de l'énoncé suivant que celles réalisées par un objet, qui sont elles-mêmes classées avant celles réalisées par d'autres fonctions grammaticales. Faut-il penser que les constituants détachés à gauche font partie de cette dernière catégorie ? Ou plutôt que leur cas n'a pas encore été

envisagé[61] ? L'étude des paramètres linguistiques qui peuvent être associés au classement des Ca, paramètres propres à chaque langue, constitue un champ de recherche important pour la théorie du centrage. M. Walker *et al* (1994) ont ainsi proposé le classement suivant pour le japonais :

Topique > Empathie > Sujet > Objet(s) > Autres[62]

Le topique est marqué morphologiquement en japonais ; or pour le français, les travaux sur la dislocation suggèrent que le détachement à gauche peut s'interpréter comme une marque positionnelle du topique[63]. Si l'entité réalisée par l'IU s'apparente au topique, elle devrait alors, comme en japonais, l'emporter sur le sujet dans le classement. Dans ce cas, la théorie prédirait que les exemples (8) et (9), où $Cr(E_{i+1})$ = l'entité réalisée par l'IU en E_i, devraient sembler plus naturels, plus cohérents que (12)-(14), où $Cr(E_{i+1})$ = l'entité réalisée par le sujet grammatical en E_i. Cette analyse est toutefois difficile à concilier avec l'intuition, et avec les restrictions qui s'imposent précisément dans les cas de reprise de l'entité réalisée par l'IU. S'il semble inapproprié, dans les exemples à l'étude, de classer les entités réalisées par les IU avant celles réalisées par le sujet, c'est peut-être parce qu'il s'agit d'entités qui fonctionnent sur des plans différents dans le discours. À la suite d'un certain nombre d'auteurs, principalement préoccupés quant à eux par la résolution d'anaphores distantes (Hahn & Strube, 1997; Hitzeman & Poesio, 1998; *inter alia*), il me semble incontournable de prendre du recul et d'envisager ces fonctionnements en relation avec l'organisation du discours sur un plan plus global. Cette démarche est d'ailleurs suggérée dans l'analyse initiale de M. Charolles, qui insiste sur la propension qu'ont les IU à étendre leur portée sur plusieurs énoncés.

6.1.5 Portée des introducteurs d'univers et structuration du discours

M. Charolles (1997) observe d'une façon générale qu'un cadre, une fois installé par une expression introductrice, est en mesure d'inclure plusieurs propositions. Il affine cette première observation par l'examen détaillé d'univers de discours introduits par l'expression *selon X,* dont il décrit le "grand pouvoir intégrateur". On pourrait être tenté de paraphraser *selon X, p* par *X dit (pense, soutient, ...) que P*, pourtant la comparaison du potentiel d'intégration de l'IU avec l'emploi de verbes *dicendi* met clairement en lumière un fonctionnement spécifique du premier. Là où le verbe *dicendi* ne peut spécifier le cadre énonciatif que de la seule subordonnée qu'il domine, ou d'une série de subordonnées à l'intérieur d'une phrase complexe à condition que soit répété le *que* de subordination, l'IU autorise une grande liberté. Ce qui m'intéresse ici, c'est que l'IU fonctionne alors comme la borne initiale d'un segment, qui sera clos par une "rupture énonciative" signalée par certaines marques ou constellations de marques. L'observation de ce fonctionnement nécessite que l'on associe à l'examen détaillé des expressions réalisant des entités dans chaque énoncé un regard plus global sur l'organisation entre segments. Je vais donc dans ce qui suit ouvrir quelques pistes pour un examen des IU dans cette perspective globale, ce qui va m'amener à les envisager tour à tour dans le cadre de la théorie du discours à laquelle se

---

[61] Les travaux typologiques de Lambrecht sur la dislocation (Lambrecht, à paraître) montrent pourtant que cette construction se retrouve dans de nombreuses langues.

[62] En japonais, l'entité signalée comme lieu de l'empathie est celle avec lequel le locuteur s'identifie, ou celle dont il adopte le point de vue (Cf. Walker, Joshi & Prince, 1998:9).

[63] La particule *wa* du japonais n'a toutefois pas pour unique fonction de marquer le topique, de même que la fonction grammaticale sujet pour l'anglais. Il s'agit, pour reprendre le terme de Davison (1984), de l'exploitation pragmatique de traits qui sont aussi des traits des constituants de phrase. Le détachement à gauche, dans la mesure où il est extérieur au fonctionnement syntaxique de la prédication, pourrait être considéré comme un marqueur plus spécifique de relations pragmatico-discursives.

rattache le centrage et dans celui du modèle de l'Architecture Textuelle présenté au chapitre 5 (1.2).

• Introducteurs d'univers et "cue phrases"

La théorie du centrage, principalement concernée par la référence pronominale, fournit un modèle plutôt atomiste du fonctionnement discursif, selon lequel la référence se construit d'énoncé en énoncé à l'intérieur d'un segment de discours. La relation entre ce focus local, qui modélise les changements d'états attentionnels à l'intérieur d'un segment de discours, et le focus global, intéressé par le niveau intersegmental, a fait l'objet de nombreux travaux dans le cadre plus global de la théorie du discours dans laquelle s'inscrit le centrage (Grosz & Sidner, 1986; Grosz & Ziv, 1998; Brennan, 1998; Passonneau, 1998; Walker, 1998; à paraître, *inter alia*). Dans ce cadre, les IU pourraient être envisagés comme un type de "cue phrases", marques explicites de la structuration du discours. B. Grosz et C. Sidner (*op.cit.*) développent cette notion de marques linguistiques qui déterminent des segments soit par rapport à la structure intentionnelle, soit en termes d'état attentionnel.

La structure intentionnelle a trait au but du discours ainsi qu'aux buts des segments qui le constituent. La particularité des buts de discours par rapport à d'autres intentions impliquées dans la communication langagière, selon ces auteurs – et on retrouve là un point également développé dans le modèle de l'Architecture textuelle –, est qu'ils doivent être reconnus par l'interlocuteur. En d'autres termes l'identification du but de discours ou du but de segment de discours est essentielle pour que l'effet visé par le locuteur soit atteint (Grosz & Sidner, 1986:178; cf. 5.1.2 : Métalangage et mise en forme matérielle). Deux relations structurelles principales sont posées entre buts de discours : *dominance* et *précédence* ("satisfaction precedence"). Si un but de segment de discours (Discourse Segment Purpose) DSP1 contribue au but principal DSP2, DSP2 domine DSP1. Si de plus l'ordre de satisfaction de ces buts compte, et que DSP1 doit être satisfait avant DSP2, DSP1 précède en satisfaction DSP2 ("DSP1 satisfaction-precedes DSP2").

Attentional Change	
(push)	*now, next, that reminds me, and, but*
(pop)	*anyway, but anyway, in any case, now back to*
(complete)	*the end, ok, fine, (paragraph break)*
True interruption	
I must interrupt, excuse me	
Flashbacks	
Oops, I forgot	
Digressions	
by the way, incidentally, speaking of, did you hear about…, that reminds me	
Satisfaction-precedes	
in the first place, first, second, finally, moreover, furthermore	
New dominance	
for example, to wit, first, second, and, moreover, furthermore, t herefore, finally	

Figure 6.1 : L'utilisation des "cue phrases" (Grosz & Sidner, 1986:198).

L'état attentionnel est une abstraction du focus d'attention des participants du discours au fur et à mesure de son déroulement. Il est représenté dynamiquement en termes d'espaces de focus et de règles de transition qui régissent les conditions d'ajouts et de retraits d'espaces. Les transitions sont déterminées par la structure intentionnelle. Chaque espace de focus contient également un but de segment de discours, pour rendre compte du fait que les participants se focalisent à la fois sur ce dont il est question et sur la raison qui les a amenés à en parler. La structure de focus est représentée comme une pile (informatique), c'est-à-dire une structure de données dans laquelle on empile les nouveaux éléments et qui limite l'accès direct au dernier élément ajouté. Les espaces en bas de la pile, quoique encore accessibles, le

sont moins immédiatement que les espaces en haut de la pile. Cette analogie de la pile amène les auteurs à décrire les transitions comme le fait de "pousser" un espace de focus en haut de la pile ("push"), d'évacuer un espace de la pile ("pop"), ou de fermer un espace ("complete"). Le centrage, on l'a vu, s'intéresse à la relation entre état attentionnel, choix d'expression référentielle et types de transition à l'intérieur d'un segment de discours. Les "cue phrases" sont des moyens abrégés et indirects par lesquels le locuteur peut informer l'interlocuteur qu'un changement de focus est imminent[64]. Le choix d'une "cue phrase" donnée est de plus indicatif du type de changement à attendre. Ce bref résumé me permet de présenter la liste non-exhaustive de cas de changements d'espace de focus, chaque cas étant accompagné de "cue phrases" caractéristiques (figure 6.1).

Dans cette optique, bien que ces "cue phrases" soient sémantiquement tout à fait différentes des exemples traités, les IU pourraient être envisagés comme indiquant un changement attentionnel de type "push". Cette interprétation me semble cependant peu satisfaisante pour un ensemble de raisons. Premièrement, le fait que, contrairement aux "cue phrases" et aux connecteurs "classiques", les IU réalisent une entité qui va pouvoir être intégrée dans le discours change fondamentalement la donne. Dans de nombreux cas, on l'a vu, l'entité réalisée par l'IU ne remplace pas, mais vient s'ajouter à l'entité focale. Si l'IU détermine bien un segment, en fournissant un critère d'interprétation des propositions dans sa portée, il y a donc lieu de se demander s'il s'agit d'un segment attentionnel. Le fonctionnement qu'on observe est complexe, et particulièrement intéressant pour l'étude de l'organisation du discours, puisqu'il touche conjointement à plusieurs de ses mécanismes. Les IU étudiés ici servent à la fois à marquer la borne initiale d'un segment, à introduire une entité de discours et à placer cette entité dans une position propre à la faire interpréter comme saillante pour l'ensemble du segment, qu'elles contribuent par là même à définir. Si ces segments définis par les IU sont autre chose que des segments attentionnels, l'interaction entre ces différentes formes de segmentation doit être comprise. Je vais proposer un autre éclairage, celui du modèle de l'Architecture textuelle, pour avancer dans cette voie.

• Introducteurs d'univers de discours et architecture textuelle

L'attention portée dans le modèle de l'architecture textuelle aux aspects visuels du texte, dans la mesure où est posée une équivalence fonctionnelle entre marques lexico-syntaxiques et marques visuelles (typographiques, dispositionnelles), a conduit à la recherche de modes de représentation des textes capables de rendre compte de ces aspects. L'*image de texte* permet de reproduire des textes longs en en faisant ressortir les marques de mise en forme matérielle pertinentes. L'image de texte ci-dessous (figure 6.2)[65], donne l'organisation générale du texte d'où est issu l'exemple (1), tout en indiquant le détail des expressions réalisant les centres des énoncés. (Les expressions qui reprennent au fil du texte l'entité réalisée dans l'IU *Pour Maurice Volkowitsch* sont en gras, celles qui reprennent le sujet de l'énoncé, *les axes et les carrefours*, sont en italiques.)

Cette image de texte rend manifeste une structuration en deux parties numérotées et titrées. Il est à noter que le rôle et le fonctionnement dans le discours des référents introduits dans les titres semblent avoir été largement laissés de côté dans la littérature. Ici, l'entité réalisée par le titre de la partie 1, *les réseaux de transports*, se retrouve dans le sujet grammatical de plusieurs phrases, sous la forme d'expressions définies ou de clitiques. Dans

---

[64] Les auteurs précisent que ces expressions sont d'une part multifonctionnelles, c'est-à-dire qu'elles peuvent avoir une fonction autre que la fonction discursive de "cue phrase"; d'autre part ambiguës dans leur fonctionnement de "cue phrases", c'est-à-dire que la même expression peut remplir plusieurs rôles discursif. Cette observation rejoint mon point de vue sur les marqueurs, formulé pour le marquage du thème en 2.2.2.

[65] Il s'agit ici d'une image de texte *a minima*. Des formes de représentation beaucoup plus précises, qui incluent les divers aspects de la mise en forme matérielle, sont en cours de développement par Christophe Luc et Mustapha Mojahid (IRIT).

certaines phrases, comme dans l'exemple (1), on a considéré qu'elle se trouvait en compétition pour le statut de Cp avec l'entité réalisée par l'IU, *Maurice Volkowitsch*. Dans la partie 1, consacrée aux réseaux de transport, ceux-ci sont examinés à travers deux auteurs. Ces références structurent la partie en deux sous-parties consacrées respectivement aux réseaux de transport vus par l'auteur 1, *M. Volkowitsch*, et par l'auteur 2, *J-M. Offner*. J'interprète la portée de l'IU *Pour Maurice Volkowitsch* comme allant jusqu'à la phrase commençant par *De son côté, Jean-Marc Offner*. L'univers introduit par *Pour Maurice Volkowitsch* détermine donc ici un segment à l'intérieur de la partie consacrée aux réseaux de transport. À l'intérieur de cette sous-partie, le Cr de chaque phrase est soit l'auteur 1, soit les réseaux de transport. Ces entités sont toutes deux accessibles et actives dans l'intégralité du segment, et semblent correspondre à deux plans imbriqués de l'organisation du texte.

LES EFFETS STRUCTURANTS DES TRANSPORTS : MYTHES ET REALITES

_____ _____ nous traiterons séparément du pouvoir structurant des transports et de celui des techniques d'information et de communication. _____.

1. Les effets structurants des réseaux de transport.

Les réseaux de transport physiques interviennent _____ _____. Pour **Maurice Volkowitsch**, *les axes et les carrefours* structurent l'espace. *Ils* introduisent _____ _____. *Les carrefours* bénéficient _____ _____. *Les réseaux de transport* structurent _____ _____. **Volkowitsch** met en garde _____ _____. **Il** estime _____ _____. **Volkowitsch** estime _____. _____, **il** relève _____. Pour **lui**, la corrélation entre la politique des réseaux et la formation des banlieues _____ _____. De son côté, Jean-Marc Offner constate _____ _____. Pour **le premier**, *le pouvoir structurant de ces réseaux* n'est _____ _____. Pour Offner, *les réseaux de transport* stimulent _____ _____.

2. Les effets structurants des télécommunications sur l'espace.

_____…

Figure 6.2. Image de texte pour l'exemple (1).

Si l'on reprend à la lumière de cette présentation l'examen de la similarité entre les IU et les "cue phrases" annonçant un changement d'état attentionnel de type "push", il est clair que l'IU *Selon Maurice Volkowitsch* ne substitue pas l'entité *Maurice Volkowitsch* à l'entité *réseaux de transports* dans la pile des espaces de focus pour le segment puisque cette dernière continue à apparaître dans des expressions qui la feront interpréter comme le Cr de plusieurs énoncés de ce segment. Par ailleurs, la segmentation en paragraphes et la titraison définissent explicitement un segment attentionnel et un but de segment de discours. Le sous-segment introduit par l'IU *Selon Maurice Volkowitsch* ne me paraît pas non plus correspondre à la définition de "new dominance" (figure 6.2) dans la mesure où, là encore, l'entité qu'il réalise n'ouvre pas un espace de focus dominé par l'espace concernant l'entité *réseaux de transports* mais oriente l'interprétation d'un sous-segment dans cet espace. Ces deux segmentations sont donc conjointes, et semblent correspondre à des niveaux différents de structuration.

Ces considérations inspirées par le modèle de l'Architecture textuelle sont tout à fait préliminaires : mon travail sur les IU a jusqu'à présent été mené dans le cadre du centrage, l'éclairage que pourrait apporter le modèle de l'Architecture textuelle fait partie des pistes à explorer, dont certaines seront définies plus précisément ci-dessous.

6.1.6 Les IU entre différents niveaux d'organisation textuelle

Cette étude a proposé une approche de la contribution des IU à la construction du modèle de discours. C'est un premier défrichage, qui a porté sur un sous-ensemble restreint d'expressions, et qui est loin d'apporter des réponses à toutes les questions posées en 6.1.3. Son cheminement a été, à partir de similarités formelles et définitionnelles entre IU et dislocations, d'explorer la possibilité que ces expressions fonctionnent comme des topiques, pour les envisager ensuite en termes de centrage. Beaucoup reste à faire pour mieux cerner le fonctionnement des IU, en particulier à travers l'examen de divers paramètres nécessitant une analyse fine et qui ont sans doute été insuffisamment pris en compte dans cette étude : il faudrait cartographier de façon beaucoup plus précise les types d'expressions référentielles dans l'IU, dans la prédication, et dans les énoncés environnants ; analyser en corpus la relation entre ces caractéristiques formelles et le fonctionnement discursif des IU notamment sur le plan de la portée ; procéder au suivi précis des transitions à l'intérieur des segments délimités par les IU. Au-delà de l'ensemble restreint d'IU étudié ici, c'est le fonctionnement des introducteurs de cadres porteurs d'une expression référentielle qui est en cause, fonctionnement apparemment paradoxal qui semble tenir à la fois de celui des topiques et de celui des "cue phrases". En effet, ils ont la particularité d'être extérieurs au contenu propositionnel, comme des "cue phrases" ou des connecteurs, en même temps qu'ils introduisent une entité qui devient disponible pour les prédications suivantes. Mon intuition est que ces prédications sont cependant limitées, par exemple aux verbes *dicendi* dans les cas envisagés plus haut. Ceci reste à vérifier en corpus.

En ce qui concerne la mise en relation des niveaux d'organisation textuelle, le fonctionnement des IU me semble intéressant à plus d'un titre. Sur le plan de la structure d'information, ils s'apparentent formellement aux dislocations et sembleraient donc être des topiques, mais ils n'en partagent pas certaines caractéristiques fondamentales. Le lien entre topiques non-liés, "scene-setting topics" et IU, pourtant convaincant au vu des caractéristiques formelles et des définitions, aurait donc besoin d'être affiné. Mais c'est leur rôle dans la segmentation du discours qui leur confère un intérêt tout particulier dans cette section sur les niveaux d'organisation : ils contribuent à la segmentation du discours en marquant la borne initiale d'un "cadre", dit M. Charolles (1997). Reste à analyser comment s'articule le segment délimité par un cadre avec ceux mis au jour par d'autres approches des textes : segments intentionnels pour la théorie de Grosz et Sidner, objets textuels pour le modèle de l'Architecture textuelle, segments résultant d'une relation rhétorique pour la RST. Le modèle de B. Grosz et C. Sidner, et le modèle du centrage qui en est issu, présentent l'intérêt de modéliser avec précision la relation entre l'introduction et le suivi des entités dont on parle avec les intentions discursives. Dans les termes du modèle hallidayen présenté dans le chapitre 5 (5.1), il s'agirait de l'interaction entre les composantes idéationnelles et interpersonnelles. Dans cette interaction, j'ai montré que le rôle des IU pose problème : ils ne correspondent ni à un état attentionnel ni à une intention, mais à une instruction de lecture pour un segment qui n'a d'autre existence qu'en relation à cette instruction.

6.2 L'articulation des niveaux d'organisation textuelle : perspectives

La réflexion sur l'articulation des niveaux d'organisation textuelle passe par la mise en relation de différents modèles du discours, avec leur focalisation propre sur certains niveaux et certaines interactions entre niveaux. Une telle mise en relation dépasse l'empan du travail d'un seul chercheur, c'est pourquoi un projet plus vaste est en train de se mettre en place, grâce à l'invitation qui m'a été faite de coordonner un numéro de la revue *Verbum* sur le thème de la cohérence textuelle[66]. J'ai répondu à cette invitation, formulée en 1998, par un projet visant à lancer des ponts entre cinq approches du discours, et à poser explicitement la question de la signalisation dans les textes. Les auteurs invités à contribuer à ce projet incluent plusieurs des concepteurs des modèles présentés (W. Mann et S. Thompson pour la RST, N. Asher pour la SDRT, J Virbel pour le MAT), et tous se sont montrés intéressés par l'idée de ce dialogue entre modèles comme façon d'avancer dans la réflexion. Le numéro sortira en mars 2001. Je reproduis ci-dessous le texte du projet, suivi de la liste des auteurs :

Numéro thématique de Verbum :

Cohérence et relations de discours à l'écrit

De nombreux théoriciens du discours font appel pour rendre compte de la cohérence textuelle à une structuration récursive mettant en jeu des segments de texte reliés par des relations. En linguistique informatique, la génération de texte est le lieu de travaux particulièrement intéressants dans la mesure où ils sont tenus de faire le lien entre des intentions discursives et des réalisations linguistiques. Ces travaux présentent l'intérêt supplémentaire d'être soumis à des exigences de précision et de formalisation. Le modèle le plus productif dans ce contexte est la Rhetorical Structure Theory (RST), proposé initialement par W. Mann & S. Thompson à la fin des années 80. Un aspect problématique de ce modèle, comme d'autres qui sont également fondés sur la notion de relation de discours, concerne l'hétérogénéité des relations. On cherche à distinguer et à articuler des relations d'ordres différents, qui amènent à postuler différents niveaux d'organisation des textes :

– des relations "idéationnelles", ou "référentielles", ou encore "sémantiques", par exemple entre un segment présentant une cause et un segment présentant l'effet relevant de cette cause;

– des relations "pragmatiques", ayant trait aux intentions du scripteur, par exemple entre deux segments dont l'un a pour but de justifier l'énonciation de l'autre;

– des relations d'ordre proprement textuel, par exemple entre un segment conclusif et le segment dont il est la conclusion.

[66] J'en remercie Michel Charolles et les membres du comité de rédaction de *Verbum*.

Par ailleurs, un problème central pour la description du fonctionnement linguistique comme pour les applications en traitement automatique des langues (génération de texte, extraction d'information à partir de documents) est celui de la signalisation des relations de discours. Les auteurs de la RST mettent en garde les chercheurs, en déclarant qu'il n'y a pas à leur avis de signalisation systématique. Il y a lieu cependant de se demander si ce constat n'est pas lié à certaines options méthodologiques : d'abord le fait de se situer en langue générale, sans chercher à examiner les régularités possibles dans des configurations genre/domaine spécifiques; deuxièmement l'impasse que font la plupart des travaux quant aux aspects visuels des textes, qui ont pourtant un rôle fondamental dans la structuration du modèle du texte lors de l'interprétation. Enfin on constate une absence de dialogue entre les nombreux travaux portant sur des connecteurs spécifiques et ceux prenant comme point de départ une relation ou une structure textuelle pour en examiner les réalisations.

Hétérogénéité des relations et niveaux d'organisation textuelle, signalisation de ces relations, voilà les thèmes sur lesquels j'envisage d'inviter des contributions pour ce numéro de Verbum, selon les problématiques esquissées ci-dessus. Pour favoriser la cohérence du numéro, je propose qu'il s'en tienne à l'écrit.

Auteurs :

– Nicholas Asher (University of Texas at Austin), Joan Busquets (Université de Bordeaux/ERSS) et Laure Vieu (Institut de Recherche en Informatique de Toulouse) sur la Segmented Discourse Representation Theory;

– John Bateman (Universités de Stirling et Bremen) sur le modèle des Conjunctive Relations;

– Corinne Rossari et Michel Charolles (Université Paris III) sur l'école de Genève;

– William Mann (SIL) et Sandra Thomson (University of California at Santa Barbara) sur la Rhetorical Sructure Theory;

– Jacques Virbel (Institut de Recherche en Informatique de Toulouse) sur le Modèle de l'Architecture Textuelle.

À suivre...

Chapitre 7

La signalisation du texte : marqueurs d'organisation textuelle

L'idée que les opérations de mise en texte laissent des traces à la surface des textes, traces qui constituent une signalisation orientant l'interprétation, fonde les travaux présentés dans ce mémoire, en parcourt toutes les étapes, et continue à motiver mes projets. Au fur et à mesure des lectures, des rencontres, des analyses, ma notion de marqueur d'organisation textuelle s'est enrichie et précisée. Je propose ici une tentative de bilan, où les connexions seront nombreuses avec les deux autres grands axes de ma réflexion, les niveaux d'organisation et leur interaction d'une part, l'impact du genre discursif et du domaine de l'autre.

Je rappelle pour commencer le choix de démarche formulé dans le Préambule (cf. Péry-Woodley, 1994) : aller préférentiellement des fonctions aux marqueurs plutôt que des marqueurs aux fonctions. Dans les années 80 en effet, quand ont commencé à prendre forme les travaux qui font l'objet de ce mémoire, l'intérêt croissant pour la dimension pragmatique des échanges langagiers et pour le dialogue, perçu comme particulièrement révélateur de la prégnance de cette dimension, a conduit de nombreux linguistes à examiner les "connecteurs", "clue words" ou "cue phrases" qui orientent l'interprétation des contenus propositionnels. Citons R. Reichman (1984), R. Cohen (1984), D. Brée & R. Smit (1986) pour l'anglais, O. Ducrot, avec la focalisation sur l'argumentation (1980; 1983a; 1983b), ou E. Roulet *et al* (1985) pour le français, *inter alia*. Là où ces recherches sont concernées par l'étude des marqueurs (les travaux de O. Ducrot et de E. Roulet ont des objectifs plus vastes), elles prennent comme point de départ des marqueurs spécifiques, toujours lexicaux, et cherchent à en cartographier le fonctionnement. On trouve ainsi des études sur *mais, donc, même, puisque* …, Et , dans les travaux sur l'anglais, sur *by the way, now, (and) so*, etc. Les marqueurs sont examinés en contexte, le plus souvent dans des corpus de dialogues, même s'il s'agit parfois de dialogue écrit (Ducrot, 1980), quelquefois dans des énoncés fabriqués (Cohen, 1984).

Cette réflexion méthodologique rejoint celle de D. Biber (1988, ch. 4) sur l'approche de la variation dans les textes, qui le conduit à opposer une approche "microscopique" et une approche "macroscopique". La première, comme celles citées ci-dessus pour les marqueurs d'organisation textuelle, est axée sur le fonctionnement de certains traits (passé *versus* présent (Schiffrin, 1981)), de certaines constructions (propositions de but (Thompson, 1983)), de marques lexicales (*actually, really* (Aijmer, 1986), Stenström (1986)). La seconde

se donne au contraire pour but l'identification "tous azimuts" des différents paramètres de la variation linguistique. Étant donné mon objectif d'identifier les différentes marques potentiellement impliquées dans l'organisation textuelle, dans le cadre d'applications informatiques en particulier, c'est une approche de type macroscopique qu'il me faut adopter. L'étude de marqueurs lexicaux pré-établis est en effet insuffisante pour deux raisons. D'abord plusieurs auteurs (voir Mann & Thompson, 1987 (cf. note 51, 5.3.1), 1988; Redeker, 1990) ont montré que le marqueur lexical n'est pas indispensable pour que soit réalisée la fonction à laquelle il est régulièrement associé (hiérarchisation, relation sémantico-pragmatique). La focalisation sur les marqueurs lexicaux a pu dans ce cas conduire à interpréter leur absence comme une absence de signalisation, alors qu'il faudrait, me semble-t-il, ouvrir la notion de signalisation pour en identifier d'autres modalités. Deuxièmement, on a souvent évoqué la plurifonctionnalité des marqueurs lexicaux. G. Redeker (1990), reprenant D. Schiffrin (1987) souligne la propension de nombreux marqueurs à associer une fonction idéationnelle et une fonction pragmatique. Elle identifie ainsi un usage sémantique et un usage pragmatique de *because*. J-M. Adam et F. Revaz (1989) illustrent le passage d'un usage chronologique à un usage logique pour un connecteur comme *alors* selon la nature narrative ou argumentative du texte ou du segment. Je suggère moi-même (6.1.1) que les mêmes expressions peuvent fonctionner comme adjoints ou comme introducteurs d'univers de discours selon leur position dans l'énoncé. Dans tous ces cas, une même forme fournit des instructions d'interprétation différentes selon l'usage qui en est fait. J'explore donc ici des approches qui cherchent à associer plusieurs traits de manière à identifier des configurations fonctionnellement fiables. Par ailleurs, je fais l'hypothèse que la signalisation de l'organisation textuelle est fonction du domaine et du genre discursif.

Le parti pris dominant qui me conduit à envisager la fonction comme première n'est toutefois pas exclusif : j'ai réalisé des études, et j'ai des chantiers en cours, où la démarche est différente. C'est le cas de l'étude comparative de la fonction du passif dans la structuration thématique décrite dans le chapitre 2 (2.2, et Péry-Woodley, 1991a). L'étude présentée dans le chapitre précédent (Péry-Woodley, à paraître) se focalise sur une classe de marqueurs, les introducteurs d'univers de discours, pour examiner sur le plan de la structure thématique des expressions initialement analysées quant à leur rôle sur un plan sémantico-pragmatique. J'ai par ailleurs entrepris une étude du rôle des clivées dans l'organisation textuelle, alors qu'elles ont surtout été envisagées par rapport à la structure d'information. Ici la focalisation sur un type de marqueur permet d'aborder l'articulation des différents niveaux d'organisation des textes.

Ce chapitre s'organise en trois sections. La première défend l'hypothèse de départ qu'il existe une signalisation de l'organisation textuelle, et cherche à formuler ce qui peut faire l'objet d'une signalisation. La seconde décrit la méthode d'identification en corpus des réalisations de cette signalisation, et précise ce que recouvre dans cette optique la dénomination *marqueur de l'organisation textuelle*. En guise d'illustration, je présente une étude en cours sur l'énumération, qui exemplifie les prises de position exposées jusque là. La troisième évoque les applications informatiques intéressées par le repérage et l'étude fine des marqueurs d'organisation textuelle.

7.1 Que marquent les marqueurs ?

La définition du texte proposée par F. Cornish et citée dans la section 1.1.2 envisage celui-ci comme "the connected sequence of verbal signs and non-verbal signals in terms of which discourse is co-constructed by the participants in the act of communication". Quels sont parmi ces signes et ces signaux ceux qui contribuent spécifiquement à l'organisation du texte plutôt qu'aux contenus propositionnels qu'il véhicule ? J'ai cherché des réponses dans plusieurs approches de la modélisation de l'organisation textuelle, qui m'ont amenée à identifier trois niveaux d'organisation :

1) la structuration liée aux entités concernées par les propositions (structure thématique dans le chapitre 2, reprise en termes de focus attentionnel dans le chapitre 6) ;

2) la structuration déterminée par les relations entre ces propositions, et, de façon récursive, entre les segments résultants (structure rhétorique, chapitre 3) ;

3) la structure produite par l'ensemble des actes textuels concernés par la mise en texte de ce matériau propositionnel et rhétorique (architecture textuelle, chapitre 5).

La recherche de marqueurs de ces différents niveaux d'organisation implique l'hypothèse forte qu'ils font effectivement l'objet d'un marquage dans le texte. Cette hypothèse s'oppose à la vision radicalement pragmatique évoquée dans la section 1.1, qui donne un rôle primordial aux facteurs contextuels et aux connaissances partagées. J'ai défendu ce parti pris en m'appuyant sur les caractéristiques propres au discours écrit, en particulier la distanciation et sa conséquence, le non-partage de la situation d'énonciation. À ces propriétés générales de l'écrit peut s'ajouter l'hypothèse d'une exigence de cohérence explicite (Reinhart, 1980) pour certains genres discursifs, tels les manuels de logiciels étudiés au chapitre 5, et d'une façon plus générale les documents techniques concernés par les applications en TAL. S'appuyant sur le principe pragmatique de cohérence qui régit l'interprétation des discours, T. Reinhart distingue les textes explicitement cohérents de ceux qui sont implicitement cohérents, distinction qui a trait non pas à leur interprétabilité mais aux types d'opérations nécessités par le processus d'interprétation dans chacun des cas. Pour être explicitement cohérent, et ne pas nécessiter (et risquer) l'application de procédures particulières pour lui attribuer une cohérence dérivée, un texte doit respecter les conditions de cohésion, non-contradiction et pertinence. J'ai conscience d'interpréter un peu librement la proposition de T. Reinhart, qui n'envisage que la cohésion alors que je fais référence à la signalisation des différents niveaux d'organisation évoqués ci-dessus. La possibilité de différentes exigences de signalisation selon le genre discursif, proposée par R. Lakoff (1984) dans le cadre d'une comparaison entre cultures orales et cultures écrites, me paraît intuitivement acceptable, et mériterait d'être examinée empiriquement.

L'hypothèse du marquage des différents niveaux de l'organisation textuelle trouve par ailleurs un appui théorique dans les travaux de Z. Harris (1968), repris par J. Virbel (1985), sur la relation ente langage et métalangage. Selon Z. Harris, le langage contient son propre métalangage, et un discours son propre métadiscours comme une partie de lui-même. Les opérateurs métalinguistiques, en relation avec leurs arguments, donnent naissance à des métaphrases qui explicitent comment cet argument (un segment de discours) doit être entendu. Les métaphrases sont diversement réductibles ou effaçables en fonction de paramètres co-textuels et contextuels, réductions et effacements qui laissent des traces spécifiques sur les arguments concernés. Les "marqueurs" seraient donc des traces du métalangage. C'est à partir d'une extension de cette conception à la notion de performatif textuel que J. Virbel (1985; 1989) et E. Pascual (1991) élaborent le modèle de représentation de l'Architecture textuelle (cf. section 5.1.2), et la notion de mise en forme matérielle, c'est-à-dire l'ensemble des traces du métalangage textuel.

Quel que soit le niveau d'organisation textuelle envisagé, structurer revient toujours à segmenter, c'est-à-dire à la fois à diviser et à rassembler. Les marqueurs sont donc dans de nombreux cas à la fois indices d'une fonction et bornes d'un segment. C'est le cas des IU du chapitre précédent, qui fournissent un critère pour l'interprétation des propositions dans leur portée et qui déterminent ainsi un segment. C'est le cas aussi des expressions de la relation d'hyperonymie, qui marquent dans le corpus étudié au chapitre 5 la borne initiale d'une définition, qui est par ailleurs une consigne. Ce deuxième exemple rappelle que co-existent dans le texte plusieurs types de segments qui correspondent à différents niveaux d'organisation. Ces segments peuvent ou non être "isomorphiques". Ainsi on a vu que les IU marquent la borne initiale d'un segment dont l'interaction avec d'autres formes d'organisation pose problème. On verra de même dans l'étude de l'énumération que la relation entre item et phrase ou paragraphe n'est pas simple.

7.2 Identifier des marqueurs d'organisation textuelle

7.2.1 Méthode et définition

Étant donné le parti pris de départ – partir de fonctions et non de marqueurs pré-identifiés –, et un cadre d'analyse dérivé de différents modèles et conçu en termes de trois principaux niveaux d'organisation, quels traits linguistiques prendre en compte et comment les identifier ? La méthode qui s'élabore au fur et à mesure de mes travaux est la suivante : la reconnaissance de la fonction fait d'abord appel simplement à la compétence de lecteur de l'analyste (cf. section 3.1.1) : c'est à partir de cette compétence que je reconnais un élément d'une proposition comme son topique, une relation comme une élaboration, un objet textuel comme une définition ou une énumération. La seconde étape est de dégager les marques linguistiques dont la récurrence avec la fonction suggère une corrélation possible. La recherche procède alors de façon itérative (cf. Rebeyrolle & Péry-Woodley, 1998; section 5.2), en exploitant l'amorce fournie par les marqueurs identifiés pour rechercher – à l'aide, pour les travaux récents, d'outils d'analyse de texte – de nouvelles occurrences dans le corpus. L'étude de ces occurrences permet de valider les marqueurs identifiés et d'enrichir la caractérisation. Les récurrences envisagées peuvent être d'ordre lexical, syntaxique, typographique, ponctuationnel, dispositionnel. C'est-à-dire que par marqueur d'organisation textuelle j'entends non pas une forme spécifique qui coderait de façon univoque une fonction textuelle, mais une configuration de traits régulièrement associée à une fonction, même si chacun de ces traits individuellement peut en même temps être attaché à une ou plusieurs autres fonctions. Dans cette optique, la caractérisation linguistique d'aspects de l'organisation textuelle ne peut se faire qu'en corpus. Intervient alors la question de l'influence des paramètres de constitution du corpus sur les marqueurs identifiés. Mon hypothèse est que certains de ces paramètres, le domaine et le genre discursif en particulier, risquent d'influer sur les marqueurs présents, ce qui conduit à un sous-thème de recherche : l'identification de zones de stabilité et de variation dans la signalisation d'une fonction dans des corpus hétérogènes selon ces paramètres (cf. 5.2 et le prochain chapitre).

7.2.2 Un exemple : les marqueurs de l'énumération

La conception des marqueurs d'organisation textuelle formulée ci-dessus s'est élaborée au fur et à mesure des recherches présentées dans ce mémoire. On a vu qu'elle doit beaucoup sur le plan théorique aux travaux de J. Virbel et au modèle de l'Architecture textuelle, avec lesquels j'ai commencé à me familiariser à partir de 1994, et sur le plan méthodologique aux travaux de D. Biber sur les types de texte, dont j'ai commencé à prendre connaissance en 1990, et qui seront présentés dans le prochain chapitre. C'est la conception que j'essaie de mettre en œuvre, et à l'épreuve, dans mes travaux actuels, et en particulier dans un chantier collectif mis en route récemment sur l'énumération (Luc *et al.,* 1999; 2000).

Cette étude de l'énumération s'inscrit dans un projet plus vaste, intitulé *Structures spatio-linguistiques du texte : traitements formels et cognitifs*, coordonné par Claudine Garcia-Debanc dans le cadre de l'action concertée incitative *Cognitique* sur le thème de la cognition spatiale[67]. Un objectif majeur du projet, qui comporte par ailleurs un volet psycholinguistique et vise des applications informatiques, est d'éclairer la relation entre la composante lexico-syntaxique du texte et son inscription visuelle dans la page. C'est la spécificité de l'écrit, du texte comme objet visuel inscrit sur un support, que le projet dans

[67] Mes collaborateurs les plus proches dans ce projet sont Jacques Virbel, Christophe Luc, Mustapha Mojahid, de l'IRIT, Christian Jacquemin du LIMSI, et Claudine Garcia-Debanc du Laboratoire Jacques-Lordat. L'approche et les observations présentées dans ce qui suit se situent dans un travail commun.

son ensemble cherche à mieux cerner. S'appuyant sur les travaux de J. Goody (1977), W. Ong (1982), et G. Nunberg (1990), J. Virbel et ses collaborateurs (Virbel, 1989; Luc *et al*, 1999) mettent en évidence l'indépendance de l'écrit par rapport à l'oral, à l'encontre de l'approche qui voudrait y voir un simple codage de la langue parlée. Ainsi, de même que l'écrit n'offre pas de représentation satisfaisante des accents, rythmes, mélodies de la langue parlée, les propriétés morphologiques et dispositionnelles des caractères et des chaînes de caractères n'ont pas de contrepartie à l'oral. Il existe donc une spatialisation du langage, qui offre d'autres possibilités tant sur le plan expressif que sur le plan cognitif. Cette spatialisation a nécessité la construction de concepts originaux – page, double page, marge, paragraphe, titre, énumération –, plus récemment la fenêtre et les possibilités de multi-fenêtrage dans les textes conçus avec les nouvelles technologies, concepts indispensables à la maîtrise des activités de lecture et d'écriture.

Dans ce cadre la tâche qui m'incombe est l'identification à partir d'un corpus des régularités de mise en forme matérielle de certains objets textuels, en particulier de l'énumération, comme je l'avais fait avec E. Pascual pour la définition. L'énumération apparaît en effet comme un objet d'étude spécialement approprié pour aborder ce type de questionnement. D'abord, parce qu'elle fait le plus souvent intervenir un formatage spécifique, elle constitue un lieu privilégié d'examen de l'interaction entre marqueurs discursifs et marqueurs visuels. Par ailleurs, l'écrit autorise la construction d'énumérations d'une longueur et d'une complexité (plusieurs niveaux d'enchâssements) difficilement envisageables à l'oral. Enfin, pour revenir aux préoccupations générales de ce chapitre, l'énumération représente sur le plan fonctionnel une sorte de coup de force textuel dont l'effet est très spécifique : le fait de reconnaître dans un segment de texte une énumération, et de reconnaître par conséquent l'intention d'énumérer chez le scripteur, amène le lecteur à organiser les éléments constitutifs, quels qu'ils soient, en fonction d'un critère de co-énumérabilité. Il s'agit donc dans un premier temps de caractériser les marqueurs – visuels et discursifs – qui conduisent à l'identification d'une énumération et ceux qui ont trait à l'explicitation du critère de co-énumérabilité. Cette caractérisation linguistique conduira à l'établissement de classes d'énumérations, qu'il faudra ensuite examiner du point de vue de leur fonction et de leur intégration dans le texte. Les paragraphes qui suivent présentent les premiers résultats de cette étude.

• Corpus

Selon la conception courante, les items d'une énumération correspondent à des entités fonctionnellement et structurellement équivalentes, et sont réalisés par le biais d'une mise en forme identique pour chacun (numérotation, puces, espacements, etc.). Le premier recueil d'énumérations[68] (Virbel, 1999) a été mené en fonction d'un principe peu standard : à partir d'observations préliminaires qui suggéraient de nombreuses déviations par rapport à cette "norme", déviations perçues comme mieux à même de faire apparaître l'essentiel du fonctionnement de cet objet, J. Virbel a pris comme principe la recherche d'objets clairement identifiables comme des énumérations tout en n'étant pas conformes à la conception courante. De façon informelle, les traits suivants ont été retenus pour définir ces cas :

– items appartenant à des classes différentes de constituants syntaxiques ou textuels ;
– factorisation (des prépositions par exemple) irrégulière ; coordination ;
– organisateurs (cardinaux, expressions adverbiales) hétérogènes ;
– borne finale difficile à déterminer en raison du statut ambigu du dernier item ;
– dépendances complexes entre éléments d'énumérations enchâssées ou adjacentes ;
– non-correspondance entre nombre d'items annoncés et nombre effectif.

[68] Un second recueil est actuellement en cours, selon des principes plus classiques, et qui permettra la prise en compte du genre discursif.

Ce recueil a porté sur des textes variés – articles et ouvrages scientifiques, presse, pages de la Toile, modes d'emploi, … – en français et en anglais. Le corpus sur lequel se fonde la première étape de l'étude est un sous-ensemble des énumérations recueillies. Il est composé de 50 énumérations présentées avec leur co-texte immédiat (1 à 2 pages).

Quelques courts exemples extraits du corpus, à titre purement illustratif, permettront de montrer que nous ne nous sommes pas focalisés sur des "monstres", mais sur des objets parfaitement interprétables et identifiables en tant qu'énumérations :

(1) Est considérée comme "lecture savante", du point de vue fonctionnel, une pratique de lecture répondant aux critères suivants :

- c'est une lecture "qualifiée",

- qui se développe sur le temps long de la recherche scientifique,

- dans un parcours forcément individualisé,

- où l'écriture se combine à la lecture, souvent dans une perspective de publication. (article scientifique)

(2) les objectifs sont de deux ordres :

– définir des règles d'acceptabilité et implémenter les algorithmes permettant de décider automatiquement l'acceptabilité ou le rejet d'une forme de surface (pour les adverbes et connecteurs temporels).

– construire les représentations des propositions élémentaires pour les temps verbaux définis ci-dessus.

– à partir de ces représentations, calculer la sémantique des nouvelles formes obtenues par adjonction d'adverbes et de connecteurs. (article scientifique)

(3) Ce sac est réalisé avec des matières premières alimentaires conformes aux normes européennes

Il est :

-Étanche,

-Imperméable aux graisses,

-Il maintient au chaud,

-Supporte des températures de + 100°C À - 40°C. (Emballage de poulet rôti)

Hétérogénéité structurelle des items pour (1), problèmes de factorisation pour (1) et (3), non-correspondance entre nombre d'items annoncé et effectif pour (2), autant de "déviances" qui ne semblent mettre en cause ni le statut d'énumération de ces objets ni leur interprétabilité en tant que telles. Il paraît d'autant plus nécessaire d'inclure de telles occurrences dans un modèle inclusif qu'elles semblent à la fois fréquentes et fonctionnellement acceptables. Seules la description d'un corpus "généraliste" et des études psycholinguistiques pourront toutefois remplacer ces impressions par des observations fondées.

• Analyse de la structure des énumérations

Les résultats de cette analyse préliminaire seront présentés en deux temps : caractérisation de la structure des énumérations d'abord, puis identification des marqueurs associés aux éléments structurels.

On note tout d'abord, si l'on appelle énumération la suite d'items, que les énumérations sont, dans le corpus, toujours précédées par une phrase introductrice que nous appelons l'*amorce*. La caractérisation de la structure des énumérations concerne donc d'une part l'amorce, d'autre part la relation entre les items.

Les amorces se répartissent en deux types syntaxiquement distincts : soit il s'agit d'une phrase incomplète dont le constituant manquant est fourni par les items (*amorce incomplète*, ex. (3)) ; soit c'est une phrase complète qui annonce les items de l'énumération (*amorce complète*, ex. (1) et (2)).

La relation entre les items peut être de type *paradigmatique* ou *syntagmatique*. Dans le premier cas, typique des énumérations "classiques", les items sont des constituants syntaxiquement homogènes (phrases, syntagmes nominaux, infinitifs comme dans l'ex. (2), adjectifs comme les deux premiers items de l'ex. (3), etc.). Dans le second cas, comme dans l'exemple (1), chaque item est un constituant et l'ensemble des items forme une phrase. Ce regard syntaxique est cependant insuffisant, car les items peuvent être des unités de divers niveaux d'organisation : une énumération peut comporter des items phrastiques et être néanmoins syntagmatique dans le cas où chaque item constitue un argument et l'ensemble une argumentation (ex. (4) ci-dessous) :

> *(4) L'argumentation présentée ci-dessous peut être résumée ainsi : (1) Il existe aujourd'hui en France un courant actif et original en psychologie du travail et en ergonomie (ici confondues), mais (2) les représentants de ce courant sont dispersés dans des structures soit petites et isolées, soit importantes mais où leur situation est marginale ; donc (3) il serait souhaitable qu'une structure institutionnelle officialise, et par là encourage et consolide, un regroupement jusqu'ici tenté de manière uniquement volontariste et épisodique.*

Je reviendrai plus bas sur les problèmes d'analyse soulevés par ce fonctionnement à différents niveaux d'organisation textuelle. Pour terminer la caractérisation de la relation entre les items, il faut noter un certain nombre de cas "hybrides" dans lesquels certains items d'une énumération globalement paradigmatique sont dans une relation syntagmatique :

> *(5) In this paper I will defend what I shall call '(nonsolipsistic) conceptual role semantics'. This approach involves the following four claims:*
>
> > *(1) The meanings of linguistic expressions are determined by the contents of the concepts and thoughts they can be used to express;*
> >
> > *(2) the contents of thoughts are determined by their construction out of concepts; and*
> >
> > *(3) the contents of concepts are determined by their 'functional role' in a person's psychology, where*
> >
> > *(4) functional role is conceived nonsolipsistically as involving relations to things in the world, including things in the past and future.*

Dans cet exemple, les trois premiers items présentent un clair parallélisme sur le plan lexico-syntaxique (SN1 *are determined by* SN2), et se distinguent en cela du quatrième, qui est une proposition relative rattachée au troisième item.

Je présente ci-dessous un tableau résumant la distribution des différents types structuraux. Ces chiffres n'ont bien sûr aucune valeur statistique, étant donné le nombre limité de cas envisagés et le principe de recueil du corpus, mais ils permettent de dégager pour ces cas les associations privilégiées entre type d'amorce et type de relation entre items (tableau 7.1) :

Énumérations	Paradigmatiques	Syntagmatiques	Hybrides	*total*
amorce complète	18	1	6	25
amorce incompl.	15	8	2	25
total	33	9	8	50

Tableau 7.1 : Types d'amorces et relations dans le corpus.

On constate que même dans ce corpus "déviant" les énumérations paradigmatiques dominent. Elles sont associées indifféremment à des amorces complètes ou incomplètes, alors que les énumérations syntagmatiques semblent dans ce corpus préférentiellement associées à des amorces incomplètes.

Pour terminer cette section sur l'analyse structurelle des énumérations, et en référence au chapitre précédent sur les niveaux d'organisation textuelle, j'évoquerai brièvement quelques problèmes posés par la structure de l'objet énumération en relation avec d'autres niveaux de structuration. Nous avons dû nous rendre à l'évidence qu'aucune correspondance nécessaire n'existe, à l'intérieur d'une énumération, entre la structure syntaxique ou la structuration en phrases ponctuationnelles et paragraphes d'une part et la structuration en items de l'autre. Ainsi, le premier item peut être inclus dans la même unité (proposition, phrase, paragraphe) que l'amorce et suivi d'items qui constituent chacun une ou plusieurs unités. Par ailleurs, la correspondance entre la structuration en items et la structure rhétorique n'est pas directe : dans l'exemple (5), le parallélisme lexico-syntaxique des trois premiers items recouvre une structure hiérarchisée puisque le premier item introduit les deux termes, *concepts* et *thoughts*, qui seront développés dans les items 2 et 3 (auquel est rattaché l'item 4). Ces deux items coordonnés (2 et 3), plus l'item 4 enchâssé syntaxiquement dans l'item 3, constituent selon les termes de la RST le satellite d'élaboration de l'item 1. On a donc une structure qui défie sur plusieurs plans à la fois la conception classique qui voudrait des items structurellement et fonctionnellement équivalents.

• Identification des traits pouvant constituer des marqueurs

Le mot énumération évoque une structure visuelle spécifique, dont les traits marquants sont le retour à la ligne entre les items et l'utilisation d'objets typographiques tels le tiret ou la puce devant chaque item. Cette structure visuelle n'est cependant pas nécessaire pour qu'il y ait énumération, on va le voir. On peut se demander si elle est par ailleurs suffisante, c'est-à-dire si il y a énumération dès qu'il y a suite d'items, même en l'absence d'amorce. La question ne se pose pas par rapport au corpus considéré ici, dans la mesure où il ne contient pas d'énumération sans amorce. Elle se pose en revanche pour la constitution d'autres corpus : le "style énumération" est par exemple très utilisé sur la Toile, sans qu'il y ait nécessairement une amorce. Ce que recouvre cette question, c'est en fait celle de la nécessité de la présence explicite du critère de co-énumérabilité, fourni en général par l'amorce. Je ne pousserai pas plus loin pour l'instant la considération de cette question, pour revenir à l'identification des marqueurs régulièrement associés aux énumérations étudiées, et à leur interaction. Ce qui va m'intéresser tout particulièrement, c'est la relation entre marqueurs lexicaux et lexico-syntaxiques d'une part, et visuels de l'autre, dans les configurations qui conduisent à l'identification d'une énumération.

Pour les deux types d'amorce, nous paraissent significatifs à ce stade les marqueurs :

– dispositionnels : retour à la ligne et blanc vertical entre amorce et premier item ;

– ponctuationnels : deux-points en fin d'amorce ;

Les amorces incomplètes sont par définition marquées sur le plan syntaxique par leur incomplétude. Les amorces complètes présentent des régularités :

– lexico-syntaxiques : le groupe nominal désignant l'hyperonyme (critère de co-énumérabilité) a pour déterminant un adjectif numéral cardinal ou un adjectif indéfini comme *quelques, certains* (ex. (2) et (5)) ;

- lexicales : présence d'une expression à valeur déictique tel que *suivant, ci-dessous* (ex. (1) et (5)). L'hyperonyme lui-même peut difficilement être pris en compte dans la mesure où il appartient à une classe lexicale ouverte, même si certains termes qu'on pourrait appeler "métadiscursifs" (*remarques, réponses, observations, claims*) forment une classe qu'on pourrait tenter de définir en extension.

Il est certain qu'aucun de ces marqueurs n'est à lui seul suffisant pour marquer l'amorce d'une énumération. C'est donc en configuration, et lorsqu'ils sont suivis des marqueurs de l'énumération à proprement parler, qu'ils peuvent signaler une amorce.

En ce qui concerne l'énumération d'items, les traits que nous avons relevés comme marqueurs potentiels sont :

– dispositionnels : espaces horizontaux, espaces verticaux entre les items ;

– ponctuationnels : ponctuation en fin d'item (souvent *;*) ;

– lexico-syntaxiques : parallélisme des énumérations paradigmatiques (infinitifs en (2), adjectifs pour les 2 premiers items de (3), SN1 *are determined by* SN2 en (5)) ;

– lexicaux/typographiques : numérotation ((4) et (5)), marques typographiques ((1), (2) et (3)), organisateurs (*Premièrement*, *Ensuite, A first issue, A second issue*)

Il s'agit donc pour l'instant de traits associés de façon récurrente sinon régulière avec les énumérations du corpus. L'identification de marqueurs à partir de ces traits suppose une étude des distributions relatives des traits, et de leurs poids respectifs, de manière à faire apparaître des configurations elles aussi récurrentes. Ce sont ces configurations qui pourront être considérées comme des marqueurs. Nous avons engagé la réflexion en ce qui concerne l'interrelation entre marques "discursives" et "visuelles". Prenons par exemple la présence d'expressions "à valeur déictique" comme trait caractéristique de l'amorce ; je dirais que c'est le cas des exemples (1), (4) et (5). Dans ces exemples, les éléments lexicaux concernés sont "suivants", "ainsi", "following". Il est clair que ces mots en eux-mêmes ne sont pas nécessairement des déictiques. Dans les amorces d'énumérations, c'est sans doute le deux-points qui clôt l'amorce, ainsi que peut-être l'identification de marqueurs d'énumération (numérotation pour (4)) dans ce qui suit, qui force une lecture déictique. Il est donc impossible de traiter séparément marqueurs visuels et marqueurs "discursifs". La même constatation s'impose en ce qui concerne les organisateurs des items : il peut s'agir de marques typographiques (puces, tirets), de numérotation, d'adverbes ordinaux (*premièrement, deuxièmement*), d'autres adverbes dénotant un ordre (*d'abord, ensuite*), de syntagmes nominaux comprenant un adjectif ordinal (*a first issue, a second issue*). Le choix de l'une ou de l'autre de ces formulations n'est certainement pas indifférent en termes d'adéquation à la situation, mais elles semblent être fonctionnellement équivalentes[69] pour ce qui est de la signalisation d'une énumération. On voit donc clairement illustré ici le principe d'équivalence fonctionnelle entre marques discursives (ici lexicales) et marques visuelles (ici typographiques) évoqué dans la présentation du modèle de l'architecture textuelle en 5.1.2.

Ce principe d'équivalence fonctionnelle a un corollaire : on peut s'attendre à ce qu'un objet textuel faisant l'objet d'une claire signalisation discursive soit moins marqué visuellement et vice versa, c'est-à-dire à ce qu'il y ait une relation inverse entre la densité de ces deux types de marques. Quelques premières observations vont en effet dans ce sens. Nous avons commencé par examiner les cas extrêmes à partir de deux critères : l'absence totale de marques dispositionnelles (énumérations en continu comme (4)), l'abondance d'organisateurs lexicaux. Les énumérations sans signalisation dispositionnelle présentent en effet d'une façon générale une forte structuration lexicale ; les énumérations où les organisateurs lexicaux sont nombreux ont pour la plupart peu de marques typographiques et dispositionnelles. Nous cherchons à présent à mettre au point des techniques plus précises pour mesurer la densité des marques en leur attribuant des poids. Comme on pouvait s'y attendre, on observe la plus faible densité de marques visuelles dans les énumérations paradigmatiques dont les items sont à la fois semblables sur le plan lexico-syntaxique (parallélisme) et précédés par des organisateurs lexicaux. Cette pondération, qui est actuellement à l'état d'ébauche, peut jouer un rôle important dans les applications computationnelles qui vont être évoquées dans la prochaine section. Pour ces applications, il sera nécessaire de mettre au point des méthodes permettant de préciser le rôle des différents éléments à l'intérieur des configurations.

[69] Il ne semble même pas y avoir de corrélation nette entre numérotation et non-permutabilité des items, et encore moins entre type d'organisateur et instruction de lecture en ce qui concerne la relation logique (ET ou OU) entre les items.

7.3 Marqueurs et applications en TAL

Ce chapitre oppose deux approches dans les travaux sur corpus autour des marqueurs. J'ai évoqué l'approche "classique" qui consiste à partir de marqueurs pré-définis pour examiner leur fonctionnement dans des textes. J'ai surtout développé l'approche inverse, qu'on pourrait appeler "de découverte", qui prend comme point de départ une fonction ou un objet textuel, identifie les traits qui lui sont régulièrement associés dans des textes, et définit comme marqueurs des configurations spécifiques de traits[70]. Ces deux approches se prêtent à des applications en traitement automatique des langues, que je vais maintenant préciser par rapport à des travaux récents ou en cours.

7.3.1 À partir des marqueurs : typage de textes, recherche d'énoncés importants

Dans le cadre de l'approche "classique", je citerai brièvement deux études qui, bien qu'elles se situent très en amont d'applications proprement dites, sont envisagées dans le contexte du traitement automatique. L'étude des introducteurs de cadres de discours (section 6.1), avant tout concernée par des problèmes d'ordre théorique, l'articulation des niveaux d'organisation textuelle et la mise en relation de différents modèles, a pris un tour plus pratique dans un projet de maîtrise de Sciences du Langage[71]. L'idée de départ était que certaines classes d'introducteurs étaient suffisamment liées à des genres discursifs pour pouvoir contribuer à l'identification du genre d'un texte. Il s'agissait donc d'évaluer l'utilisabilité des introducteurs pour le typage automatique des textes. Ce typage a des enjeux importants pour le TAL dans la mesure où de nombreux aspects du fonctionnement linguistique et textuel sont réguliers à l'intérieur d'un type de texte. Je décrirai dans le chapitre 8 la méthodologie élaborée par D. Biber (1988) pour regrouper des textes qui partagent certains traits linguistiques. Un pré-traitement exploitant cette méthodologie pour attribuer un profil à un texte à partir de cooccurrences de traits analysables de façon automatique constituerait donc une avancée considérable (cf. Illouz *et al.,* 1999; Habert *et al.,* 2000). L'étude des introducteurs de cadres représente une tentative d'extension de cette méthode. Le point de départ en était l'hypothèse d'une corrélation entre types d'univers – et d'expressions introductrices – et genre discursif (ou plus exactement configurations genre/domaine, cf. prochain chapitre). L'étude des introducteurs dans un corpus diversifié a confirmé cette hypothèse. Le problème de traitement reste cependant complexe, ces expressions étant pour une bonne part des adverbes et des syntagmes prépositionnels, c'est-à-dire qu'elles sont au moins partiellement constituées de lexèmes de classes ouvertes.

Le second projet, également mis en œuvre par des étudiantes de maîtrise de Sciences du Langage sous ma direction[72], a trait au fonctionnement en discours d'une structure examinée dans la section 2.2.2 sur le plan de la structure thématique : la construction clivée. Ici, l'idée de départ, issue des observations de Borkin sur la saillance de la phrase clivée tout entière – et non seulement de l'élément clivé – dans le discours environnant, est que les clivées pourraient signaler des énoncés importants, et ainsi constituer un marqueur utilisable pour certaines approches du résumé automatique de textes. La première tâche par rapport à cet objectif est de préciser le fonctionnement des clivées, sur le plan interne d'abord – catégorie syntaxique de l'élément clivé, statut présuppositionnel de la "relative" –, en discours ensuite, par la comparaison avec la non-clivée correspondante et la mise en relation avec différents aspects de l'organisation du texte.

---

[70] On retrouve cette approche "de découverte" dans l'extraction de relations sémantiques (cf. Condamines & Rebeyrolles, à paraître; Hearst, 1998; Morin, 1999).

[71] Mémoire réalisé par Mai Ho Dac, juin 1999.

[72] Deux mémoires, l'un terminé en septembre 1999 (Emmanuelle Labbe), le second en cours (Hélène Corvisier).

7.3.2 Des fonctions aux marqueurs : génération de texte, extraction d'information

Ce qu'on peut appeler la "modélisation" d'un objet textuel – définition, énumération -, c'est-à-dire l'identification des configurations de traits qui en signalent la présence, concerne deux grands domaines d'application : la génération de texte et l'extraction d'information à partir de documents textuels. En génération, il s'agit de construire la grammaire de certains objets textuels, de manière à générer des textes "réalistes" ; en extraction on cherche à repérer automatiquement ces objets dans les textes. Les deux applications nécessitent l'approche "de découverte" décrite dans ce qui précède puisqu'on ignore *a priori* ce qui signale ces objets. La prise en compte des marques visuelles au même titre que les marques "discursives", défendue sur le plan théorique dans le cadre du modèle de l'architecture textuelle, apparaît comme une évidence en relation avec ces visées applicatives. L'idée de pondérer les marques (7.2.2) trouve aussi toute sa justification par rapport à ces applications, dans la mesure où elle devrait permettre de déterminer un seuil de signalisation minimale pour ces objets.

L'exploitation de cette approche en génération est actuellement poursuivie par Christophe Luc (Luc, 1998; Luc *et al.,*1999; 2000). L'extraction me concerne plus directement, bien que là encore mes propres recherches se focalisent sur la modélisation préalable. Je citerai pour en illustrer le potentiel applicatif le travail de C. Jacquemin et C. Bush (à paraître), mené conjointement au mien dans le cadre du projet sur les structures spatio-linguistiques du texte présenté en 7.2.2. L'objectif visé est l'extraction de données lexicales spécifiques, les *entités nommées,* c'est-à-dire les noms propres, dans les pages de la Toile. Pour la constitution de ressources lexicales à partir de corpus, les noms propres représentent une problème particulier de par leur renouvellement permanent. C. Jacquemin et C. Bush proposent une technique d'extraction automatique de ces noms qui permet en outre de leur attacher un nom de classe, un *genus,* en exploitant la forme des énumérations. Ils s'intéressent à des énumérations telles que (6), qui listent des noms d'entités dont le type est donné dans l'amorce :

(6) The following international organizations are collaborating on the Project:

> International Commission on Non-Ionizing Radiation Protection (ICNIRP)

> International Agency for Research on Cancer (IARC)

> United Nations Environment Prgramme (UNEP)

La technique permet d'extraire les trois entités nommées tout en leur attachant le *genus* "international organizations".

L'approche de la notion de marqueur développée dans ce chapitre peut se résumer comme suit :

– les marqueurs sont des configurations de traits régulièrement associées à une fonction ou à un objet textuel ;

– ces configurations de traits caractéristiques regroupent des traits lexico-syntaxiques, typographiques, dispositionnels, ponctuationnels ;

– l'équivalence fonctionnelle entre des traits "discursifs" et des traits "visuels" débouche sur l'affirmation du statut linguistique de ces derniers, le texte écrit étant un objet visuel ;

– les marqueurs ne sont pas donnés mais doivent faire l'objet d'une procédure de découverte en corpus.

Ce dernier point entraîne des conséquences qui seront reprises et développées dans le chapitre suivant : l'identification des marqueurs ne se fait pas en langue mais en corpus, c'est-à-dire dans des recueils de textes qui sont les traces de discours, liés à des situations, à des interlocuteurs, à des visées, à des domaines particuliers. L'impact potentiel sur les

marqueurs identifiés des paramètres qui contribuent à distinguer des genres discursifs doit maintenant être examiné.

Chapitre 8

Domaines, types et genres dans les corpus : penser la variation

Les corpus jouent un rôle central dans l'ensemble des travaux présentés dans ce mémoire. Ce choix est sans doute en partie lié à des facteurs "historiques" : ma carrière de linguiste a en effet débuté au sein de problématiques dialectologiques (locutions prépositionnelles dans l'anglais de l'ouest de l'Irlande), puis psycholinguistiques (dysphasie), qui ne pouvaient s'envisager que sur corpus. Par ailleurs, les corpus faisaient un retour en force dans les études sur l'anglais au début des années 80 (cf. Aarts & Meijs, 1984, *inter alia*). Des facteurs plus directement liés à l'objet et aux objectifs des travaux présentés ici – la focalisation sur le texte, l'optique fonctionnaliste qui découle de ce choix, et la démarche dominante qui va des fonctions vers les marqueurs – rendent indispensable le recours aux corpus. J'ai opté initialement pour des corpus dits "expérimentaux", c'est-à-dire composés de textes rédigés spécifiquement pour les besoins de l'étude en fonction de consignes précises (chap. 2, 3, 4), pour m'orienter ensuite vers des corpus de textes préexistants composés essentiellement de ce qu'on pourrait appeler des "documents de travail", des textes dont on peut cerner le rôle dans un univers professionnel donné. Ce choix s'est trouvé renforcé par l'inscription de mes recherches dans la théorie des sous-langages de Z. Harris (cf. 5.1). D'emblée, la notion de variation dans les réalisations langagières a été présente dans mes recherches, qui envisagent le corpus non pas comme une simple source de données pour l'étude de la langue mais comme un ensemble d'échantillons de populations dont je tente de décrire certains fonctionnements. J'ai donc été amenée à m'interroger, avec mes collègues de UMIST d'abord, puis de l'opération Sémantique et Corpus à Toulouse, sur la relation entre les données de départ et les résultats d'analyse de corpus, sur les façons de penser la variation dans les corpus, sur les problèmes propres aux études sur corpus au niveau du texte (Péry-Woodley, 1994; 1995; Condamines *et al.,*1999). Je reprends dans ce dernier chapitre des éléments de cette réflexion, dont le lecteur apercevra vite combien elle doit aux travaux de Douglas Biber (1988; 1989; 1990; 1992; 1993a; 1993b; 1993c; 1995; 1998; Biber & Finegan, 1993), ainsi qu'à un dialogue de longue date avec Benoît Habert (Habert, 1985; 1995; 1998; Habert & Jacquemin, 1993;.Habert & Salem, 1995; Habert *et al.*, 1995; 1997; 1998; 2000).

La première section fait le lien avec les travaux présentés dans ce mémoire en évoquant spécifiquement le niveau textuel, Les sections suivantes traitent de façon plus générale de la variation : sa pertinence en linguistique comme en TAl, sa modélisation.

8.1 Variation et niveau textuel

8.1.1 Marqueurs d'organisation textuelle

Des travaux qui prennent en compte la variation dans les réalisations langagières seront cités dans la section suivante, concernant les niveaux lexical, morpho-syntaxique et syntaxique. Cette prise en compte me paraît tout aussi nécessaire pour le niveau du texte : dans une perspective "microscopique", on constate un impact de paramètres liés au genre sur le fonctionnement de marqueurs spécifiques. Ce sera mon premier exemple. Il est issu des travaux de J-M. Adam et F. Revaz sur les marqueurs de structuration de texte, où sont distingués les fonctionnements de connecteurs comme *enfin*, *alors* selon le type de texte ou la séquence textuelle où ils apparaissent (Adam & Revaz, 1989). Les auteurs font état d'une surdétermination globale liée à la mise en texte, qui entraîne qu'un marqueur comme *enfin* sera tour à tour :

– un marqueur temporel dans une séquence narrative ;
– un marqueur signalant et soulignant le dernier élément dans une énumération ;
– un marqueur de reformulation récapitulative dans une description ;
– un marqueur de reprise énonciative dans l'interaction spontanée.

De même, *alors* change de fonction et de valeur selon qu'il se trouve dans une séquence narrative ou dans une séquence argumentative. Bien qu'elles se situent dans un cadre théorique et méthodologique différent du mien (typologie en termes rhétoriques, notion de séquence, focalisation sur les marqueurs lexicaux), ces observations me paraissent intéressantes à plus d'un titre : elles présentent des exemples très nets de variabilité fonctionnelle d'une série de marqueurs en lien avec la visée discursive d'un segment ; elles posent le problème de l'hétérogénéité interne des textes, problème également abordé par R. Grishman et R. Kittredge (1986, cf. 8.3.2), et par D. Biber et E. Finegan dans une étude des différences linguistiques systématiques associées aux variations de visée discursive des différentes sections d'articles de recherche expérimentale (1994).

Dans une perspective "macroscopique", on peut s'attendre à ce que les marqueurs associés à une fonction ou à un objet varient suivant des paramètres situationnels. Le second exemple provient de mes propres travaux, en collaboration avec Josette Rebeyrolle (Péry-Woodley & Rebeyrolle, 1998; Rebeyrolle & Péry-Woodley, 1998; cf. 5.2.3). Nous avons constaté dans certains corpus des préférences régulières dans la formulation des définitions, préférences que nous avons pu associer à des caractéristiques extra-linguistiques de ces textes. Le corpus de géomorphologie se démarque ainsi des autres sous-corpus (manuels de logiciels, guide pour le développement de projets en génie logiciel) par l'emploi exclusif de définitions où la relation d'hyperonymie ne se fait que par l'assertion (et non par le biais de formes nominales complexes). Ce choix peut être interprété comme lié à un objectif central pour ce manuel didactique : introduire les termes du domaine et construire la taxinomie des objets. Là où ce même corpus privilégie les modifieurs (exprimant les *differentiae)* de type adjectif ou participe passé, les modifieurs des définitions du domaine logiciel sont principalement des subordonnées relatives. Dans ces modes d'emploi, le modifieur est régulièrement la partie la plus développée des définitions, qui, on l'a vu (5.1.1), peuvent fonctionner comme des consignes précisément grâce à ces éléments de description fonctionnelle[73].

[73] Ces travaux préliminaires, qui portaient sur des corpus limités, n'ont pas donné lieu à une quantification, mais seulement à l'observation de tendances. Ils nous ont permis d'expérimenter une démarche que Josette Rebeyrolle poursuit maintenant de manière plus systématique et sur de plus gros corpus dans le cadre de sa thèse de doctorat (sous la direction de A. Borillo ; cf Rebeyrolle, 2000).

La démarche adoptée part donc d'une hypothèse de variabilité, pour suivre les étapes suivantes :

– recherche des configurations de marques dans des sous-corpus homogènes en termes de domaine-genre discursif ;

– identification des zones de variabilités et d'invariance à partir de la comparaison des configurations caractéristiques des sous-corpus.

Cette démarche correspond à celle prônée par J. McNaught (1993) pour le TAL : plutôt qu'une spécialisation après-coup de systèmes généralistes fondés sur des grammaires et des lexiques généraux, il recommande une démarche ascendante, qui construirait des niveaux de généralisation à partir de descriptions de systèmes linguistiques spécialisés. Elle se démarque de l'approche de Hearst (1992, 1998), ou de celle de Berri *et al.* (1996) ou Cartier (1997), qui font explicitement ou implicitement l'hypothèse de l'invariabilité des marqueurs (de relations sémantiques principalement).

Cette méthode, utilisée à petite échelle pour l'étude de la définition, présente un intérêt certain pour le nouveau chantier sur l'énumération (cf. chapitre 7), où les formulations vont du tout discursif au tout visuel. Le lien entre domaine-genre et mise en forme matérielle des textes a jusqu'à présent peu été étudié. Les corpus que nous avons commencé à examiner laissent déjà apercevoir des différences marquées, avec en particulier le caractère très visuel des pages de la Toile. Le fait d'envisager le texte en tant qu'objet visuel inscrit sur un support physique m'amènera à ajouter le type de support aux paramètres situationnels proposés par D. Biber pour modéliser le genre discursif, qui seront exposés en 8.3.

8.1.2 Corpus et texte : quelques problèmes

Avant d'ouvrir ce chapitre à des questions d'ordre plus général concernant d'abord la prise en compte de la variation dans l'analyse linguistique et le TAL, puis des propositions pour sa modélisation, j'évoquerai trois problèmes soulevés par l'étude en corpus de l'unité texte. Ils concernent le principe d'exhaustivité, les corpus bilingues, et la prise en compte de cette unité fonctionnelle dans la constitution de corpus.

Le principe d'exhaustivité, ou "principle of total accountability" est défini par G. Leech (1992) dans son "paradigme de la recherche empirique en linguistique sur corpus informatisés" :

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

Ce principe s'entend bien sûr dans un cadre théorique qui constitue en observables un corpus construit en fonction d'une problématique spécifique. Il représente alors un atout majeur de la linguistique de corpus, mais peut poser problème lorsque les règles qui sous-tendent les fonctionnements à l'étude sont en partie *acquises* et en partie *apprises*. On *acquiert* sa langue maternelle mais on *apprend* à l'écrire, on *apprend* à rédiger des textes, on *apprend* à mettre en forme une énumération. Pour tout ce qui relève de l'acquisition, le corpus libère en effet de l'attitude normative qui peut brider l'introspection. Les objets textuels dont je cherche à caractériser la signalisation relèvent quant à eux à la fois de l'acquisition de la langue et de l'apprentissage de la mise en texte. Pour illustrer, je reproduis ci-dessous un exemple d'énumération cité dans le chapitre précédent :

> *(1) Ce sac est réalisé avec des matières premières alimentaires conformes aux normes européennes*
>
> *Il est :*
>
> *-Étanche,*
>
> *-Imperméable aux graisses,*

-Il maintient au chaud,

-Supporte des températures de + 100°C À - 40°C. (Emballage de poulet rôti)

Faut-il faire entrer une énumération comme (1) dans le modèle ou l'exclure comme mal formée ? La seconde solution est en fait difficile à mettre en œuvre dans la mesure où elle suppose connues les règles que l'on cherche précisément à induire. Le niveau du texte pose ainsi d'une façon spécifique le problème de la prise en compte de l'"attesté impossible" soulevé par B. Habert (2000) : "le recours renouvelé aux corpus confronte effectivement à des énoncés que l'on juge *a priori* impossibles", C'est l'articulation entre les régularités observées en corpus et les règles postulées qui est en jeu. B. Habert propose une classification de ces "attestés impossibles" fondée sur une interprétation de leur origine: a) erreurs, lapsus ; b) transgression délibérée et donc chargée de sens ; variation interne à la langue ; évolution des règles. Ce qui est propre au niveau du texte, cependant, c'est l'articulation entre mise en œuvre d'une compétence linguistique acquise et application de règles de mise en texte apprises. Le type d'application détermine certaines décisions : on pourra choisir d'exclure certaines occurrences dans le contexte d'une modélisation à partir de corpus pour la génération automatique de textes ; si l'on modélise au contraire pour l'extraction, il faudra opter pour la couverture la plus large. Mais on a quitté la linguistique pour l'ingénierie linguistique.

Ma deuxième remarque concerne la constitution de corpus bilingues pour la recherche au niveau du texte. Le problème est celui de la comparabilité de textes produits dans des contextes culturels différents. Un exemple : je me propose d'étendre à l'anglais l'étude du fonctionnement textuel de la clivée, évoquée au chapitre 7. Comme le passif, la clivée constitue un objet intéressant pour l'approche contrastive dans la mesure où on a pour les deux langues la même relation syntaxique entre phrase active/phrase passive d'une part, phrase canonique/phrase clivée d'autre part. J'ai montré pour le passif, à partir d'un corpus "expérimental", que cette correspondance syntaxique pouvait conduire à une sorte de "faux ami textuel" pour des anglophones écrivant en français, dans la mesure où le passif ne "fait" pas la même chose sur le plan du texte en français et en anglais (section 2.2.2.). Le recueil d'un corpus bilingue permettant une étude contrastive sur le fonctionnement de la clivée en discours n'est pas simple : les "corpus bilingues" dont une langue est la traduction de l'autre (Hansard) sont peu fiables en raison du risque de calque au niveau de la signalisation textuelle. La presse quotidienne, facile à obtenir dans les deux langues, pose des problèmes d'appariement : le "paysage" de la presse britannique ne se superpose pas aisément à celui de la presse française par exemple. Quels paramètres prendre en compte pour cette comparaison ?

Finalement, un problème pour la constitution de corpus de référence : comment concilier une bonne pratique d'échantillonnage et le respect de l'unité texte ? Les corpus pour l'anglais, y compris le récent British National Corpus, sont faits d'échantillons pris à divers moments du texte pour ne pas privilégier certaines formes, associées par exemple aux débuts de texte. La conséquence est bien évidemment que l'organisation textuelle n'y est plus accessible. Pour les linguistes du texte, un corpus idéal serait peut-être fait d'échantillons correctement sélectionnés et calibrés, mais dont certains tout au moins donneraient accès au texte intégral.

8.2 L'insoutenable variabilité des corpus

La disponibilité croissante de bases de textes sur support électronique accompagnées d'interfaces de recherche conduit un nombre également croissant de linguistes de toutes disciplines, et ne se réclamant pas nécessairement des linguistiques de corpus, à travailler sur des exemples extraits de textes plutôt que fabriqués. On utilise donc la base Frantext, ou les grands quotidiens disponibles sur CD-Rom, comme source d'exemples pour des études dans tous les domaines de la linguistique. Beaucoup de ces études, malgré ce changement

méthodologique, restent toutefois centrées sur la *langue*, et traitent le recours au corpus avant tout comme une manière commode de se constituer des données pour aborder la *langue*. Dans l'univers du TAL, la possibilité de recueillir et de traiter des milliards de mots pousse parfois les chercheurs à accepter sans sourciller la devise "more data is better data". Il me semble important de s'interroger sur le statut de ces données, sur leur adéquation à l'objet d'étude, sur leur représentativité. Les questions sont complexes : le fait que des milliers d'articles du Monde deviennent accessibles grâce à la collection de CD-Roms regroupant plusieurs années de ce quotidien ne change rien quant à leur aptitude à représenter la diversité des potentialités du français. Ils sont tous issus du même journal. En revanche, si l'on désire envisager Le Monde comme représentant un usage non marqué, une "langue générale", il faudra s'interroger non seulement sur sa représentativité mais aussi sur son homogénéité (les rubriques sportives y parlent-elles la même langue que les éditoriaux ?). Il y a lieu de se demander s'il existe des domaines de la linguistique descriptive pour lesquels l'origine des données n'importe pas. La question se pose *a fortiori* en TAL, puisque sont visés sur la base d'observations à partir de corpus des traitements informatisés dont la qualité et l'efficacité va dépendre de l'adéquation de la modélisation initiale aux textes à traiter.

L'objectif principal de mon article "Quels corpus pour quels traitements automatiques ?" (Péry-Woodley, 1995), paru dans un numéro spécial de la revue TAL intitulé *Traitements probabilistes et corpus,* était précisément d'attirer sur ces questions l'attention de la communauté alors en train de se constituer autour des linguistiques de corpus dans le monde francophone. Ma formation et mes orientations de recherche me donnent en effet une conscience aiguë du fait que tout corpus est constitué de textes, c'est-à-dire de traces de discours ancrés dans des situations de communication particulières qui influent de diverses façons sur les réalisations langagières. Deux grands facteurs conduisent souvent à occulter la variation dans les travaux sur corpus : d'une part les progrès technologiques, qui permettent de constituer et de traiter des masses énormes de texte, dont on se contente de supposer qu'elle doivent contenir "un peu de tout" ; de l'autre la difficulté de modéliser la variation d'une façon utile pour la construction et l'exploitation de corpus réutilisables. Prôner la prise en compte de la variation passe donc par deux types d'arguments : d'abord démontrer sa pertinence dans l'exploitation des résultats d'analyses de corpus en linguistique et en TAL, les risques scientifiques qu'on encourt en l'ignorant, les avantages pratiques potentiels (pour le TAL) ; ensuite proposer des démarches théoriques – modélisation des facteurs de variation – et méthodologiques – constitution de corpus – qui la rendent possible. Ces deux axes seront abordés tour à tour dans cette section et la suivante (8.3).

8.2.1 La variation pèse dans la description et les traitements

Je m'inscris ici, inévitablement, dans le débat toujours vif sur les différents types de données pour la linguistique. J'éviterai cependant d'envisager la question en termes dichotomiques : données introspectives contre corpus. Il semble plus fertile de mener, comme l'a fait de façon exemplaire P. Corbin (1980) pour l'introspection, une réflexion sur la nature des données obtenues par ces méthodes et sur les pratiques descriptives appropriées à chacune, de manière à en voir les complémentarités. Ainsi que le note C. Fillmore pour clore sa caricature des deux sortes de linguistes – "the armchair linguist" et "the corpus linguist", "the two kinds of linguists, wherever possible, should exist in the same body" (Fillmore, 1991:35). Cette réflexion, qui est en cours dans la communauté des linguistiques de corpus[74], est également nécessaire pour fonder les pratiques de recueil des données, et de constitution de ressources réutilisables. Centrée particulièrement dans cette section sur le

[74] Cf. l'atelier thématique *Corpus et TAL : Pour une réflexion méthodologique*, organisé et animé par C. Fabre, A. Condamines et moi-même dans le cadre du colloque TALN'99 (Condamines *et al.,* 1999).

paradoxe entre les visées généralisatrices inhérentes au travail de linguiste et la nature accidentelle et spécifique des discours dont les traces constituent les corpus, elle va passer par la considération d'exemples qui illustrent le poids de la variation à différents niveaux d'analyse linguistique. En exergue, je citerai la mise en garde qu'adresse D. Biber aux linguistes sur corpus tentés de généraliser leurs observations :

> Global generalizations are often not accurate at all, because there is no adequate overall linguistic characterization of the entire language; rather, there are marked linguistic differences across registers (or sublanguages). Thus a complete description of the language often entails a composite analysis of features as they function in various registers. (1993b:220).

C'est à D. Biber aussi que seront empruntés plusieurs des exemples ci-dessous, qui intéressent la linguistique et le TAL, aux niveaux lexical, morpho-syntaxique et syntaxique.

Niveau lexical

Le recueil de données pour la lexicographie a sans doute été un des domaines pionniers de l'approche corpus, en particulier avec le projet COBUILD en Grande-Bretagne (Sinclair, 1987; Sinclair *et al.,*1987). Au début des années 90, J. Sinclair (1991), comme K. Church et R. Mercer dans leur introduction au numéro de *Computational Linguistics* consacré aux corpus (1993), défendent les très gros corpus en lexicographie pour des raisons de coût et de faisabilité. Ces derniers montrent à l'aide d'exemples précis – occurrences de *imaginable*, collocations de *strong* – comment les gros corpus fournissent une abondance de données là où les petits corpus, (Brown, Bible), n'en donnent pas du tout (1993:18-19). On peut toutefois s'interroger sur la valeur de ces données lorsqu'il s'agit non seulement de recueillir des occurrences mais de délimiter de façon fine les usages et les préférences sémantiques ou syntaxiques pour la constitution d'une entrée de dictionnaire. D. Biber aborde également la lexicographie, mais avec une prise de position différente :

> Similar to the patterns for grammatical structures, for many words there is no general pattern of use that holds across the whole language; rather different word senses and collocational patterns are strongly preferred in different registers. (1993b:226).

Il illustre cette affirmation par l'analyse d'un groupe d'adjectifs ayant trait à la certitude en anglais (*certain*, *sure* et *definite*), d'abord sur le plan des fréquences, ensuite en examinant certaines collocations[75]. La comparaison se fait selon deux axes : entre écrit et oral ; entre textes de fiction et articles en sciences sociales (tableau 8.1).

	Oral	**Ecrit** Fiction	Articles en sciences sociales
Fréquence (rang)	*sure* *certain* *definite* fréquences beaucoup plus élevées qu'à l'écrit (surtout *sure*)	*sure* *certain* *definite*	*certain* *definite*
Usages préférés		*sure* = certitude *pron+be+certain*	*certain* = imprécision

Tableau 8.1 : Utilisation de certain/sure/definite *selon le corpus (d'après D. Biber).*

[75] Cette analyse est menée dans un corpus écrit, le corpus London-Lund, composé de textes répartis en dix "catégories" différentes, et dans un corpus oral, le corpus Longman-Lancaster, qui comprend six "catégories" principales.

D. Biber signale d'importantes différences de fréquences (calculées par million de mots) : *sure* s'avère considérablement plus fréquent à l'oral, où les occurrences des trois mots confondus dépassent de beaucoup leur nombre dans le corpus écrit. Dans celui-ci, les usages sont très contrastés : en sciences sociales, *sure* est rare, *certain* est fréquent, et *definite* aussi par rapport à sa faible fréquence dans le corpus. Le schéma s'inverse pour la fiction, où *certain* cède la première place à *sure*, et où *definite* disparaît presque. Outre ces fréquences brutes, l'étude des collocations met en évidence des différences d'utilisation de ces mots en contexte dans les deux sous-corpus. Là aussi les observations se fondent sur des fréquences normalisées (par million de mots) ; il en ressort que *certain* est utilisé en sciences sociales principalement comme déterminant (*a certain kind of..., in certain cases…*), et donc pour exprimer l'incertitude (ou tout au moins l'imprécision), alors qu'on le trouve beaucoup dans des constructions du type *pronom* + *be* + *certain* dans les textes de fiction, constructions totalement absentes du corpus de sciences sociales. En revanche, *sure* est pratiquement réservé à l'expression de la certitude, ce qui explique sa rareté en sciences sociales. D. Biber en conclut qu'un corpus limité à un seul registre ne permettrait qu'une analyse partielle de l'usage, et que la généralisation de cette analyse à la langue dans son ensemble serait incorrecte.

Niveau morpho-syntaxique

L'ambiguïté grammaticale, ou polycatégorie, dont la résolution est un des problèmes centraux de l'annotation morpho-syntaxique automatique, est le plus souvent considérée *en langue* et non *en discours*. L'expression *en discours* fait ici référence non seulement au contexte immédiat qui permet de désambiguïser, mais aussi à l'appartenance à un type de contexte, qui permet de faire des prédictions. D. Biber donne quelques exemples frappants de variations du fonctionnement de lexèmes ou de classes selon le registre (Biber, 1993b). Bien que ses observations portent sur l'anglais, langue où les problèmes de polycatégorie sont sans doute assez différents de ceux qui se posent en français, sa méthodologie et ses conclusions paraissent assez pertinentes pour que j'en présente quelques extraits.

Son étude sur deux "genres" du corpus LOB (textes de fiction et textes expositifs) concerne deux types d'information fréquemment utilisés pour l'annotation morpho-syntaxique : la probabilité relative de chaque catégorie grammaticale pour les lexèmes ambigus, et la probabilité relative de séquences de catégories pour les groupes ambigus. En voici quelques résultats :

– les distributions des formes pouvant être soit un verbe, soit un nom (*trust, rule*, formes en *-ing*), soit un verbe, soit un adjectif (*secure*), soit un nom, soit un adjectif (*major, representative*) ont des distributions clairement différenciées (tableau 8.2) :

Type d'ambiguïté grammaticale	**Textes de fiction** catégorie dominante	**Textes expositifs** catégorie dominante
nom/verbe ex. *trust, rule, -ing*	verbe	nom
verbe/adjectif ex. *secure*	verbe	adjectif
nom/adjectif ex. *major, representative*	nom	adjectif

Tableau 8.2 : Distribution des formes polycatégorielles dans deux "genres" (d'après D. Biber).

– *that* a à peu près la même distribution dans les deux sous-corpus en tant que pronom relatif, mais une probabilité beaucoup plus grande d'être un démonstratif dans les textes de fiction et une conjonction dans les textes expositifs ;

– séquences de catégories : la copule *be* est le plus souvent suivie d'un passif dans les textes expositifs, d'une forme progressive dans les textes de fiction ;

– rattachement des syntagmes prépositionnels : le rattachement à un élément verbal domine dans les textes de fiction (78,7%), alors que dans les textes expositifs les syntagmes prépositionnels se rattachent en proportion égale à des éléments nominaux ou verbaux.

Niveau syntaxique

J-P. Sueur plaidait dès 1982 pour une intégration des études statistiques des faits linguistiques dans une syntaxe alors dominée par le générativisme. Les jugements de grammaticalité, disait-il, "ne sont peut-être que la construction d'un corpus, ou plutôt la projection de l'image qu'on se fait d'un corpus sur quelques exemples" (Sueur, 1982:149). Un corpus construit et documenté donne lieu à des généralisations plus modestes mais mieux motivées. De façon indépendante et *a posteriori*, les travaux de D. Biber illustrent et confirment cette intuition : pour un grammairien qui s'interroge sur la grammaticalité de la suppression de *that* ("*that*-deletion") dans les relatives et complétives, par exemple, il est intéressant de savoir que les occurrences de ce phénomène sont en corrélation très forte avec l'implication du locuteur dans le texte et qu'il est généralement absent des textes à visée principalement informative (Biber, 1988). Les jugements de grammaticalité hors contexte ont tendance à occulter la variabilité liée aux registres, ce qui limite leur utilité et leur généralisabilité.

En ce qui concerne la variabilité liée au domaine, J. McNaught (1993) met l'accent sur ce qu'il perçoit comme une des caractéristiques les plus marquées du fonctionnement lexico-syntaxique des sous-langages : des verbes qui sont des homographes de verbes en langue générale ont fréquemment dans un sous-langage donné des structures prédicat-arguments complètement différentes de celles qui leur correspondent en langue générale ou dans d'autres sous-langages. Ceci a trait à la relation extrêmement serrée entre syntaxe, sémantique et domaine conceptuel qui caractérise les sous-langages (cf. 5.1.3.). B. Habert *et al* (1997) donnent un exemple issu du domaine de la vinification où l'omission du complément d'un verbe transitif est rendue systématique par sa totale prédictabilité : *on sucre* y est acceptable, mais l'explicitation de l'argument dans **on sucre le moût* produit un énoncé inacceptable.

Dans les exemples ci-dessus, les questions posées sont classiques dans le sens où elles ont trait à la grammaticalité de certaines constructions. Les réponses le sont moins : il ne s'agit plus d'une distinction binaire grammatical/agrammatical, mais d'une analyse faisant intervenir les conditions de production et l'importance relative des différentes réalisations selon ces conditions. On s'achemine, comme ont pu le dire C. Blanche-Benveniste (1996), ou B. Habert *et al* (1997) (après J-P. Sueur (1982)), vers une vision probabiliste de la grammaire. Les frontières deviennent plus floues entre linguistique descriptive et linguistique de l'usage, comme en témoigne l'analyse des phrases passives dans le corpus Brown résumée dans le tableau 8.3 ci-dessous (Francis & Kucera, 1982).

L'examen de ce tableau permet de développer deux apports spécifiques de l'analyse de corpus, à partir d'observations qui semblent aussi révélatrices de ce qu'est le passif qu'une modélisation purement syntaxique. D'abord la quantification des occurrences met en lumière la rareté du passif, qui ne concerne que 11% des phrases du corpus tous genres confondus (voir aussi Halliday, 1991). S'ajoute à cela une observation qui va à l'encontre d'une vision purement syntaxique qui ferait des énoncés au passif de simples "variantes" d'énoncés actifs : W. Francis et H. Kucera notent que la grande majorité de ces passifs (85%) n'ont pas d'agent. Là où l'agent est exprimé, il est presque toujours de type non humain comme dans *may be achieved by several methods*. Il apparaît donc clairement que le passif "fait" dans les textes quelque chose de foncièrement différent de ce que "fait" la forme

active. L'analyse de corpus fournit ici des éléments essentiels pour une approche fonctionnelle de cette construction.

Genre discursif	% phrases actives	% phrases passives
1. Informative prose		
A. Press: reportage	87.34	12.66
B. Press: editorial	88.83	11.17
C. Press: reviews	90.72	9.28
D. Religion	88.23	11.77
E. Skills and hobbies	85.75	14.25
F. Popular lore	87.49	12.51
G. Belles lettres	89.43	10.57
H. Miscellaneous	75.85	24.15
J. Learned	78.05	**21.95**
		14.21
2. Imaginative prose		
K. General fiction	95.21	4.79
L. Mystery and detective	96.19	3.81
M. Science fiction	93.40	6.60
N. Adventure / western	98.41	3.59
P. Romance / love story	96.68	3.32
Q. Humour	93.13	6.87
		4.83
Whole corpus	*88.93*	***11.07***

Tableau 8.3 : Phrases actives et passives dans le corpus Brown.

Le deuxième apport a trait aux données que fournit le tableau sur la distribution des phrases passives dans les subdivisions du corpus. Le corpus Brown a été conçu pour représenter la gamme des réalisations de l'anglais américain écrit, en fonction d'une classification en "genres". On constate que le passif est extrêmement rare dans les textes de fiction (4,83% des phrases), mais qu'il est fortement associé en revanche à un genre informatif, les textes "savants" ou scientifiques (21,95% des phrases). C. Blanche-Benveniste (1996) évoque des résultats similaires pour le français. Là encore, l'analyse de corpus diversifiés fournit des données importantes pour une analyse fonctionnelle telle que celle présentée dans le chapitre 2. Elle n'est pas sans pertinence non plus pour le traitement automatique, en particulier pour l'étiquetage probabiliste, comme le montre D. Biber en contrastant les "préférences" catégorielles de formes ambiguës en fonction du genre (1993b) : son analyse du corpus LOB, corpus conçu suivant les mêmes principes d'échantillonnage que le corpus Brown mais pour l'anglais britannique, lui permet de comparer les probabilités d'étiquetage de formes polycatégorielles selon qu'elles apparaissent dans des textes de fiction ou des textes expositifs. Pour un groupe de formes en *-ed*, qui peuvent être soit des verbes au passé "simple", au passé "composé" (*perfect*), ou à la forme passive, soit des adjectifs, il constate une probabilité de 78 % en moyenne qu'il s'agisse de verbes au passé dans les textes de fiction contre 65% qu'il s'agisse de verbes au passif dans les textes expositifs.

Sur cette base, il suggère que les systèmes d'analyse automatique fondés sur des techniques probabilistes, plutôt que d'utiliser un ensemble unique de probabilités pour un traitement "généraliste, devraient partir de probabilités calculées séparément pour ce qu'il appelle les "registres majeurs", qu'on peut interpréter comme des regroupements de genres. C'est ce type d'amélioration que propose N. Smith en 1996 pour l'étiqueteur CLAWS en

évoquant : " the tuning of the tagger to each major variety of text in the [BNC] corpus" (Smith, 1996:149). Ce "réglage" concernerait le lexique, la matrice de transitions d'étiquettes, et les règles. G. Illouz (1999) vont dans le même sens à partir de la constatation de différences marquées dans la performance d'étiqueteurs pour le français selon le genre discursif.

8.2.2 Problèmes de représentativité et d'hétérogénéité

De même que l'efficacité d'un étiqueteur dépend de la coïncidence entre le corpus d'entraînement et le corpus à traiter, la validité des résultats de l'analyse de corpus pour une recherche particulière est fonction de l'adéquation des données aux objectifs de cette recherche. L'évaluation de cette adéquation, dont P. Corbin (1980) met au jour la difficulté dans le cas de l'analyse syntaxique sur la base de données introspectives, est d'autant plus complexe en ce qui concerne les études sur corpus, beaucoup plus diversifiées dans leurs objets et leurs objectifs. Il est intéressant de noter que dans cet article de 1980, qui commence par opposer "linguistique de bureau" et "linguistique de terrain", P. Corbin envisage l'appel à des corpus dans la linguistique de terrain uniquement dans une perspective sociolinguistique ("la prise en charge des aspects sociaux des variations langagières" p.121). Dans cette optique, les corpus sont construits pour représenter des populations humaines. Dans la conception plus courante aujourd'hui des linguistiques de corpus, c'est une langue – ou un sous-langage – qu'on cherche à décrire à partir de ses manifestations, et la notion de "populations de textes" (Biber, 1993a) devient centrale. Celle-ci permet de penser le problème de l'adéquation des données non seulement en termes de représentativité du corpus par rapport aux populations concernées, mais aussi en termes de pertinence de ces populations par rapport à l'objectif de la recherche. Je vais envisager trois cas de complexité croissante pour formuler l'articulation entre ces deux aspects.

Premier cas : la population de textes concernée par l'étude est entièrement contenue dans le corpus. C'est le degré zéro de la constitution de corpus : l'œuvre complète de Marguerite Yourcenar, l'ensemble des textes existants en grec de l'époque socratique, ou en ancien français d'une période donnée. Le choix de textes est entièrement déterminé dès le départ, à partir du choix de la population. L'analyse de corpus dans ce cas vise exclusivement la description des usages à l'intérieur du corpus. Il y a correspondance totale entre l'étendue du corpus et la portée des observations.

Second cas : le corpus doit représenter une population de textes spécifiques. On cherche par exemple à construire la terminologie d'un domaine technique. Au delà d'une description des usages dans le corpus, le travail d'analyse vise alors une modélisation, c'est-à-dire une description susceptible d'être généralisée à l'ensemble des textes produits dans ce domaine. Le choix de population est déterminé par l'objectif. La validité des résultats dépend fondamentalement de la représentativité des textes constituant le corpus par rapport à cette population. Se pose aussi le problème de l'homogénéité du corpus : ce type d'étude présuppose en effet le plus souvent que le domaine est le seul facteur de variation. Il y a lieu de s'interroger sur l'impact d'autres facteurs, tel le genre discursif.

Troisième cas : le corpus est envisagé comme un échantillon représentatif de "la langue". C'est la position des recherches sur la "langue générale" pour lesquelles le corpus est essentiellement une mine de données attestées reflétant le potentiel de la langue. Elles se sentent généralement peu concernées par les problèmes de constitution de corpus. Leur objectif se situe à l'opposé de la visée modélisatrice : les résultats de l'analyse ne sont pas censés caractériser le corpus, ni même une population de textes, par rapport aux traits étudiés, mais refléter le fonctionnement de la langue. On peut donc se demander quelles seraient les spécifications d'un corpus pour ce type d'étude. En quoi consisterait un corpus de "langue générale" ? Deux réponses ont cours : le "corpus équilibré", qui aurait "de tout un peu", et le corpus de textes "non-marqués".

La construction d'un "corpus équilibré" nécessiterait de déterminer ce qu'est "tout" et ce qu'est "un peu". C'est-à-dire d'une part un modèle aussi complet que possible de la variation langagière, de l'autre des principes d'échantillonnage. En fait, la recherche de corpus équilibrés semble bien constituer une impasse : la notion d'équilibre s'apparente à celle de "langue générale", et elle paraît tout aussi insaisissable. Elle suppose également une recherche d'exhaustivité irréaliste, étant donné qu'il n'existe pas d'inventaire exhaustif des populations concernées (Habert, 2000). Dans une perspective plus modeste, des corpus conçus à partir de principes de diversification et d'équilibre ont été constitués pour l'anglais[76]. Ils n'exonèrent en rien de la nécessité de prendre en compte la variation, mais ils facilitent cette prise en compte en présentant de façon convenablement documentée des échantillons de textes de différents genres discursifs. Le principe sous-jacent n'est donc pas de prendre le corpus diversifié dans son ensemble comme représentant globalement la "langue générale", mais de permettre de traiter séparément des sous-corpus, pour aboutir à une analyse "composite", selon le terme de D. Biber.

L'autre angle d'approche dans la recherche d'un corpus de langue générale est de sélectionner une population supposée non-marquée, telle une certaine presse quotidienne pour reprendre l'exemple évoqué plus haut. Un tel choix repose sur deux hypothèses problématiques : le corpus est représentatif et il est homogène. L'hypothèse de représentativité est invérifiable dans la mesure où la langue générale demeure une abstraction, qui ne se laisse saisir ni par les corpus ni par l'introspection. Avec une lucidité et une indépendance d'esprit peu communes, P. Corbin résume le problème du point du vue de l'introspection :

> Il est illusoire, (…), de prétendre décrire *LA langue*, conçue comme le plus petit dénominateur linguistique commun à tous les membres d'une communauté linguistique (par exemple le français dit *standard*). À supposer qu'un tel noyau existe, c'est une abstraction à laquelle ne se réduit le savoir linguistique d'aucun locuteur particulier : il n'est pas accessible par introspection. (1980:155)

Du point de vue des linguistiques de corpus, ne pourraient être considérés comme non-marqués sur le plan du genre que des textes dont on ne pourrait deviner l'origine ou l'intention ("textoïdes"), ce qui ne semble pas être le cas pour les articles de presse quels qu'ils soient. Lorsqu'on a affaire à des textes attestés inscrits dans des situations de communication, l'existence de corrélats linguistiques de divers paramètres de ces situations a été démontrée par les travaux de D. Biber (cf. 8.3 infra), ce qui met en cause la notion de textes non-marqués.

Quant à l'hypothèse d'homogénéité, il est clair qu'à l'intérieur d'un même quotidien cohabitent des articles représentatifs de domaines et de visées discursives très diversifiés. Il est possible que cette diversité n'ait pas de pertinence par rapport à un objectif de recherche particulier, l'important étant de pouvoir gérer une hétérogénéité connue et assumée.

P. Corbin conclut l'article cité en définissant la spécificité et les limites de l'intuition comme mode de connaissance :

> repérer des potentialités linguistiques sans savoir dans quelles conditions ni par quels locuteurs elles sont susceptibles d'être converties en pratiques langagières effectives. (op.cit.:160)

Les limites des corpus sont à l'inverse de représenter des actualisations de la langue qu'on ne peut envisager indépendamment des paramètres qui déterminent les pratiques langagières effectives. Les linguistiques de corpus, quels qu'en soient les objectifs, exigent donc des corpus construits en fonction d'un modèle de la variation langagière, et aussi documentés que possible quant aux paramètres de cette variation. La construction d'un

[76] Voir Habert *et al.* (1997) pour une liste des grands corpus disponibles pour l'anglais.

modèle de la variation, qui passe par l'identification des paramètres pertinents, fait l'objet de la prochaine section.

8.3 Modéliser la variation pour pouvoir la gérer et l'utiliser

Exception faite de l'école systémique autour de M. Halliday, et de quelques études en syntaxe fonctionnelle (Givón, 1979), les travaux sur l'organisation textuelle dont je me suis inspirée se préoccupent assez peu de la diversité des réalisations textuelles – structures ou marqueurs – en fonction de paramètres extra-linguistiques. Pour ma part, j'ai très vite ressenti le besoin de structurer le foisonnement des productions textuelles : si un objectif de mes recherches est d'identifier des marqueurs d'organisation dans des textes, ce n'est pas dans une visée étroitement descriptive, mais à partir d'une hypothèse de généralisabilité des observations sur corpus à des familles de textes. Je le formulais comme suit en 1994 :

> La démarche typologique apparaît comme centrale pour la linguistique du discours : elle est une façon de penser la systématicité en parole. Chaque énonciation est certes unique, mais il est possible de regrouper ces spécimens en types selon des critères pertinents. Ainsi, si les marques de structuration thématique ou rhétorique ne font pas partie de la grammaire générale de la langue, leur utilisation est sans doute jusqu'à un certain point prévisible à l'intérieur d'un *type de texte* donné. (Péry-Woodley 1994:9)

Cette position de principe laisse toutefois ouvert le problème de la méthode. Les approches typologiques "classiques", qui classent les textes en fonction d'un modèle de la communication – modèle fonctionnel, socio-institutionnel ou issu de la rhétorique classique – pour caractériser ensuite ces classes sur le plan linguistique, échappent difficilement à la circularité (cf. Péry-Woodley, 1993:98-99). Pour y voir clair, il faut impérativement traiter de façon séparée les caractéristiques internes (linguistiques) et les caractéristiques externes (fonctionnelles, situationnelles). C'est ce qu'a su faire D. Biber en distinguant clairement types (linguistiques) et genres (situationnels), et en élaborant une méthodologie originale pour induire les types de texte à partir d'analyses de corpus.

8.3.1 La typologie "émergente" de D. Biber

La démarche typologique de D. Biber (1988; 1989) s'inscrit dans l'approche dite "macroscopique" introduite dans le chapitre précédent, son but étant l'étude systématique de la variation linguistique, plutôt que la validation de types prédéfinis. Elle est "émergente" dans la mesure où l'identification des types n'est pas donnée *a priori*, mais émerge du traitement statistique de la caractérisation linguistique des textes. À partir de deux corpus de l'anglais (LOB et London-Lund), structurés en "genres" écrits et parlés tels que reportages, textes scientifiques, romans d'aventure, conversations téléphoniques, D. Biber procède à une analyse des cooccurrences de 67 traits linguistiques associés aux conditions d'énonciation, au niveau d'abstraction, au degré d'intégration ou de fragmentation du discours, ou encore au degré de détachement ou d'implication du locuteur. La première étape passe par le regroupement en *facteurs* de traits linguistiques qui sont fréquemment en cooccurrence dans les textes (ex. temps du passé + 3e personne + aspect accompli + verbes dicendi □+ participes présents) ; ces facteurs font ensuite l'objet d'une interprétation en termes de *dimensions* textuelles, à travers l'identification des fonctions communicatives maximalement communes aux traits qui les constituent (ex. la narration). Les dimensions opposent deux pôles, qui correspondent chacun à un ensemble de traits (orientation narrative *versus* non-narrative). Elles constituent ainsi des échelles sur lesquelles se situent les textes (ex. plus ou moins orientés vers la narration).

Les pôles de regroupement de traits définissent ainsi cinq *dimensions* majeures. Le tableau 8.4 en donne la liste avec en regard les principaux traits linguistiques qui en caractérisent le premier pôle :

Dimension	**Traits caractéristiques positifs (1er pôle)**
Production impliquée *versus* production à visée informative	*do* comme pro-verbe, 1e et 2e personne, *be* comme verbe principal, présent, démonstratifs, contractions (*don't*)
Orientation narrative *versus* non-narrative	passé, 3e personne, participes présents, verbes *dicendi*
Référence explicite *versus* dépendante de la situation	propositions relatives objet et sujet, coordination phrastique, nominalisations
Visée persuasive apparente	infinitifs, modaux (prédiction, nécessité, possibilité), verbes de persuasion, subordination conditionnelle
Information abstraite *versus* non-abstraite	connecteurs, passifs, subordonnées réduites, propositions circonstancielles

Tableau 8.4 : "Dimensions" de la typologie de D. Biber (1988).

Dans une deuxième étape, des techniques de classification automatique permettent à D. Biber de rapprocher les textes en fonction de leurs coordonnées sur ces cinq échelles. Il identifie huit regroupements qui sont alors considérés comme des "types" de textes :

1) Interaction interpersonnelle intime
2) Interaction à but informatif
3) Exposé scientifique
4) Exposé savant
5) Exposé narratif
6) Fiction narrative
7) Reportage situé
8) Persuasion impliquée.

Ces types sont déterminés de façon à ce que : (i) les textes appartenant à chaque type partagent le maximum de caractéristiques linguistiques ; (ii) les différents types soient le plus distincts possible. S'ils se prêtent à une interprétation fonctionnelle, bien que la méthodologie soit tout à fait différente de celle des typologies fonctionnelles classiques, c'est, selon D. Biber, que la cooccurrence de traits vient refléter une fonction commune.

Un apport essentiel de cette approche est de fonder une distinction nette entre les types de textes, caractérisés de façon linguistique, et les registres ou genres, qui correspondent à des paramètres extra-linguistiques. En pointant sur un graphe l'ensemble des scores pour une dimension linguistique pour les différents genres présents dans le corpus, D. Biber en expose le degré de cohérence linguistique. Il constate par exemple que les lettres professionnelles sont, contrairement à l'intuition, beaucoup moins "typées" linguistiquement que les lettres personnelles. En effet, alors que ces dernières ont des objectifs entièrement interactionnels et affectifs, les lettres professionnelles ont des objectifs à la fois informationnels et interactionnels, et penchent tantôt d'un côté tantôt de l'autre. Les corpus LOB et London-Lund utilisés par D. Biber présentent une partition en genres et sous-genres parfois discutables dans la mesure où elle mélange des paramètres liés à la situation et au domaine. D. Biber a ensuite œuvré dans le sens d'une meilleure définition des genres pour la constitution de corpus en proposant une grille des facteurs extra-linguistiques susceptibles de jouer un rôle dans la variation langagière (prochaine section).

Il était nécessaire d'exposer de façon assez complète ce travail typologique de D. Biber dans la mesure où il représente une avancée marquante dans la façon d'envisager la classification des corpus. Lorsqu'on travaille sur une langue "minoritaire" et relativement démunie sur le plan des ressources en corpus comme le français, on ne peut cependant bénéficier dans l'immédiat de cette avancée. On est loin en effet de pouvoir procéder à une adaptation de la méthodologie biberienne pour le français, non pas qu'on manque de travaux descriptifs qui pourraient servir de base à la constitution d'une liste de traits pour une typologie émergente, mais parce qu'on ne dispose actuellement d'aucun corpus de référence. Deux projets sont toutefois à l'étude qui se situent directement dans la lignée des travaux de D. Biber : le premier est un projet de "corpus clé pour le français actuel", qui s'inspire du récent British National Corpus (Habert, 1998) ; le second propose un outil de profilage automatique de textes, qui permettrait de calibrer les textes d'un corpus hétérogène en fonction de traits de niveaux variés, et de positionner un nouveau texte par rapport aux regroupements obtenus sur un corpus préexistant (Illouz *et al.,*1999; Habert *et al.*, 2000).

Pour ce qui est de la caractérisation externe des textes dans la constitution de corpus, j'ai évoqué à plusieurs reprises dans l'exposé de mes travaux la notion de *configuration domaine/genre*. Je vais chercher à préciser dans ce qui suit ces deux paramètres importants pour la classification extralinguistique des textes spécialisés qui sont mon objet d'étude dans ma recherche de marqueurs d'organisation textuelle.

8.3.2 Domaines : des langues de spécialité aux sous-langages

L'idée que les textes varient dans plusieurs de leurs dimensions (lexicale, syntaxique) en fonction des domaines de connaissances ou d'expérience auxquels ils ont trait est depuis longtemps explorée et exploitée en didactique des langues et parmi les spécialistes de la rédaction technique (cf. Swales, 1990). On parle alors généralement de langues de spécialité. Plus récemment, et plus spécifiquement dans les milieux concernés par le TAL, est apparu le terme de sous-langage (cf. 5.1.3.). Bien qu'il y ait recoupement, ainsi que l'indique une définition "informelle" selon laquelle un sous-langage est "the language used by a particular community of speakers, say, those concerned with a particular subject matter or those engaged in a specialized occupation." (Sager 1986:2), la recherche sur les sous-langages a des objectifs applicatifs et des fondements théoriques spécifiques.

Ce sont sans doute les objectifs applicatifs qui expliquent le mieux l'expansion de ces recherches au cours des quinze dernières années : la description détaillée de tous les niveaux linguistiques qui serait nécessaire au traitement de toute une langue apparaît comme une tâche impossible. Heureusement, constatent R. Grishman et R. Kittredge,

> many language processing problems are effectively restricted to the language used in a particular domain (...). The variety of language used in a given science or technology not only is much smaller than the whole language, but is also more clearly systematic in structure and meaning. (1986:ix).

Ces caractéristiques – "dimensions" réduites, systématicité structurelle et sémantique – renvoient au modèle linguistique dans lequel prend place la notion de sous-langage. Il s'agit du modèle progressivement développé par Z. Harris des années 60 aux années 90 (Harris, 1968; 1982; 1988; 1991) et qui isole par leur comportement syntaxique la métalangue, elle-même un sous-langage, et les "sous-langages de domaine". Je reprends brièvement les caractéristiques principales des sous-langages selon Z. Harris :

– ils ont un lexique restreint ;

– ils ont une syntaxe restreinte, dans laquelle chaque type de phrase est caractérisable comme une combinaison particulière de sous-classes de mots ;

– les restrictions de sélection, qui ne peuvent faire l'objet de règles au niveau de la langue dans son ensemble, font partie de la grammaire des sous-langages.

On comprend qu'au delà de l'intérêt strictement applicatif – c'est dans les domaines techniques que se situent la plupart des besoins en TAL –, de nombreux linguistes se soient intéressés à ces microcosmes linguistiques que semblent présenter les sous-langages, en fonction de leurs caractéristiques de fermeture et de systématicité. Les travaux sur les sous-langages se fondent évidemment sur des corpus de textes issus de domaines scientifiques ou techniques (pharmacologie, médecine, espace, météorologie, pour ne citer que les plus connus), mais il est clair que l'intérêt de tels corpus dépasse les frontières de ces domaines. Pour ce qui me concerne, le choix de travailler sur les sous-langages est fondé sur l'hypothèse d'une extension des régularités observées par Z. Harris et ses collaborateurs (Sager, *op.cit*; Sager *et al.,*1987; Ryckman, 1990) au domaine textuel.

Toutefois, si la notion de sous-langage contribue de façon importante à la compréhension et à la description des spécificités lexicales et syntaxiques liées au domaine, elle n'épuise pas les possibilités de variation dans les textes, ainsi que le remarquent C. Montgomery et B. Glover dans leur contribution à l'ouvrage fondateur sur les sous-langages coordonné par R. Grishman et R. Kittredge :

> Clearly, a given sublanguage will never be totally distinct from other sublanguages in semantic domain and/or grammatical and discourse structure; some features will be shared by other related sublanguages. For example, the sublanguage of cookbooks and the sublanguages exemplified by other technical manuals such as aviation hydraulics share many grammatical and format features (…), for they all involve the specification of procedures for carrying out tasks and require formats that are conceived of as optimal for communicating a procedural type of information. (1986:158)

La question posée est celle des similarités linguistiques entre des textes qui appartiennent à des sous-langages différents mais qui sont apparentés sur le plan de la fonction discursive. R. Grishman et R. Kittredge soulèvent un problème apparenté en évoquant les séquences discursives composant un texte – par exemple un manuel comportant des descriptions et des consignes – et qui, bien qu'appartenant à un même sous-langage, sembleraient nécessiter des grammaires distinctes (1986:xvi ; cf. aussi Biber & Finegan, 1994). La section suivante va explorer ces autres aspects de la variation, et proposer des modèles potentiellement exploitables pour la constitution et la documentation de corpus, ainsi qu'une méthodologie visant l'automatisation de l'identification du type de texte.

8.3.3 Genres discursifs

Il existe en effet une autre grande classe de facteurs de variation : ceux qui ont trait à la "posture" énonciative et aux types d'opérations discursives effectuées, elles-mêmes plus ou moins déterminées par les objectifs communicationnels et les situations dans lesquelles s'inscrivent les textes. Plusieurs classifications sont possibles selon que l'accent est mis sur l'un ou l'autre paramètre. Comme je l'ai évoqué dans la section précédente, D. Biber propose un modèle paramétrique qui tient compte de deux grandes sources de variation : celle liée à l'usage et celle liée aux interlocuteurs. En effet, seul un modèle paramétrique pourra rendre compte du fait que s'il existe bien des prototypes nets, les frontières entre registres ou genres sont inévitablement floues, et qu'il serait vain de chercher à en établir une liste exhaustive. La notion de registre est donc spécifiée comme une notion continue, et non discrète.

Je reproduis ci-dessous (figure 8.1) la liste simplifiée des paramètres situationnels qui doivent être vus comme des "strates d'échantillonnage" hiérarchisés et complémentaires (Biber, 1993a:245).

---

1. Canal : écrit / parlé / écrit lu
2. Format : publié / non publié (+ divers formats sous "publié")
3. Cadre : institutionnel / autre cadre public / privé-personnel
4. Destinataire :

a. pluralité : non compté / pluriel / individuel / soi-même
b. présence : présent / absent
c. interaction : aucune / peu / beaucoup
d. connaissances partagées : générales / spécialisées / personnelles
5. Destinateur :
a. variation démographique : sexe, âge, profession, etc.
b. statut : individu / institution dont l'identité est connue
6. Factualité : informatif-factuel / intermédiaire / imaginaire
7. Objectifs : persuader, amuser, édifier, informer, expliquer, donner des consignes, raconter, décrire, enregistrer, se révéler, améliorer les relations interpersonnelles, …
8. Thèmes : …

Figure 8.1 : Paramètres situationnels de D. Biber (1993a).

L'auteur reconnaît lui-même le flou qui entoure les deux derniers paramètres. Le problème est que ces paramètres sont censés couvrir toute la diversité des situations de production langagière. En fait, les textes situés dans des cadres professionnels sont très peu représentés dans les corpus sur lesquels ont porté les études typologiques de D. Biber. Il se pourrait pourtant que la détermination des paramètres situationnels pertinents soit elle aussi plus simple dans un sous-langage : les textes produits dans un cadre professionnel donné ont des dénominations qui correspondent à une nomenclature close. La notion de *genre*, qui désigne les catégories intuitives selon lesquelles les usagers de la langue reconnaissent, interprètent, produisent des documents, prend des contours plus précis dans le cadre professionnel sous la forme de *document de travail* : procès-verbal, convocation, compte-rendu d'hospitalisation, manuel de références… Ainsi, dans mon cadre de travail, le monde universitaire, *article* se distingue précisément de *notes de lecture*, *mémoire* de *thèse*… À chaque document est attachée une configuration spécifique des paramètres situationnels ci-dessus. Le *procès-verbal de réunion*, pour prendre un autre exemple, est écrit, publié (en interne), appartient au cadre institutionnel, a des destinataires pluriels, absents, non-intéragissants, partageant avec le scripteur des connaissances spécialisées, il est informatif-factuel, et son objectif est un mélange bien spécifique de récit, d'informations et de consignes. De même qu'on fait appel à des experts d'un domaine pour en établir l'ontologie, il serait utile pour la constitution de corpus spécialisés d'interroger des professionnels sur la nomenclature des documents qu'ils utilisent et produisent. C'est en rapport avec ces catégories, qui correspondent à une compétence textuelle, qu'il sera intéressant d'identifier des caractéristiques linguistiques. Ce que j'ai appelé configuration genre-domaine réfère donc au croisement d'un document de travail avec un domaine de connaissance : manuels de référence et informatique par exemple.

Chapitre 9

Conclusion

Ce mémoire se termine. Il a présenté des travaux échelonnés sur plus de dix ans, qui reflètent certaines évolutions de la linguistique au cours de cette période tant dans la conception de son objet que dans ses méthodes. J'esquisse en guise de conclusion quelques grandes lignes de ces évolutions, de façon partielle et inévitablement partiale, en lien avec mes recherches.

9.1 De la langue aux discours

La délimitation saussurienne de l'objet de la linguistique, à travers la dichotomie *langue-parole*, a longtemps eu pour effet d'opposer système et usage du système (cf. Halliday, 1991). En démontrant l'impact du contexte linguistique et extra-linguistique sur la structure syntaxique des phrases, les grammairiens fonctionnalistes ouvrent dès les années 60 une brèche dans la muraille théorique qui séparait système et usage. C'est le cas, à la suite des linguistes de l'École de Prague (cf. Firbas, 1964), de M. Halliday, qui introduit la notion de structure d'information pour théoriser l'influence structurante des contextes dans lesquels sont énoncées les propositions, envisagées comme les représentations conceptuelles d'états de choses (Halliday, 1967). Cette notion permet de rendre compte d'alternatives syntaxiques dont l'occurrence sélective (dislocation, clivée, etc. *versus* phrase canonique) ne peut s'expliquer qu'en termes pragmatiques, par rapport au statut présuppositionnel des entités introduites, et à leur degré d'identifiabilité et d'activation. Ainsi, il devient clair que la description de certains fonctionnements syntaxiques nécessite qu'on les envisage comme inscrits dans des usages, c'est-à-dire dans leur dimension discursive (Lambrecht, 1994).

Au delà de cet éclairage pragmatique de la syntaxe, l'ouverture vers l'usage met en cause le statut de la phrase comme unité maximale des études linguistiques. Je citerai à nouveau M. Halliday, ou T. van Dijk, dans une perspective plus sémantique, parmi les linguistes qui ont mis le texte, vu comme unité fonctionnelle formant un tout qui se tient, sur le devant de la scène linguistique (Halliday & Hasan, 1976; van Dijk, 1977). Le texte venant à être perçu comme une unité proprement linguistique, l'étude de ses modes de construction rentre dans le champ de la linguistique. Si le texte est une unité linguistique, c'est en tant qu'ensemble de signaux permettant la (re)construction d'un discours, c'est-à-dire l'élaboration par un lecteur d'un modèle interprétatif à partir des traces textuelles de la construction d'un discours par le scripteur. De nombreux aspects de l'organisation textuelle ne peuvent s'envisager que dans cette perspective interactionnelle : c'est le cas de tout ce qui concerne le topique et le choix d'expressions référentielles, anaphoriques ou non.

Se réclamer des linguistiques du discours, même en laissant de côté "l'analyse de discours" axée sur l'interprétation, peut ainsi prendre au moins deux significations : dans l'une, le regard s'ouvre aux facteurs extra-linguistiques, mais l'unité reste la phrase, comme c'est le cas dans la syntaxe fonctionnelle ; dans l'autre, l'unité d'analyse change et l'objectif inclut les relations inter-phrastiques, nécessairement en lien avec les aspects pragmatiques. En ce qui me concerne, je me reconnais dans ces deux courants : le premier par l'attention au détail des structures dans les textes, le second par le fait de prendre l'unité texte comme objet d'étude.

9.2 De l'exemple construit aux textes attestés

L'ouverture de la linguistique au discours s'est faite conjointement à l'évolution des données utilisées pour la description et la construction théorique. A la prédominance de l'introspection comme mode de production de données succède actuellement un "boom" des corpus. Cette évolution contribue elle aussi à mettre en question la dichotomie langue-parole, ainsi que l'argumente M. Halliday (1991) par le biais d'une analogie rafraîchissante avec la relation climat-temps (qu'il fait) : le météorologue observe et mesure au jour le jour les variations de température, de direction du vent, de pression atmosphérique ; le climatologue modélise en termes probabilistes le potentiel climatique d'une zone donnée. La différence entre temps (qu'il fait) et climat est liée au fait qu'il y a deux observateurs qui utilisent des échelles de temps différentes. N'objectivisons pas la distinction méthodologique, s'insurge M. Halliday : le système et ses instances ne sont pas des phénomènes distincts, et plus nous observons le temps qu'il fait (les textes pour le linguistes), plus nous serons capables de modéliser le climat (le système linguistique).

Cette analogie permet aussi de rendre compte d'un autre aspect de l'influence des corpus sur l'objet et les objectifs de la linguistique : de même que la climatologie construit des modèles climatiques par zones, pour lesquelles sont calculées les probabilités, les linguistiques de corpus mettent de plus en plus l'accent sur la nature fragmentée et probabiliste de la grammaire. Les zones sont délimitées par des configurations de paramètres situationnels (genre et domaine) ; par ailleurs les règles sont de plus en plus perçues comme constituant non pas des systèmes fermés, mais des modèles probabilistes (cf. Blanche-Benveniste, 1996). Avec cette vision éclatée de la grammaire comme un ensemble de modèles probabilistes, on est donc bien loin de l'idée d'un système unique modélisé à partir de jugements de grammaticalité essentiellement binaires.

9.3 Du parsage aux traitements robustes pour la recherche d'information

Le traitement automatique a suivi une évolution parallèle, bien que spécifique. Ce domaine a longtemps été prioritairement axé sur l'analyse syntaxique complète, envisagée comme la première phase indispensable de tout traitement. L'étiquetage et l'analyse syntaxique ne se concevaient que sur la base de règles, et une bonne partie de la recherche se focalisait sur l'élaboration de formalismes. Les applications visées se résumaient pratiquement aux trois "classiques" : traduction, compréhension, génération (cf. Fuchs *et al.*, 1993). Les dix dernières années voient l'arrivée des grands corpus, le développement des approches probabilistes pour l'étiquetage morpho-syntaxique, la prise de conscience de l'importance du lexique, l'explosion des applications, particulièrement dans le domaine de la recherche d'information dans des bases textuelles. Le passage de la compréhension automatique à la recherche d'information est symptomatique de cette évolution : plutôt que de viser une analyse exhaustive qui réunirait composantes syntaxique, sémantique et pragmatique pour aboutir à une représentation complète du "sens", on extrait d'ensembles de textes homogènes des informations permettant de répondre à des questions très spécifiques, de manière à remplir des formulaires ("templates"). Les techniques changent, mais aussi les

objectifs : ces derniers se diversifient, se morcellent, collent davantage aux applications. Il devient difficile de parler de traitement automatique au singulier.

Cette évolution fait la part belle aux traitements partiels, analyse syntaxique robuste, recherche de marqueurs de surface. Elle accorde également une place importante aux corpus, que ce soit pour l'apprentissage des étiqueteurs, pour le repérage de réalisations lexico-syntaxiques spécifiques pour la recherche d'information ou pour l'acquisition lexicale. Là encore, on semble s'éloigner d'une conception rationaliste et uniciste, ayant pour objectif la modélisation de LA langue, pour s'orienter vers un double empirisme, empirisme de la description de données attestées rassemblées dans des corpus plus ou moins construits, empirisme d'une modélisation en lien très étroit avec des applications spécialisées. Cette évolution a de quoi réjouir, parce qu'elle entraîne une prise en compte de l'hétérogénéité des réalisations linguistiques, hétérogénéité qui semble être une caractéristique importante de toute langue ; elle a peut-être aussi de quoi inquiéter, en raison de l'éparpillement des recherches, et du manque d'enracinement dans des modèles théoriques qui permettraient de "faire du sens", sur le plan de la langue, à partir des données observées.

Aarts, J. & Meijs, W., Eds. (1984). *Corpus linguistics I. Recent developments in the use of computer corpora in English language research.* Amsterdam: Rodopi.

Adam, J.-M. (1991). Cadre théorique d'une typologie séquentielle. *Etudes de linguistique appliquée*, (83), 7-18.

Adam, J.-M. & Revaz, F. (1989). Aspects de la Structuration du Texte Descriptif: Les Marqueurs d'Enumération et de Reformulation. *Langue Française*, (81), 59-98.

Aijmer, K. (1986). Why is *actually* so popular in spoken English? In G. Tottie & I. Bäcklund, Eds., *English in speech and writing: a symposium*, pp. 119-130. Stockholm: Almqvist & Wiksell.

Allwright, R.L., Woodley, M.-P. & Allwright, J.M. (1988). Investigating reformulation as a practical strategy for the teaching of academic writing. *Applied Linguistics*, **9**(3), 236-256.

Atlani, F. (1984). ON l'illusioniste. In A. Grésillon & J-L. Lebrave, Eds., *La langue au ras du texte*, pp. 13-30. Lille: Presses Universitaires de Lille.

Bateman, J. & Rondhuis, K.J. (1997). "Coherence Relations". Towards a General Specification. *Discourse Processes*, **24**(1), 3-50.

Beacco, J.-C., Ed. (1992). *Ethnolinguistique de l'écrit.* Langages 105. Paris: Larousse.

de Beaugrande, R. & Dressler, W. (1981). *Introduction to Text Linguistics.* London & New York: Longman.

Beekman, J. & Callow, J. (1974). *Translating the Word of God.* Grand Rapids, MI: Zondervan Publishing House.

Berri, J., Cartier, E., Desclés, J.-P., Jackiewicz, A. & Minel, J.-L. (1996). SAFIR, système automatique de filtrage de textes. In *Actes, TALN'96*.

Berthoud, A-C. & Mondada, L. (1991). Stratégies et marques d'introduction et de réintroduction d'un objet dans la conversation. *Bulletin CILA*, (54), 159-179.

Berthoud, A.-C. (1996). *Paroles à propos. Approche énonciative et interactive du topic.* Paris: Ophrys.

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

Biber, D. (1989). A typology of English texts. *Linguistics*, (27), 3-43.

Biber, D. (1990). Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing*, **5**, 257-269.

Biber, D. (1992). Using computer-based text corpora to analyze the referential stragegies of spoken and written texts. In J. Svartvik, Ed. *Directions in corpus linguistics*, pp. 213-252. Berlin, New York: Mouton de Gruyter.

Biber, D. (1993a). Representativeness in corpus design. *Literary and Linguistic Computing*, **8**, 243-257.

Biber, D. (1993b). Using register-diversified corpora for general language studies. *Computational Linguistics*, **19**(2), 219-241.

Biber, D. (1993c). Co-occurrence patterns among collocations: a tool for corpus-based lexical knowledge acquisition. *Computational Linguistics*, **19**(3), 531-538.

Biber, D. (1995). *Dimensions of register variation : A cross-linguistic comparison.* Cambridge: Cambridge University Press.

Biber, D. & Finegan, E. (1993). Intra-textual variation within medical research articles. In N. Oostdijk & P. de Haan, Eds., *Corpus-based research into language. In honour of Jan Aarts*, pp. 201-221. Amsterdam/Atlanta: Rodopi.

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 of Jan Aarts*, pp. 201-221. Amsterdam/Atlanta: Rodopi.

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

Blanche-Benveniste, C. (1996). De l'utilité du corpus linguistique. *Revue Française de Linguistique Appliquée*, **I**(2), 25-42.

Borkin, A. (1984). *Problems in Form and Function.* London: Ablex.

Brennan, S.E., Friedman, M.W. & Pollard, C.J. (1987). A Centering Approach to Pronouns. In *Actes, 25th Annual Meeting of the Association for Computational Linguistics*, pp. 155-162. Stanford, CA.

Brennan, S.E. (1998). Centering as a Psychological Resource for Achieving Joint Reference in Spontaneous Discourse. In M. Walker, A. Joshi & E. Prince, Eds., *Centering Theory in Discourse*, pp. 227-249. Oxford: Clarendon Press.

Brée, D.S. & Smit, R.A. (1986). Linking Propositions. In *Actes, COLING-86*, pp. 177-180.

Bronckart, J.-P., Bain, D., Schneuwly, B., Davaud, C. & Pasquier, A. (1985). *Le fonctionnement des discours.* Neuchâtel: Delachaux et Niestlé.

Candel, D. (1994). Le discours définitoire : variations discursives chez les scientifiques. In S. Moirand, M. Ali Bouacha, J.-C. Beacco & A. Collinot, Eds., *Parcours linguistiques de discours spécialisés*, Berne, Paris, New York: Peter Lang.

Cartier, E. (1997). La définition dans les textes scientifiques et techniques : présentation d'un outil d'extraction automatique des relations définitoires. In *Actes, TIA'97 (Terminologie et Intelligence Artificielle)*, Toulouse.

Chafe, W. (1970). *Meaning and the Structure of Language.* Chicago: University of Chicago Press.

Chafe, W.L. (1976). Givenness, Contrastiveness, Definiteness, Subjects, Topics and Point of View. In C. N. Li, Ed. *Subject and Topic*, New-York: Academic Press.

Charolles, M. (1978). Introduction aux problèmes de la cohérence des textes. *Langue Française*, **38**, 7-41.

Charolles, M. (1983). Coherence as a Principle in the Interpretation of Discourse. *Text*, **3**, 71-97.

Charolles, M. (1995). Cohésion, cohérence et pertinence du discours. *Travaux de Linguistique*, **29**(112), 125-151.

Charolles, M. (1997). *L'encadrement du discours : Univers, champs, domaines et espaces*. Cahier de Recherche Linguistique 6, Université de Nancy 2.

Charolles, M., Petöfi, J.S. & Sözer, E. (1986). *Research in Text Connexity and Text Coherence. A Survey.* Hamburg: Buske.

Church, K.W. & Mercer, R.L. (1993). Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, **19**(1), 1-24.

Clark, H.H. & Haviland, S.E. (1977). Comprehension and the given-new contract. In R. O. Freedle, Ed. *Discourse Production and Comprehension*, Norwood, New Jersey: Ablex.

Clyne, M. (1981). Culture and discourse structure. *Journal of Pragmatics*, **5**, 61-66.

Cohen, R. (1984). A Computational Theory of the Function of Clue Words in Argument Understanding. In *Actes, COLING-84*, pp. 251-258.

Combettes, B. (1983). *Pour une grammaire textuelle : la progression thématique.* Bruxelles: Duculot/De Back.

Combettes, B. (1998). *Les constructions détachées en français.* Paris: Ophrys.

Condamines, A., Fabre, C. & Péry-Woodley, M.-P., Eds. (1999). *Corpus et TAL : pour une réflexion méthodologique.* TALN'99. Cargese: ATALA.

Condamines, A. & Rebeyrolle, J. (à paraître). Searching for and identifying conceptual relationships via a corpus-based approach to Terminological Knowledge Bases (CTKB): methods and results. In D. Bourigault, M.-C. L'Homme & C. Jacquemin, Eds., *Recent Advances in Computational Terminology* , Amsterdam: John Benjamins.

Corbin, P. (1980). De la production des données en linguistique introspective. In A. M. Dessaux-Berthoneau, Ed. *Théories linguistiques et traditions grammaticales*, pp. 121-179. Lyon: Presses Universitaires de Lyon.

Cornish, F. (1986). *Anaphoric Relations in English and French. A Discourse Perspective.* London, Sydney, Dover New Hampshire: Croom Helm.

Cornish, F. (1998). The Functional Grammar conception of (discourse) anaphora: a sympathetic critique. In *Actes, 8t International Conference on Functional Grammar*, Amsterdam.

Cornish, F. (1999). *Anaphora, Discourse and Understanding. Evidence from English and French.* Oxford: Clarendon Press.

Cornish, F. (à paraître). L'accessibilité cognitive des référents, le centrage d'attention, et la structuration du discours : une vue d'ensemble. *Verbum*, **22**(1),

Dachelet, R. (1994). *Sur la notion de sous-langage*. Doctorat en sciences du langage, Université de Paris VIII.

Danes, F. (1966). A three-level approach to syntax. In F. Danes *et al.*, Ed. *Travaux linguistiques de Prague*, pp. 225-240. University of Alabama Press.

Danes, F. (1974). Functional Sentence Perspective and the Organization of Text: Different types of Thematic Progression. In F. Danes, Ed. *Papers on Functional Sentence Perspective*, The Hague: Mouton.

Daniel, M.P., Nicaud, L., Prince, V. & Péry-Woodley, M.-P. (1992). Apport du style linguistique à la modélisation cognitive d'un élève. In C. Frasson, G. Gauthier & G. I. McCalla, Eds., *Intelligent Tutoring Systems*, pp. 252-259. Berlin: Springer-Verlag.

Davison, A. (1984). Syntactic markedness and the definition of sentence topic. *Language*, **60**(4), 797-846.

van Dijk, T.A. (1977). *Text and Context. Explorations in the Semantics and Pragmatics of Discourse.* London, New York: Longman.

van Dijk, T.A. (1981). *Studies in the Pragmatics of Discourse.* The Hague, Paris, New-York: Mouton.

Dik, S.C. (1997). *The Theory of Functional Grammar.* Berlin: de Gruyter.

Ducrot, O. (1983a). *Puisque* : essai de description polyphonique. *Revue Romane*, **24**, 166-185.

Ducrot, O. (1983b). Opérateurs argumentatifs et visée argumentative. *Cahiers de Linguistique Française*, **5**, 7-36.

Ducrot, O. *et al.* (1980). *Les mots du discours.* Paris: Editions de Minuit.

Enkvist, N.E. (1985). A Parametric View of Word Order. In E. Sözer, Ed. *Text Connexity, Text Coherence*, pp. 320-336. Hamburg: Helmut Buske.

Fillmore, C.J. (1992). "Corpus linguistics" or "Computer-aided armchair linguistics". In J. Svartvik, Ed. *Directions in corpus linguistics: Proceedings of Nobel Symposium 82* , pp. 35-60. Berlin, New York: Mouton de Gruyter.

Firbas, J. (1964). On defining the theme in functional sentence analysis. *Travaux Linguistiques de Prague*, **1**, 267-280.

Firbas, J. (1972). On the Interplay of Prosodic and Non-Prosodic Means of Functional Sentence Perspective. In U. Fried, Ed. *The Prague School of Linguistics and Language Teaching*, pp. 77-94. London: Oxford University Press.

Firbas, J. (1986). Thoughts on Functional Sentence Perspective, intonation and emotiveness. *Brno Studies in English*, **16**, 11-48.

Flowerdew, J.L. (1992). Salience in the performance of one speech act: the case of definitions. *Discourse Processes*, **15**, 165-181.

Fradin, B. (1990). Approche des constructions à détachement. Inventaire. *Revue Romane*, **52**(1), 3-34.

Francis, W.N. & Kucera, H. (1982). *Frequency analysis of English usage: lexicon and grammar.* Boston: Houghton Mifflin.

Givón, T. (1979). *On Understanding Grammar.* New-York, San Francisco, London: Academic Press.

Givón, T. (1983). Topic continuity in discourse: An introduction. In T. Givón, Ed. *Topic continuity in discourse*, Amsterdam/Philadelphia: John Benjamins.

Goody, J. (1977). *The Domestication of the Savage Mind.* Cambridge: Cambridge University Press.

Gordon, P.C., Grosz, B.J. & Gilliom, L.A. (1993). Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, **17**, 311-347.

Grice, H.P. (1975). Logic and conversation. In P. Cole & J. Morgan, Eds., *Syntax and Semantics 3: Speech Acts*, New York: Academic Press.

Grimes, J.E. (1975). *The Thread of Discourse.* The Hague: Mouton.

Grishman, R. & Kittredge, R., Eds. (1986). *Analyzing language in restricted domains. Sublanguage description and processing.* Hillsdale, N.J.: Laurence Erlbaum.

Grosz, B.J. (1977). The representation and use of focus in a system for understanding dialogs. In *Actes, 5th International Joint Conference on Artificial Intelligence*, Cambridge, MA.

Grosz, B.J. & Sidner, C.L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, **12**(3), 175-204.

Grosz, B.J., Joshi, A. & Weinstein, S. (1995). Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, **21**(2), 203-225.

Grosz, B.J. & Ziv, Y. (1998). Centering, global focus, and right-dislocation. In M. Walker, A. Joshi & E. Prince, Eds., *Centering Theory in Discourse*, pp. 293-308. Oxford: Clarendon Press.

Gundel, J.K. (1985). Shared knowledge' and topicality. *Journal of Pragmatics*, **9**, 83-107.

Gundel, J.K. (à paraître). Statut cognitif et formes des anaphoriques indirects. *Verbum*, **22**(1).

Gundel, J.K. (1988). Universals of topic-comment structure. In M. Hammond, E. Moravcsik & J. Wirth, Eds., *Studies in Linguistic Typology*, pp. 209-239. Amsterdam/Philadelphia: John Benjamins.

Gundel, J.K., Hedberg, N. & Zacharski, R. (1993). Cognitive Status and the Form of Referring Expressions. *Language*, **69**, 274-307.

Gundel, J.K. (1998). Centering Theory and the Givenness Hierarchy: Towards a Synthesis. In M. Walker, A. Joshi & E. Prince, Eds., *Centering Theory in Discourse*, pp. 183-198. Oxford: Clarendon Press.

Habert, B. (1985). Etudes des formes "spécifiques" et typologie des énoncés (les résolutions générales des congrès de la CFTC-CFDT de 1945 à 1979). *Mots*, **11**, 127-154.

Habert, B. (1995). Introduction. *TAL*, **36**(1-2), 3-6. Traitements probabilistes et corpus, Benoît Habert, (resp.).

Habert, B. (1998). *Un corpus clé pour le français actuel.* http://www.biomath.jussieu.fr/CLEF/

Habert, B. & Jacquemin, C. (1993). Noms composés, termes, dénominations complexes : problématiques linguistiques et traitements automatiques. *TAL*, **34**(2), 5-42.

Habert, B., Barbaud, P., Dupuis, F. & Jacquemin, C. (1995). Simplifier des arbres d'analyse pour dégager les comportements syntactico-sémantiques des formes d'un corpus. *Cahiers de Grammaire*, (20), 1-32.

Habert, B. & Salem, A. (1995). L'utilisation de catégories multiples pour l'analyse quantitative de données textuelles. *TAL*, **36**(1-2), 249-276.

Habert, B., Nazarenko, A. & Salem, A. (1997). *Les linguistiques de corpus*. Paris: Armand Colin.

Habert, B., Fabre, C. & Issac, F. (1998). *De l'écrit au numérique. Constituer, normaliser et exploiter les corpus électroniques.* Paris: Masson.

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

Hahn, U. & Strube, M. (1997). Centering in-the-large: Computing referential discourse segments. In *Actes, 35th Annual Meeting of the ACL*, pp. 104-111..

Halliday, M.A.K. (1967a). Notes on Transitivity and Theme in English. Part 1. *Journal of Linguistics*, **3**(1), 37-81.

Halliday, M.A.K. (1967b). Notes on Transitivity and Theme in English. Part 2. *Journal of Linguistics*, **3**(2), 199-244.

Halliday, M.A.K. (1968). Notes on Transitivity and Theme in English. Part 3. *Journal of Linguistics*, **4**(2), 179-215.

Halliday, M.A.K. (1980). Text semantics and clause grammar: some patterns of realization. In J. E. Copeland & P. Davis, Eds., *The 7th LACUS Forum*, Columbia, S.C.: Hornbeam Press.

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

Halliday, M.A.K. (1991). Corpus studies and probabilistic grammar. In K. Aijmer & B. Altenberg, Eds., *English corpus linguistics. Studies in honour of Jan Svartvik*, pp. 30-43. London: Longman.

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

Harris, Z.S. (1968). *Mathematical Structures of Language.* New York: Wiley & Sons.

Harris, Z.S. (1982). *A grammar of English on mathematical principles.* New York: Wiley-Interscience.

Harris, Z.S. (1988). *Language and information.* New York: Columbia University Press.

Harris, Z.S. (1991). *A theory of language and information; A mathematical approach.* Oxford: Clarendon Press.

Harris, Z., Gottfried, M., Ryckman, T., Mattick Jr, P., Daladier, A., Harris, T.N.*, et al.* (1989). *The form of information in science. Analysis of an immunology sublanguage.* Dordrecht: Kluwer Acdemic Publishers.

Hathout, N. (1996). Pour la construction d'une base de connaissances lexicographiques à partir du Trésor de la Langue Française : Les marqueurs superficiels dans les définitions spécialisées. *Cahiers de Lexicologie*, **68**(1), 137-173.

Hearst, M.A. (1998). Automated discovery of WordNet relations. In C. Fellbaum, Ed. *WordNet: an electronic lexical database* , pp. 131-151. Boston: MIT Press.

Hinds, J. (1983). Contrastive Rhetoric: Japanese and English. *Text*, **3**, 183-196.

Hitzeman, J. & Poesio, M. (1998). Long distance pronominalisation and global focus. In *Actes, COLING-ACL'98*, Montreal, ACL.

Hobbs, J.R. (1985). *On the Coherence and Structure of Discourse*. CSLI-85-37, Center for the Study of Language and Information.

Hoey, M. (1983). *On the Surface of Discourse.* London: Allen & Unwin.

Hovy, E. (1990). Parsimonious and profligate approaches to the question of discourse structure relations. In *Actes, 5th International Workshop on Language Generation*, Pittsburgh, PA.

Hunt, K.W. (1965). *Grammatical structures written at three grade levels*. NCTE Research Report 3, NCTE.

Hunt, K.W. (1970). *Syntactic maturity in schoolchildren and adults.* Chicago: University of Chicago Press.

Hurewitz, F. (1998). A quantitative look at discourse coherence. In M. Walker, A. Joshi & E. Prince, Eds., *Centering Theory in Discourse*, pp. 273-292. Oxford: Clarendon Press.

Illouz, G. (1999). Méta-étiqueteur adaptatif : vers une utilisation pragmatique des ressources linguistiques. In *Actes, TALN'99 (Traitement Automatique des Langues Naturelles)*, pp. 185-194. Cargese, ATALA.

Illouz, G., Habert, B., Fleury, S., Folch, H., Heiden, S. & Lafon, P. (1999). Maîtriser les déluges de données hétérogènes. In A. Condamines, C. Fabre & M.-P. Péry-Woodley, Eds., *Actes, Corpus et TAL : pour une réflexion méthodologique. TALN'99 (Traitement Automatique des Langues Naturelles)*, pp. 37-46. Cargese.

Jacquemin, C. & Bush, C. (2000). Combining lexical and formatting cues for named entity acquisition from the Web. In *Actes, COLING'2000* (soumis).

Kameyama, M. (1998). Intrasentential Centering: A Case Study. In M. Walker, A. Joshi & E. Prince, Eds., *Centering Theory in Discourse*, pp. 89-114. Oxford: Clarendon Press.

Kass, R. (1989). Student Modeling in Intelligent Tutoring Systems. In A. Kobsa & W. Wahlster, Eds., *User Models in Dialog Systems*, Berlin, Heidelberg, New York: Springer Verlag.

Keenan, E.L. (1985). Passive in the world's languages. In T. Shopen, Ed. *Language Typology and Syntactic Description*, Cambridge: Cambridge University Press.

Kieras, D.E. (1981). The role of major referents and sentence topics in the construction of passage macrostructure. *Discourse Processes*, (4), 1-15.

Kobsa, A. & Wahlster, W. (1988). Preface. *Computational Linguistics. Special Issue on User Modeling*, **14**(3), 1-4.

Lakoff, R. (1984). The Pragmatics of Subordination. In *Actes, 10th Annual Meeting of the Berkeley Linguistics Society*, pp. 481-492. Berkeley, CA, Berkeley Linguistics Society.

Lambrecht, K. (1981). *Topic, antitopic and verb-agreement in non-standard French.* Amsterdam/Philadelphia: John Benjamins.

Lambrecht, K. (1982). *Discourse Pragmatics.* University of California, Berkeley.

Lambrecht, K. (1987). On the status of SVO sentences in French discourse. In R. Tomlin, Ed. *Coherence and grounding in discourse*, pp. 217-262. Amsterdam/Philadelphia: John Benjamins.

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

Lambrecht, K. (1998). Sur la relation formelle et fonctionnelle entre topiques et vocatifs. *Langues*, **1**(1), 34-45.

Lambrecht, K. (à paraître). Dislocation. In M. Haspelmath, Ed. *Language Typology and Language Universals*, Berlin, New York: Walter de Gruyter.

Landelle, M. (1988). *Analyse syntaxique de l'expression de la segmentation : Approche linguistique pour un traitement informatique des structures textuelles*. DEA de linguistique, Université de Toulouse 2.

Lautamatti, L. (1978). Some Observations on Cohesion and Coherence in Simplified Texts. In J. O. Östman, Ed. *Cohesion and Semantics*, pp. 165-181. Åbo, Finland: Research Institute of the Åbo Akademi Foundation.

Lautamatti, L. (1987). Observations on the development of the topic of simplified discourse. In U. Connor & R. B. Kaplan, Eds., *Writing across Languages: Analysis of L2 Text*, pp. 87-114. Reading, MA: Addison-Wesley.

Leech, G. (1991). The state of the art in corpus linguistics. In K. Aijmer & B. Altenberg, Eds., *English corpus linguistics,* pp. 8-29. London: Longman.

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

Longacre, R.E. (1976). *An Anatomy of Speech Notions.* Lisse: The Peter de Ridder Press.

Longacre, R.E. (1979). The paragraph as a grammatical unit. In T. Givón, Ed. *Discourse and Syntax*, pp. 115-134. New York, London: Academic Press.

Luc, C. (1998a). Contraintes sur l'architecture textuelle. *Document Numérique*, **2**(2), 203-219.

Luc, C. (1998b). Types de contraintes architecturales sur la composition d'objets textuels. In *Actes, CIDE'98 (Colloque International sur le Document Électronique)*, pp. 15-30. Rabat, Maroc.

Luc, C., Mojahid, M. & Virbel, J. (1999). Connaissances structurelles et modèles nécessaires à la génération de textes formatés. In *Actes, GAT'99 (Génération Automatique de Textes)*, pp. 157-170. Grenoble.

Luc, C., Mojahid, M., Virbel, J., Garcia-Debanc, C. & Péry-Woodley, M.-P. (1999). A linguistic approach to some parameters of layout: A study of enumerations. In R. Power & D. Scott, Eds., *Actes, AAAI 1999 Fall Symposium: "Using Layout for the Generation, Understanding or Retrieval of Documents"*, pp. 20-29. North Falmouth, Massachussetts.

Luc, C., Mojahid, M., Péry-Woodley, M.-P. & Virbel, J. (2000). Les énumérations : structures visuelles, syntaxiques et rhétoriques. In *Actes, CIDE'2000 (Colloque International sur le Document Électronique)*, Lyon, France.

Maier, E. & Hovy, E. (1993). Organising discourse structure relations using metafunctions. In H. Horacek & M. Zock, Eds., *New Concepts in Natural Language Generation*, pp. 69-86. London: Pinter.

Mann, W.C. & Thompson, S.A. (1986). Relational propositions in discourse. *Discourse Processes*, **9**(1), 57-90.

Mann, W.C. & Thompson, S.A. (1987). Rhetorical structure theory: a theory of text organization. In L. Polanyi, Ed. *The Structure of Discourse*, Norwood, N.J.: Ablex.

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

Mann, W.C. & Thompson, S.A., Eds. (1992). *Discourse description. Diverse linguistic analyses of a fund-raising text.* Pragmatics and Beyond. Amsterdam, Philadelphia: John Benjamins.

Mann, W.C., Matthiessen, C. & Thompson, S.A. (1989). *Rhetorical structure theory and text analysis*. ISI/RR-89-242, Information Sciences Institute.

Martin, J.R. (1983). Conjunction: the Logic of English Text. In J. R. Martin, Ed. *Papers in Text Linguistics, Vol. 45*, pp. 1-72. Hamburg: Helmut Buske Verlag.

Martin, R. (1983). *La logique du sens.* Paris: PUF.

Martin, R. (1990). La définition "naturelle". In J. Chaurand & F. Mazière, Eds., *La définition*, pp. 86-96. Paris: Larousse.

Matthiessen, C. & Thompson, S.A. (1988). The structure of discourse and "subordination". In J. Haiman & S. A. Thompson, Eds., *Clause Combining in Grammar and Discourse*, Amsterdam: Benjamins.

McKeown, K. (1985). *Text Generation: Using discourse strategies and focus constraints to generate natural language text.* Cambridge UK: Cambridge University Press.

McNaught, J. (1993). User needs for textual corpora in natural language processing. *Literary & Linguistics Computing*, **8**(4), 227-234.

Montgomery, C.A. & Glover, B.C. (1986). A sublanguage for reporting and analysis of space events. In R. Grishman & R. Kittredge, Eds., *Analysing language in restricted domains. Sublanguage description and processing*, pp. 129-162. Hillsdale, N.J.: Lawrence Erlbaum.

Morin, E. (1999). Using lexico-syntactic patterns to extract semantic relations between terms form a technical corpus. In P. Sandrini, Ed. *Actes, 5th International Congress on Terminology and Knowledge Engineering (TKE'99)*, pp. 268-278. Innsbrück, Austria.

Nicaud, L. & Prince, V. (1990). TEDDI : An ITS for definitions learning. In *Actes, PRICAI'90*.

Nunberg, G. (1990). *The Linguistics of Punctuation.* Menlo Park: Center for the Study of Language and Information.

O'Brien, T. (1995). Rhetorical Structure Analysis and the case of the inaccurate incoherent source-hopper. *Applied Linguistics*, **16**(4), 442-482,

Ong, W.J. (1982). *Orality and Literacy: The Technologizing of the Word.* London: Methuen.

Pascual, E. (1991). *Représentation de l'architecture textuelle et génération de texte*. Doctorat d'informatique, Université Paul Sabatier, Toulouse.

Pascual, E. & Péry-Woodley, M.-P. (1995). La définition dans le texte. In J.-L. Nespoulous & J. Virbel, Eds., *Textes de type consigne – Perception, action, cognition*, pp. 65-88. Toulouse: PRESCOT.

Pascual, E. & Péry-Woodley, M.-P. (1997a). Définition et action dans les textes procéduraux. In E. Pascual, J.-L. Nespoulous & J. Virbel, Eds., *Le texte procédural : langage, action et cognition*, pp. 223-248. Toulouse: PRESCOT.

Pascual, E. & Péry-Woodley, M.-P. (1997b). Modèles de texte pour la définition. In *Actes, Premières Journées Scientifiques et Techniques du Réseau francophone de l'Ingénierie de la Langue de l'AUPELF-UREF*, pp. 137-146. Avignon, AUPELF-UREF.

Pascual, E. & Péry-Woodley, M.-P. (1997c). Modélisation des définitions dans les textes à consignes. In J. Virbel, J.-M. Cellier & J.-L. Nespoulous, Eds., *Cognition, Discours procédural, Action*, pp. 37-55. Toulouse: PRESCOT.

Passonneau, R.J. (1998). Interaction of Discourse Structure with Explicitness of Discourse Anaphoric Noun Phrases. In M. Walker, A. Joshi & E. Prince, Eds., *Centering Theory in Discourse*, pp. 327-358. Oxford: Clarendon Press.

Péry-Woodley, M.-P. (1989). *Textual designs: Signalling coherence in first and second language academic writing*. Doctorat (Ph.D.) de linguistique, Université de Lancaster. (édité en 1991 comme Notes et Documents LIMSI 91-1, CNRS/Université Paris XI)

Péry-Woodley, M.-P. (1990a). Contrasting discourses: contrastive analysis and a discourse approach to writing. *Language Teaching*, **23**(3), 143-151.

Péry-Woodley, M.-P. (1990b). De la langue aux discours: recherches sur l'analyse des textes d'apprenants. In B. Schneuwly & J.-P. Bronckart, Eds., *Diversifier l'enseignement du français écrit (IVe Colloque International de Didactique du Français Langue Maternelle)*, pp. 329-335. Neuchâtel, Paris: Delachaux & Niestlé.

Péry-Woodley, M.-P. (1991a). French and English passives in the construction of text. *Journal of French Language Studies*, **1**(1), 55-70.

Péry-Woodley, M.-P. (1991b). Writing in L1 and L2: analysing and evaluating learners' texts. *Language Teaching*, **24**(2), 69-83.

Péry-Woodley, M.-P. (1993a). *Les écrits dans l'apprentissage. Clés pour analyser les productions des apprenants.* F Références. Paris: Hachette.

Péry-Woodley, M.-P. (1993b). *Textual clues for user modelling in an intelligent tutoring system*. Notes et Documents LIMSI 93-21, CNRS/Université Paris XI.

Péry-Woodley, M.-P. (1994). Une pragmatique à fleur de texte: marques superficielles des opérations de mise en texte. In S. Moirand, A. Ali Bouacha, J.-C. Beacco & A. Collinot, Eds., *Parcours linguistiques de discours spécialisés*, pp. 337-348. Berne: Peter Lang.

Péry-Woodley, M.-P. (1995). Quels corpus pour quels traitements automatiques? *TAL*, **36**(1-2), 213-232.

Péry-Woodley, M.-P. (1998). Signalling in written text: a corpus-based approach. In M. Stede, L. Wanner & E. Hovy, Eds., *Actes, COLING 98 (Workshop on Discourse Relations and Discourse Markers)*, pp. 79-85. Montreal, ACL.

Péry-Woodley, M.-P. (à paraître). *Cadrer* ou *centrer* son discours ? Introducteurs de cadres et centrage. *Verbum*, **22**(1).

Péry-Woodley, M.-P. (1996b). Syntaxe et discours : l'expression du but. In S. T. Paribakht, H. Seguin, M.-C. Tréville & R. Williamson, Eds., *L'expression orale et écrite : recherche, enseignement, technologie*, pp. 20-25. Ottawa: University of Ottawa.

Péry-Woodley, M.-P. & Rebeyrolle, J. (1998). Domain and genre in sublanguage text: Definitional microtexts in three corpora. In A. Rubio, N. Gallardo, R. Castro & A. Tejada, Eds., *Actes, First International Conference on Language Resources and Evaluation*, pp. 997-992. Granada, ELRA.

Prince, E.F. (1981). Toward a Taxonomy of Given-New Information. In P. Cole, Ed. *Radical Pragmatics*, pp. 223-255. New-York: Academic Press.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language.* London & New York: Longman.

Rebeyrolle, J. & Péry-Woodley, M.-P. (1998). Repérage d'objets textuels fonctionnels pour le filtrage d'information: le cas de la définition. In *Actes, Rencontre Internationale sur l'Extraction, le Filtrage et le Résumé Automatiques*, pp. 19-30. Sfax, Tunisie.

Rebeyrolle, J. (à paraître). Utilisation de contextes définitoires pour l'acquisition de connaissances à partir de textes. In *Actes, Ingénierie des Connaissances (IC'2000)*, Toulouse.

Redeker, G. (1990). Ideational and pragmatic markers of discourse structure. *Journal of Pragmatics*, (14), 367-381.

Reichman, R. (1984). Extended person-machine interface. *Artificial Intelligence*, **22**, 157-218.

Reinhart, T. (1980). Conditions for Text Coherence. *Poetics Today*, **1**, 161-180.

Régent, O. (1985). A comparative approach to the learning of specialized written discourse. In P. Riley, Ed. *Discourse and Learning*, London, New York: Longman.

Régent, O. (1992). Pratiques de communication en médecine: contextes anglais et français. *Langages*, **26**(105), 66-75.

Riegel, M. & Tamba, I. (1987). Définition directe et indirecte dans le langage ordinaire : les énoncés définitoires copulatifs. *Langue Française*, (73), 29-53.

Riegel, M. (1990). La définition, acte du langage ordinaire - De la forme aux interprétations. In J. Chaurand & F. Mazière, Eds., *La définition*, pp. 97-110. Paris: Larousse.

Rodrigues Faria Coracini, M-J. (1992). L'hétérogénéité dans le discours scientifique français et brésilien: un effet persuasif. *Langages*, (105), 76-86.

Roulet, E., Auchlin, A., Moeschler, J., Rubattel, C. & Schelling, M. (1985). *L'articulation du discours en français contemporain.* Berne: Peter Lang.

Ryckman, T. (1990). De la structure d'une langue aux structures de l'information dans le discours et dans les sous-langages scientifiques. *Langages*, (99), 21-28.

Sager, N. (1986). Sublanguage: Linguistic phenomenon, computational tool. In R. Grishman & R. Kittredge, Eds., *Analyzing language in restricted domains. Sublanguage description and processing*, pp. 1-18. Hillsdale, N.J.: Laurence Erlbaum.

Sager, N., Friedman, C. & Lyman, M.S., Eds. (1987). *Medical language processing. Computer management of narrative data.* Reading, MA.: Addison-Wesley.

Schiffrin, D. (1981). Tense variation in narrative. *Language*, **57**, 45-62.

Schiffrin, D. (1987). *Discourse markers.* Cambridge: Cambridge University Press.

Shopen, T. & Williams, J.M. (1981). *Styles and Variables in English.* Cambridge, MA: Winthrop.

Sinclair, J., Ed. (1987). *Looking Up: An Account of the Cobuild Project in Lexical Computing.* London: Collins.

Sinclair, J., Hanks, P., Fox, G., Moon, R. & Stock, P., Eds. (1987). *Collins COBUILD English Language Dictionary.* Glasgow: Collins.

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Sleeman, D. & Brown, J.S. (1982). *Intelligent Tutoring Systems.* New York: Academic Press.

Smith, N. (1997). Improving a Tagger. In R. Garside, G. Leech & A. McEnery, Eds., *Corpus Annotation: Linguistic Information from Computer Text Corpora*, pp. 137-150. London: Addison Wesley.

Stenström, A.-B. (1986). What does *really* really do? In G. Tottie & I. Bäcklund, Eds., *English in speech and writing: a symposium*, pp. 149-164. Stockholm: Almqvist and Wiksell.

Strawson, P. (1971). *Logico-linguistic Papers.* London: Methuen.

Sueur, J.-P. (1982). Pour une grammaire du discours. *Mots*, (5), 143-185.

Swales, J. (1990). *Genre analysis: English in academic and research settings.* Cambridge: Cambridge University Press.

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

Virbel, J. (1985). Langage et métalangage dans le texte du point de vue de l'édition en informtique textuelle. *Cahiers de Grammaire*, **10**, 5-72.

Virbel, J. (1989). The Contribution of Linguistic Knowledge to the Interpretation of Text Structures. In J. André, V. Quint & R. K. Furuta, Eds., *Structured Documents*, pp. 161-181. Cambridge: CUP.

Virbel, J. (1999). *Structures textuelles - planches, fascicule 1 : Énumérations*. Rapport IRIT, Toulouse.

Walker, M.A. (à paraître). Vers un modèle de l'interaction du centrage avec la structure globale du discours. *Verbum*, **22**(1),

Walker, M.A., Iida, M. & Cote, S. (1994). Japanese Discourse and the Process of Centering. *Computational Linguistics*, **20**(2), 193-232.

Walker, M., Joshi, A. & Prince, E. (1998). Centering in Naturally Occurring Discourse: An Overview. In M. Walker, A. Joshi & E. Prince, Eds., *Centering Theory in Discourse*, pp. 1-28. Oxford: Clarendon Press.

Werth, P. (1984). *Focus, Coherence and Emphasis*. Dover, New Hampshire: Croom Helm.

Annexes

1. Corpus Étudiants : textes dont sont extraits les exemples de la Partie I :

ALM-2

[1] Using the statistics, I intend to explain the phenomena over which there has been much controversy about higher education in Great Britain over recent months, 2] and I will go on to suggest in which areas higher education is expanding and contracting, and give my own opinion as to the direction which higher education should be taking.

[3] Table one shows the rate of scholarization of the population aged between nineteen and twenty-four years. [4] Compared with the United States and Italy, which have twenty-six point seven and nineteen point two percent repectively of the population engaged in higher education, Great Britain's percentage of its population which participates in higher education is only nine point three. [5] In fact, of all the countries mentioned, it is Great Britain which has the lowest percentage of its population aged beween nineteen and twenty-four years engaged in higher education.

[6] The average age at which pupils enter and leave the education system is represented in the figures of table two. [7] The United States' figures represent the longest time for pupils to stay in the education system: [8] that is sixteen point seven years on average are spent by a pupil at school. [9] However, Great Britain's figures show that British pupils spend the shortest time in the education system, that is only thirteen point two years. [10] So it is again the British education system which seems to be contracting the most compared with the other countries.

[11] The average rate of annual increase of public expenditure on higher education is shown in table three. [12] The figures clearly reveal that the British education system suffers the most in relation to the annual rate of increase of public expenditure, the British figure being only eight point twelve percent compared with Sweden and France, for example, whose governments spend an annual increase of twenty-eight point nine and twenty-four point seven percent on higher education respectively. [13] Again it is the British education system which is contracting the most in view of the amount of money consecrated to it.

[14] Finally, the figures in table four show the percentage of Gross National Product which is consecrated to public expenditure in the field of higher education. [15] The statistics clearly show that in countries such as the United States and Sweden, there has been a great change in the percentage of GNP between 1965 and 1974, whereas in Great Britain there has been very little change over these nine years as regards the percentage of GNP consecrated to public expenses for the sake of the higher education system, namely from nought point six-four percent in 1965 to nought point eight-nine percent in 1974, which is only a nought point two-five percent increase.

[16] It is clear from the figures shown in the four tables that the British education system is suffering the most in relation to the education systems in other countries. [17] There has been much controversy particularly in recent months over the future of higher education in Great Britain, [18] and one only has to gaze at the figures to wonder why.

[19] Although there has been an increase in the percentage of GNP which is consecrated for public expenses for higher education, and although the average rate of annual increase of public expenses for higher education is eight point twelve percent, the increases are only slight when compared with the other countries. [20] There is no doubt then as to whether higher education is expanding or contracting in Great Britain, [21] for it is clear that the system is contracting to reduce its cost and the number of students.

ALM-3

[1] The future of higher education at present hangs in the balance because of cuts in governement expenditure. [2] It is certainly a question of considerable controversy whether government policy to reduce the amount of money reserved for higher education is a sound one, or if, in the long run, it is going to be to the detriment of the country.

[3] In my opinion the standards of higher education will utlimately determine the future of the country, [4] for it is the educated youth of today who will be the leaders of the country tomorrow. [5] The obvious path for higher education to take is one of expansion, [6] and an attempt to improve standards is essential if Britain is to continue to compete with other leading countries in the spheres of education, trade, science, and so on. [7] In 1976, Britain had the smallest percentage of the 19 to 24 age group in higher education, in a survey taken in 7 countries. [8] At the same time, Britain has the smallest growth rate of expenditure on higher education, although, as a percentage of her Gross National Product, she has a relatively high expenditure. [9] All signs indicate, however, that Britain is being left behind with regard to higher education.

[10] At a time of high unemployment, an obvious alternative to the dole queues is to take a course in higher education after leaving school. [11] Extra qualifications lead to improved chances in one's chosen profession, and in theory at least, bring about more rapid promotion. [12] Why then should today's school leavers be deprived of the opportunity to further their education? [13] In a developed country such as Britain, we accept that it is our right to be able to attend institutions of higher education. [14] If the government cuts continue, however, we shall find ourselves in a situation where only an elite few can benefit from these institutions. [15] Entrance requirements will automatically be raised, so that only the academically brilliant will succeed in obtaining a university or college place. [16] Is it really such a bad thing? [17] I hear you ask. [18] Naturally we want our academic standards to be maintained, if not improved, [19] but at a time when there is population growth, so more children are staying on at school to take Alevels, and so find themselves eligible for entance into higher education, it is wrong to reduce the number of places available.

[20] Of course, we must accept that at a time of economic recession, such as Britain is at present experiencing, there is less money available to be spent on higher education. [21] If cuts must be made, then, the government must examine carefully the areas which will be affected, not only in terms of numbers of students, but also in terms of advancement in technology, and up-to-date information, which may be held back if the amount of money available to higher education institutions is limited. [22] Expansions should be made in areas of education which are going to directly influence the progress of the country, not only in the sciences, but also in arts and humanities courses which are going to aid the country if it is going to continue to compete with the rest of the world. [23] The governement could revise the higher education system, by making many course more relevant to "the outside world', by including practical experience in more courses, by introducing more courses in which students study more than one subject, and where the subjects complement each other and will be of benefit to both the student, and to his future employer.

[24] The question of higher education is obviously a difficult one for the government to respond to in a manner to suit everyone. [25] However, it seems to be equally obvious that attempts should be made to expand, both in numbers of students, because more school-leavers are today academically capable of pursuing a higher education course than in the past – [26] there is more motivation, more parental encouragement to name but two reasons for attending an institute of higher education, and in the amount of money given to higher education, in order to improve standards and offer better opportunities.

[27] An investment in higher education is an investement in the future.

ALM-12

Higher education

[1] The importance of higher education in our modern society is an extremely complex problem forming part of the serious economic difficulties which Great Britain is at present encountering. [2] It is argued that in these times of unceasing progress and ever-expanding technology, education at a higher level and especially scientific teaching is essential to cater for the increasing demands of modern life.

[3] In order to promote and maintain this modern lifestyle many people feel that higher education in Great Britain must continue to expand. [4] However one could also say that it is important to realise the intrinsic value of education in its development and expansion of the character as well as the mind of an individual. [5] Education is a preparation for the future [6] and the necessity for intelligent, educated people in any society is undisputed. [7] It could also be said that education is a legal right which must be fought for in order for it to be upheld. [8] In a free capitalist society such as our own the aesthetic value of education must be appreciated. [9] Everyone should be entitled, if they wish, to do something which they enjoy. [10] This therefore implies that it is not only those technological and scientific courses in higher education which must be maintained because of their obvious necessity in an expanding age, but also those courses such as art, music and languages whose future is threatened merely because they have no obvious function. [11] This problem also raises the issue of unemployment as an argument for the continued expansion of higher education. [12] Many people have said that the integration of many young people into higher education leads to a reduction in the unemployment rate and is also advantageous in that it leads to the development of a more intelligent individual who is better able to eventually find employment.

[13] However it is argued by many that there is an inevitable need to reduce the rate of expenditure on higher education. [14] In the bleak economic climate of today it must be admitted that many young people graduating from higher education are still unable to find employment. [15] It is possible that their skills are too specialised and are superfluous to the needs of our modern age. [16] Within the past few years governement expenditure on universities and colleges has gradually been cut back to try to combat the ever-increasing rate of inflation. [17] It has been said that these cuts are essential as the percentage of young people in higher education is very low in proportion to the vast amounts of money poured into it each year. [18] This money would perhaps be put to better use in catering for the majority and for example improving public services such as the health service.

[19] In conclusion the current standard of education must be maintained at a level.

FLE-2

[1] On n'a qu'à étudier les statistiques pour voir qu'il est évident que les opinions divergent de nos jours sur la direction que devrait prendre l'enseignement supérieur.

[2] Selon les données O.C.D.E. des années soixante-dix, de différents pays ont des tendances bien différentes en ce qui concerne l'enseignement supérieur, soit dans le domaine de l'âge moyen à l'entrée et à la sortie du système d'enseignement, soit dans le domaine du

taux de scolarisation de la population âgée de dix-neuf à vingt-quatre ans. [3] Il est clair, par exemple, qu'en Amérique 26,7 pour cent des gens âgés de dix-neuf à vingt-quatre ans participent dans l'enseignement supérieur, tandis que seulement 9,3 pour cent des gens en Grande-Bretagne qui ont le même âge se trouvent dans l'enseignement supérieur. [4] Il y a beaucoup d'autres exemples qui montrent que le système de l'enseignement supérieur varie de pays en pays, [5] et en observant les tendances de la Grande-Bretagne il s'agit de voir s'il y a toujours une expansion ou une réduction des crédits et des effectifs en tenant compte des statistiques des autres pays.

[6] D'abord, le taux de scolarisation de la population âgée de dix-neuf à vingt-quatre ans pour la Grande-Bretagne est de 9,3 pour cent, [7] et c'est un taux assez bas si l'on considère les statistiques pour les autres pays; [8] par exemple aux Etats-Unis, 26,7 pour cent de la population âgée de dix-neuf à vingt-quatre ans est scolarisée, [9] et pour l'Italie, 19,2 pour cent est la statistique pour le même aspect de l'enseignement supérieur. [10] Selon ces statistiques, c'est la Grande-Bretagne qui a le moins de gens âgés de dix-neuf à vingt-quatre ans dans l'enseignement supérieur.

[11] Quand on considère l'aspect de l'âge moyen à l'entrée et à la sortie du système d'enseignement, il est évident que les élèves américains passent le plus de temps à l'école, de 4,8 ans à 21,5 ans, et que dans la Grande-Bretagne seulement 13,2 années sont passées à l'école.

[12] En ce qui concerne le taux moyen d'accroissement annuel des dépenses publiques relatives à l'enseignement supérieur public de 1965 à 1970, il est à voir que ce sont la Suède et la France qui dépensent le plus d'argent pour l'enseignement supérieur et que c'est la Grande-Bretagne qui dépense seulement 8,12 pour cent d'argent par an pour l'enseignement supérieur public.

[13] En considérant le pourcentage du P.N.B. consacré aux dépenses publiques dans l'enseignement supérieur public, il est intéressant de voir que de 1965 à 1974, il y a eu un grand changement aux Etats-Unis en ce qui concerne le pourcentage du P.N.B. - de 0,90 à 1,62, [14] mais en Grande-Bretagne il y a eu peu de changement par rapport aux autres pays, seulement de 0,64 à 0,89 pour cent.

[15] Tout bien pesé, il est évident que la Grande-Bretagne est en train de réduire les crédits de l'enseignement supérieur, [16] et par rapport aux autres pays c'est la Grande-Bretagne qui consacre le moins de dépenses et d'efforts en essayant d'améliorer le système de l'enseignement supérieur public.

FLE-6

[1] En ce moment, le problème de l'enseignement supérieur est très aigu. [2] Il s'agit d'une situation très compliquée, [3] mais la raison la plus importante est celle de la position économique à présent. [4] Au moment où il y a un très haut niveau de chômage et où l'importation et l'exportation des biens sont très instables, les gouvernements trouvent difficile de justifier les dépenses pour l'enseignement supérieur. [5] Les ministres disent que l'on n'a pas assez d'argent à dépenser sur ce qui ne produit rien en particulier, c'est-à-dire des biens concrets que l'on peut vendre ou acheter.

[6] Il y a certains qui sont tout à fait contre cette raison que donne le gouvernement, surtout celui de la Grande-Bretagne. [7] On dit que le gouvernement paie bien d'autres choses qui ne sont pas vendables; par exemple l'argent que l'on dépense sur les armements nucléaires. [8] On dirait que ce n'est pas aussi important que l'enseignement, et qu'il faut réduire la dépense pour les armements nucléaires pour augmenter celle pour l'enseignement. [9] On dit que les recherches nucléaires sont inutiles, [10] car cela ne ménera qu'à la destruction du monde, tandis que l'enseignement supérieur a une importance beaucoup plus valable pour la vie actuelle. [11] On voit, dans les informations données, que le taux moyen d'accroissement annuel des dépenses publiques relatives à l'enseignement supérieur de la

Grande-Bretagne est le plus bas pendant les années 1965-1970. [12] On dirait que c'est à cause de la dépense sur d'autres choses, par exemple les armements nucléaires.

[13] Cependant, il y a d'autres qui disent qu'une expansion continue de l'enseignement supérieur est tout à fait impossible, parce que ça coûte trop cher. [14] On dirait que le taux élevé du chômage et tout ce que ça implique veut dire qu'il faut diminuer la dépense de l'argent sur l'enseignement supérieur, pour augmenter celle sur le chômage; par exemple pour créer des emplois pour les jeunes etc.. [15] On dirait que ce n'est pas juste de dépenser beaucoup d'argent sur ce qui ne produit rien pour bénéficier à tout le pays. [16] Il paraît des données que la situation en Grande-Bretagne est très aigue, ou bien, plus que celle en d'autres pays, par exemple la Suède, ou la France. [17] Il paraît que ces pays-là ont augmenté considérablement leur dépense de l'argent sur l'enseignement supérieur [18] et on ne peut que croire que la situation économique dans ces pays n'est pas aussi aigue que celle en Grande-Bretagne. [19] Encore une raison pour ce phénomène est celle des systèmes différents. [20] En Grande-Bretagne, on est très fier de son système d'enseignement supérieur que certains croient être le meilleur du monde. [21] Il est très difficile de gagner une place, [22] et une fois là, il faut travailler très dur pour obtenir des licences. [23] En autres pays, le système est plus laxe. [24] Tout le monde peut entrer dans une université, [25] et les licences ne sont pas difficiles à obtenir.

[26] C'est pour ces raisons que je crois qu'il faut maintenir les dépenses sur l'enseignement supérieur au niveau présent.

FLE-8

[1] A mon avis le gouvernement de Grande-Bretagne doit continuer à augmenter des crédits et des effectifs par rapport à l'enseignement supérieur, car une réduction ne serait pas un moyen effectif de faire des économies.

[2] En Grande-Bretagne moins de la population âgée de 19 à 24 ans font l'enseignement supérieur qu'en France (9,3% en grande-Bretagne, 12,3% en France). [3] Mais le pourcentage du P.N.B. consacré aux dépenses publiques dans l'enseignement supérieur est plus grand en Grande-Bretagne (0,89%) qu'en France (0,43%). [4] Au Japon 14,7% de la population âgée de 19 à 24 ans ont été admis au système d'enseignement supérieur, [5] mais le pourcentage du P.N.B. n'est que 0,38%.

[6] La population de la Grande-Bretagne passe moins d'années dans le système d'enseignement que celle de la plupart des autres pays. [7] (L'âge à l'entrée au système d'enseignement est 4,5 en Grande-Bretagne en comparaison avec 3,4 en France, tandis que l'âge à la sortie de celle-ci est 17,7 et 18,9 de celle-là). [8] Mais tous les autres pays, sauf les Etats-Unis, ont consacré moins de leur P.N.B. à l'enseignement supérieur public.

[9] Les autres pays ont augmenté leur dépenses publiques relatives à l'enseignement supérieur plus que la Grande-Bretagne pendant les années 1965-70. [10] (Le taux moyen d'acroissement annuel des dépenses relatives à l'enseignement supérieur est 24,71 en France, 18,07 au Japon, 28.09 en Suède, mais seulement 8.12 en Grande-Bretagne). [11] Mais notre pourcentage du P.N.B. consacré aux dépenses dans l'enseignement supérieur est quand même plus grand que celui de presque tous nos voisins. [12] (Il est de 0,89% en Grande-Bretagne mais seulement 0,43% en France, et 0,38% au Japon). [13] Bien que la Suède ait beaucoup augmenté ses dépenses publiques relatives à l'enseignement supérieur, le pourcentage de son P.N.B. a diminué de 0,78% en 1970 à 0,63% en 1974.

[14] Selon ces statistiques, il ne vaut pas la peine réduire des crédits et des effectifs relatifs à l'enseignement supérieur en Grande-Bretagne. [15] Par contre, il nous faut augmenter nos dépenses afin d'améliorer le système d'enseignement.

FLE-14

[1] Il faut admettre que les pourcentages britanniques dans les deux premières données semblent tout à fait catastrophiques, [2] mais bien qu'on ne puisse pas s'enorgueillir de ces chiffres, je crois qu'on peut présenter une explication valable. [3] Il faut se rappeler que les collèges et universités américains (qui paraissent avoir les chiffres les plus 'louables') ne sont pas aussi sélectifs que leurs équivalents anglais, [4] c'est à dire que si, en Grande-Bretagne, on créait des centaines de nouveaux collèges, on perdrait la qualité de notre enseignement supérieur.

[5] En la troisième donnée, on voit des chiffres qui indiquent pourquoi on voit les taux de scolarisation si bas en Angleterre en 1976 : [6] pendant les années 1965-70 on n'a pas eu un taux d'accroissement tellement grand. [7] Ainsi, on ne pouvait pas remédier à cette situation immédiatement, [8] et les effets de ce manque d'accroissement se manifestent toujours en 1976.

[9] Pourtant, il est en la quatrième donnée qu'on trouve les statistiques les plus inquiétantes: [10] en Angleterre, nous dépensons plus d'argent que les autres pays (sauf les Etats-Unis), [11] mais très peu de jeunes gens reçoivent l'enseignement supérieur. [12] On a l'argent... [13] mais on ne l'utilise pas sagement; [14] on peut dire que les Britanniques opérent un système qui n'est pas efficace. [15] Après tout, les autres pays arrivent à enseigner plus d'étudiants avec moins d'argent.

FLM-2

[1] Le taux de scolarisation pour la classe d'âge de 19 à 24 ans est en France l'un des plus bas. [2] Et l'on se rend compte grâce à la 2e statistique sur l'âge de la sortie de l'enseignement, que les Français sont les premiers, ou plutôt parmi les premiers à quitter l'enseignement. [3] C'est à dire qu'ils vont jusqu'au baccalauréat.

[4] Mais le baccalauréat n'est plus en 1982 un diplôme qui a une grande valeur. [5] En tous cas il ne débouche sur rien.

[6] Les Etats-Unis ont la plus grande proportion de jeunes dans l'enseignement supérieur (2 ,7%) [7] et ils quittent celui-ci à l'âge moyen de 21 ans et demi. [8] Mais cela est normal [9] car les Etats-Unis ont une population jeune si on la compare avec celle des pays d'Europe. [10] Et puis le pourcentage du PNB qu'ils destinent à l'enseignement supérieur est le plus important. [11] Cependant, pour ce qui est de la France, on remarque que pour la période 65 à 70 elle fait partie des pays qui ont fait le plus gros effort en faveur de l'enseignement supérieur.

[12] Mais depuis les années 70 la France a ralenti l'accroissement de ses dépenses pour ce secteur d'une facon remarquable. [13] D'autres pays comme la Grande-Bretagne, qui ont aussi été touchés assez durement par la crise ont pourtant un pourcentage du PNB accordé à l'enseignement supérieur plus important.

[14] La France pourrait donc faire un effort supplémentaire pour celui-ci, si elle ne veut pas avoir sur les bras des tas de jeunes sans aucune qualification.

[15] Les jeunes ne devraient pas faire les frais de ces réductions d'horaires, de la limitation dans le choix des matières qui sont la comséquence du manque d'argent à l'université. [16] C'est leur enlever autant d''armes' dont ils auraient pu se servir plus tard dans la vie active.

[17] La réduction des effectifs aurait pour conséquence, à mon avis de produire des gens aigris, (car n'ayant pas pu mener à bien leurs études, mal préparés à rentrer dans la vie active). [18] Cela n'aiderait surement pas le pays à remonter la pente.

[19] La réduction des crédits aurait pour conséquences de voir certaines petites universités reculer ou même être obligées de supprimer de grands pans de l'enseignement qui y était dispensé jusqu'ici.

FLM-5

[1] Ces différents tableaux nous informent sur la scolarisation de plusieurs pays développés, et les dépenses publiques relatives à l'enseignement supérieur.

[2] D'après le premier tableau, nous pouvons constater qu'en général, peu de jeunes de 19 à 24 ans suivent des études supérieures. [3] A part les USA qui se démarquent avec un pourcentage presque égal à 30, le taux d'étudiants reste faible pour les autres pays. [4] En effet, les cycles courts sont accessibles à tous alors que les cycles longs présentent des problèmes de tout ordre (concours d'entrée, prix des études) et ne sont donc possibles qu'à un certain nombre de personnes. [5] L'accès à l'enseignement supérieur n'est pas simple bien qu'il le semble à première vue; [6] bien sûr, toute personne bachelière peut s'inscrire et suivre ses études [7] mais à quoi cela sert-il si nous n'avons aucune garantie d'emploi à la sortie? Problème d'insécurité.

[8] Le second tableau nous renseigne sur le nombre moyen d'années passées à étudier. [9] C'est encore aux USA que l'on étudie le plus longtemps [10] et c'est en RFA que l'on étudie le moins longtemps. [11] Ce tableau signifierait-il que les étudiants de certains pays soient moins instruits que d'autres? [12] Il me semblerait plutôt juste de dire que les pays comme la RFA ont choisi une méthode efficace pour enseigner, c'est à dire une méthode rentable et certainement moins onéreuse. [13] Pour être plus objectif, il faudrait compter les années d'études à partir de l'entrée en primaire et non en maternelle. [14] Les petits Français y passent 2 ans, [15] mais la maternelle donne simplement l'épanouissement général.

[16] L'accroissement des dépenses publiques pour l'enseignement supérieur est fort élevé en France, [17] mais encore faut-il connaître les dépenses réelles. [18] En nous référant au 4ème tableau, on constate qu'elles représentent une faible part du PNB (à peine 5%), comparées aux USA ou à la GB. [19] De plus, il faudrait ajouter que ces dépenses sont faites dans des secteurs scientifiques (importance de l'informatique) et que les branches littéraires sont assez délaissées. [20] Ce qui est très surprenant, c'est que la France avec une scolarité supérieure à celle de l'Allemagne, consacre une part moindre de son PNB aux dépenses publiques.

FLM-7

[1] L'enseignement supérieur et notamment sa nécessaire adaptation au monde contemporain constitue actuellement un problème préoccupant dans la plupart des pays dits industrialisés. [2] La nouvelle conjoncture est en effet susceptible de remettre en cause une institution, qui dans la plupart des pays considérés, est vieille de plusieurs siècles et qui, n'ayant guère évolué dans ses structures, s'adapte de plus en plus difficilement aux exigences et aux besoins de nos sociétés actuelles. [3] Il n'est pas rare d'entendre parler de la crise que connait l'université pour ne citer qu'elle, considérée par beaucoup comme "une vieille dame malade" qu'il convient donc de soigner. [4] Toutefois, il est nécessaire avant d'étudier quels sont les remèdes envisageables, d'analyser d'un peu plus près à travers les statistiques fournies par l'OCDE, la situation actuelle de l'enseignement supérieur.

[5] Il convient tout d'abord de noter que ces statistiques ne concernent que des pays industrialisés ce qui permet un recoupement plus facile, mais restreint les éléments de comparaison entre différents systèmes économiques. [6] Ceci dit si des disparités quelquefois importantes peuvent apparaître entre les différents pays, il n'en reste pas moins qu'une même tendance globale existe. [7] En ce qui concerne le premier élément de considération, c'est à dire le taux de scolarisation de la population âgée de 19 à 24 ans disons que ce taux avoisine environ les 15%, les Etats Unis formant l'exception avec 26,7% des personnes de 19 à 24 ans scolarisées, ce qui peut être rapproché des résultats du second

tableau où les Etats Unis présentent l'âge moyen de sortie du système d'enseignement le plus élevé (21,5%). [8] La durée de la scolarité serait ainsi la plus importante aux Etats Unis. [9] La Grande-Bretagne est également, mais dans un sens nettement négatif, un cas particulier: [10] c'est elle en effet qui possède le taux de scolarisation le plus bas et de beaucoup par rapport aux autres pays de même qu'elle connait l'âge moyen de sortie du système le plus faible. [11] Notons sans qu'il n'y ait assurément de relation de cause à effet que c'est également le pays dont le taux d'accroissement des dépenses publiques pour l'enseignement supérieur est le plus bas. [12] En dehors de ces deux cas précis les autres pays suivent semble-t-il une tendance équivalente. [13] On peut cependant être surpris par les taux extrèmement peu élevés de la part du PNB consacré à l'enseignement supérieur public par chacun des pays. [14] Mais ces statistiques ne font pas apparaître ce qui est selon moi l'aspect le plus révélateur de l'enseignement supérieur de ces dernières années à savoir sa démocratisation. [15] De plus en plus de gens ont aujourd'hui accès à cet enseignement de même que la durée moyenne des études est de plus en plus longue. [16] L'enseignement supérieur est donc une section numériquement en expansion. [17] Le vrai problème est alors de savoir comment faire face à ce problème d'effectifs qui s'accompagne bien sûr pour le secteur public d'un problème de financement budgétaire.

[18] Il paraît assez difficile de réduire le nombre des effectifs ou de fermer les portes de l'enseignement supérieur, [19] ce serait foncièrement injuste. [20] Toute personne doit pouvoir avoir libre accès aux études supérieures qui doivent être un droit et non une necessité. [21] Chaque pays se doit donc d'assurer à ses citoyens cette possibilité. [22] D'autant plus que selon moi la démocratisation de l'enseignement supérieur est un facteur de progrès pour un pays; de progrès scientifique, technologique peut-être mais surtout de progrès vers une meilleure qualité de vie, une meilleure connaissance, un progrès "culturel" en quelque sorte. [23] Mais naturellement l'administration centrale d'un pays ne rentre pas dans des considérations d'ordre philosophique [24] elle réfléchit en termes de conjoncture économique: [25] il est en effet inutile de préparer des gens à des travaux ou à des emplois où ils ne trouveraient aucun débouchés.

[26] La situation de crise fausse donc le problème, [27] l'enseignement supérieur ne doit pas former des chômeurs. [28] A chaque pays de faire alors qu'il n'en soit pas ainsi, en commençant peut-être par accorder une plus grande place à l'enseignement dans ses dépenses publiques. [29] Il faut pouvoir arràter de réfléchir en terme de rendement ou de possibilités économiques et commencer à réfléchir en pensant au long terme. [30] L'investissement dans la matière grise doit être selon moi un bon investissement qu'aucun pays ne devrait négliger.

FLM-11

[1] Les nombreux problèmes que pose actuellement l'enseignement supérieur conduisent à s'interroger sur sa valeur même et la nouvelle orientation qu'il devrait prendre. [2] D'après notre expérience du système actuel, doit-on opérer une réduction des crédits et des effectifs ou bien continuer à développer?

[3] En premier lieu, une incohérence frappe l'enseignement supérieur: [4] de plus en plus, les élèves sont poussés à reculer leur âge d'entrée dans la vie active; [5] une majorité d'écoliers tentent de passer le baccalauréat; [6] en conséquence, le nombre d'étudiants tend à s'accroître bien que les crédits se réduisent de plus en plus dans l'enseignement supérieur. [7] Le pourcentage du PNB consacré aux dépenses publiques dans l'enseignement supérieur n'a que très peu augmenté entre 1970 et 74; [8] à l'opposé, bien que ne connaissant pas les chiffres exacts, je suppose que le pourcentage des étudiants croît bien plus rapidement. [9] Ce phénomène ne se limite pas à la France [10] et dans un pays comme la Suède on voit même le taux baisser entre 1970 et 74. [11] D'après une demande en diplômes de plus en plus importante dans la vie active, les subventions accordées par l'état à l'enseignement supérieur devraient elles aussi augmenter à taux égal.

[12] Un autre problème se pose: [13] selon le type d'études, les crédits différent; [14] actuellement, l'université doit répondre aux exigences de l'économie et du secteur industriel; [15] le secteur littéraire est donc l'un des plus défavorisé. [16] Ce fait est intolérable, [17] on encourage le développement de la science, de l'informatique... [18] mais que fait-on de la culture? [19] Aucun domaine ne devrait recevoir de plus grande faveur de la part de l'état.

[20] Le problème du manque de crédits se pose dans la majorité des facultés: [21] on supprime des cours, [22] on ferme les bibliothèques à des heures où les étudiants pourraient y travailler, [23] les classes sont surchargées... [24] Ceci résulte d'une inadéquation entre les crédits accordés et le nombre croissant de la population étudiante (bien que le taux de scolarisation de la population entre 19 et 24 ans soit faible: 12,3% pour la France).

[25] On doit continuer à augmenter crédits et effectifs mais en les employant mieux; [26] l'enseignement supérieur devrait avoir une orientation plus proche de la vie active; [27] un nombre considérable de jeunes quittent l'université après une première année d'études (souvent parce qu'un diplôme n'est pas toujours synonyme de travail) et restent alors sans formation.

[28] L'argent investi par le gouvernement dans l'enseignement supérieur ne l'est pas toujours fait à bon escient. [29] L'université devrait changer son orientation et permettre d'envisager plus sérieusement un avenir dans la vie active.

FLM-13

Commentaire

[1] 1) Pour ce qui est du 1er tableau de statistiques, on remarque que les USA ont un pourcentage, de beaucoup, plus élevé que ceux des autres pays: 26,7%. [2] C'est donc que la fréquentation des Universités est beaucoup plus importante aux USA qu'en Grande Bretagne, par exemple, où le taux est le moins élevé: 9,3%. [3] La France vient en 3ème position avec un pourcentage de 12,3.

[4] La Faculté semble donc beaucoup plus accessible aux USA: [5] les jeunes étudiants américains de 19 à 24 ans semblent pouvoir et vouloir plus volontiers prolonger leurs études dans l'enseignement supérieur.

[6] Il est reconnu que l'entrée dans les facultés Britanniques est très difficile et la poursuite des études supérieures en GB assez onéreuse. [7] Un tel pourcentage aux USA peut alors peut être s'expliquer par un moindre coût des études supérieures ou par le fait que les familles des étudiants sont plus aisées financièrement en moyenne et donc en mesure de subvenir aux besoins de leurs enfants si ceux-ci désirent poursuivre leurs études. [8] Je pense aussi qu'aux Etats-Unis, le fait de suivre des études universitaires est beaucoup plus répandu qu'en GB ou même en France; [9] en France, par exemple, on aurait plutôt tendance après le baccalauréat, à vouloir entrer directement dans la vie active, ou alors, à poursuivre des études courtes d'un ou deux ans.

[10] Il est possible que l'opinion publique considère que les études universitaires ne préparent pas suffisamment à l'entrée dans la vie professionnelle, et de ce fait les délaisse au profit des formations, plus courtes, mais, leur semble t-il, plus à même de les préparer à leur emploi futur.

[11] 2) Dans le 2ème tableau de statistiques, il est frappant de constater qu'en France, l'âge moyen d'entrée dans le système scolaire est particulièrement bas: 3,4. [12] Les enfants français sont donc très tôt scolarisés; [13] si l'on compare avec l'Allemagne Fédérale, l'âge moyen d'entrée y est de 6,5, [14] c'est près du double: [15] c'est une différence considérable. [16] En règle générale, dans tous les autres pays figurant dans ce tableau, sont scolarisés plus tardivement qu'en France. [17] Il est vrai qu'en France, pour entrer à la Maternelle l'âge minimum est de 3 ans: [18] c'est plus tôt qu'ailleurs. [19] Et les parents en profitent donc en majorité et les mettent à l'école dès leur plus jeune âge.

[20] Pour ce qui de l'âge moyen de sortie du système d'enseignement, c'est aux USA qu'il est le plus élevé: 21,5. [21] Cela rejoint un petit peu ce que nous avons dit dans la 1ère partie, à savoir , que les jeunes Américains restent plus longtemps dans le système d'enseignement: [22] en effet nous avons vu qu'ils étaient nombreux à prolonger leurs études au delà de 19 ans: 26,7%. [23] C'est en GB que l'âge de sortie est le moins élevé: 17,7. [24] Ces jeunes entrent donc plus rapidement que dans les autres pays dans la vie active.

[25] 3) On constate, dans ce 3ème tableau, que c'est en Suède que le taux d'accroissemenrt des dépenses publiques, relatives à l'enseignement supérieur, est le plus élevé: 28,09%. [26] C'est donc que dans les années 65-70, il s'est produit une forte augmentation dans les sommes allouées à l'enseignement supérieur; [27] en France aussi: le taux est relativement important: 24,71%. [28] C'est en GB que ce taux est le plus bas avec 8,12%: [29] les sommes allouées à l'enseignement supérieur ont très peu augmenté: [30] cela représente le 1/3 du taux en Suède.

[31] 4) On peut constater que c'est aux USA, et ce en 65, en 70 comme en 74, que le pourcentage du PNB consacré aux dépnses publiques dans l'enseignement supérieur est le plus important. [32] En 5 ans: de 65 à 70 le pourcentage a beaucoup augmenté alors qu'entre 70 et 74, c'est à dire en 4 ans, il a très peu augmenté: [33] il a donc tendance à se stabiliser.

[34] En 65 et 70, c'est en France que le pourcentage est le plus bas, avec une très faible progression, mais régulière et qui se poursuit en 74; [35] mais en 74, c'est l'Italie qui détient le plus faible pourcentage, suivie du Japon: [36] on peut noter que c'est le Japon qui connait la plus faible progression des pourcentages au fil des ans.

2. Exemple du corpus Reformulation :

Texte original	Reformulation
Leyford, town close to London, was a declining industrial town about twenty five years ago, but now our town is a beautiful thriving tourist and commercial centre. What made Leyford change so completely ?	*A quarter of a century ago, Leyford, a small town situated close to London, was a declining industrial town. Now it is a beautiful and thriving tourist and commercial centre. What has caused Leyford to change so radically ?*
The trigger of the change was the world recession. After the Second World War, the industrial foreign competition became severe. Japan and West Germany began to produce many good cheap machines such as cars and typewriters. As a result British industry declined and many factories closed. In this town there were five factories, but now there are none. The closure of factories caused the flood of unemployed people for a job in London. There were so many people in London that the government felt it necessary to decentralize overpopulated London.	*The change was triggered, ironically, by the world recession. After the Second World War, foreign industrial competition became severe as Japan and West Germany, in particular, began to produce good cheap manufactured goods (cars, typewriters, etc.). As a result, British industry declined and many factories closed all over the country. In Leyford itself there were once five factories, but all five were eventually forced to close. These closures, in Leyford and elsewhere, caused a great flood of unemployed people to leave their home and look for work in London. Eventually London became so overpopulated that the government felt it necessary to adopt a policy of decentralisation. It is this decentralisation policy that has enabled Leyford to recover so successfully from the period of industrial decline.*

3.Extraits du corpus MIEL (PSYCHO, INFO1, INFO2, MANA)

Definitions (or extracts) used as examples in the text, sorted by attack and structure type.

1. Attaque 1: "histoire"

L'animal est placé dans sa cage (PSYCHO).

Lorsque deux mots ont un rapport sémantique assez proche, l'énoncé d'un des deux mots va activer dans la mémoire le second mot (PSYCHO).

Après extinction et une période de repos, on présente de nouveau le stimulus conditionnel à l'animal, on constate de nouveau la réaction conditionnelle (PSYCHO).

2. Attaque 2: "action"

Tri: ordonner un ensemble d'éléments donné selon un ordre précis (INFO-2).

Itération: répéter un ensemble d'instructions jusqu'à ce qu'un ensemble de conditions soit éventuellement rempli.(INFO-1)

Trier: faire apparaître une relation d'ordre dans un ensemble d'éléments (INFO-2).

Le tableau de financement regroupe les emplois d'un côté (...) et les ressources (...) dans un tableau (MANA).

Un tri permet de classer des éléments selon un certain ordre (INFO-2).

Le tableau de financement sert à calculer la variation du fonds de roulement (MANA).

3. Attaque 3: "générique (processus)"

Un tri est un mécanisme qui permet d'ordonner un ensemble d'éléments de même type suivant un critère donné (INFO-2).

Phénomène par lequel la perception d'un objet est inchangée malgré les changements de stimulation physique qui la provoque (PSYCHO).

4. Attaque 4: "synonyme (concept)"

Le conditionnement est l'étude des modifications d'un comportement après présentation de différents stimuli (PSYCHO).

Apprentissage d'un type de comportement en réponse à un stimulus donné (PSYCHO).

C'est l'acquisition d'un comportement nouveau en réponse à un stimulus neutre (ne provoquant pas de réponse initialement) (PSYCHO).

Un tri est l'ordonnancement d'éléments dans un ordre voulu (INFO-2).

Nouvel apprentissage beaucoup plus rapide que le premier. (PSYCHO)

Le tableau de financement est un récapitulatif des emplois stables et des ressources durables que possède l'entreprise (MANA).

Réapparition d'une réponse conditionnelle, ayant subi une extinction, sans renforcement(PSYCHO).

5. Structure: Identification

Module objet: c'est le résultat de la compilation d'un programme source (INFO-1).

Après avoir observé une extinction et après une période de repos si on recommence une série de tests, on observe une RC à la présentation de SC: c*'est la récupération spontanée de l'apprentissage. (PSYCHO)*

Module objet : Un programme, tel qu'il est écrit dans un langage évolué, n'est pas compréhensible par l'ordinateur. Il doit pour cela être traduit en une séquence d'instructions plus élémentaires que celui-ci pourra comprendre. Un module objet est une telle séquence d'instructions. (INFO-1)

Module objet: c'est le résultat de la compilation d'un programme source, le module objet peut être lié pour obtenir du code exécutable. (INFO-1).

C'est l'acquisition d'un comportement nouveau en réponse à un stimulus neutre (ne provoquant pas de réponse initialement). Il s'obtient à l'aide d'expériences répétées et par des renforcements du stimulus neutre (appelé stimulus conditionnel). (PSYCHO).

6. Structure: explicitation-illustration

Un tri permet de classer des éléments suivant un certain ordre (croissant ou décroissant). Il peut se faire sur des éléments numériques, alpha-numériques ou alphabétiques. On peut trier des éléments suivant plusieurs méthodes: quicksort, dichotomie, insertion, heapsort, sélection, méthode bulle. (INFO2)

Un tri permet d'ordonner des données suivant une relation d'ordre préalablement établie. D'un point de vue algorithmique nous avons plusieurs principes de tri qui sont plus ou moins performants. Pour exemple le tri "bulle" qui est très lent, ou le heapsort qui lui est très performant, c'est à dire très rapide. (INFO2)

7. Structure: situation-explication

Le tableau de financement est un récapitulatif des emplois stables et des ressources durables que possède l'entreprise. Il permet d'obtenir le fond de roulement net et global et en cela il est un bon indicateur de la sécurité de la situation financière de l'entreprise. Si FRN est supérieur à la partie structurelle du BFRE alors il est suffisant. (MANA)

Après avoir observé une extinction et après une période de repos si on recommence une série de tests, on observe une RC à la présentation de SC: c'est la récupération spontanée de l'apprentissage. (PSYCHO)

C'est un désapprentissage. Si l'on ne présente plus que le stimulus neutre il n'y aura plus de réponse. Pour Pavlov après le conditionnement on ne présente plus que le son le chien ne salivera plus. (PSYCHO)

Après une extinction et un temps de repos, si on représente le stimulus conditionnel, on observe le comportement initialement induit par le conditionnement : c'est le phénomène de la récupération spontanée. (PSYCHO)

Après extinction et une période de repos, on présente de nouveau le stimulus conditionnel à l'animal, on constate de nouveau la réaction conditionnelle. (PSYCHO)

4. Corpus LOG1 : vue d'ensemble

Le montage ci-après illustre la présentation du corpus LOG1 (manuel du logiciel SATO), en fournissant des extraits de cinq des huit sous-sections qui composent la partie 6.1 (Analyseur) du chapitre décrivant les commandes du programme d'interrogation (chap. 6).

Bibliographie

Aarts, J. & Meijs, W., Eds. (1984). *Corpus linguistics I. Recent developments in the use of computer corpora in English language research.* Amsterdam: Rodopi.

Adam, J.-M. (1991). Cadre théorique d'une typologie séquentielle. *Etudes de linguistique appliquée*, (83), 7-18.

Adam, J.-M. & Revaz, F. (1989). Aspects de la Structuration du Texte Descriptif: Les Marqueurs d'Enumération et de Reformulation. *Langue Française*, (81), 59-98.

Aijmer, K. (1986). Why is *actually* so popular in spoken English? In G. Tottie & I. Bäcklund, Eds., *English in speech and writing: a symposium*, pp. 119-130. Stockholm: Almqvist & Wiksell.

Allwright, R.L., Woodley, M.-P. & Allwright, J.M. (1988). Investigating reformulation as a practical strategy for the teaching of academic writing. *Applied Linguistics*, **9**(3), 236-256.

Atlani, F. (1984). ON l'illusioniste. In A. Grésillon & J-L. Lebrave, Eds., *La langue au ras du texte*, pp. 13-30. Lille: Presses Universitaires de Lille.

Bateman, J. & Rondhuis, K.J. (1997). "Coherence Relations". Towards a General Specification. *Discourse Processes*, **24**(1), 3-50.

Beacco, J.-C., Ed. (1992). *Ethnolinguistique de l'écrit.* Langages 105. Paris: Larousse.

de Beaugrande, R. & Dressler, W. (1981). *Introduction to Text Linguistics*. London & New York: Longman.

Beekman, J. & Callow, J. (1974). *Translating the Word of God.* Grand Rapids, MI: Zondervan Publishing House.

Berri, J., Cartier, E., Desclés, J.-P., Jackiewicz, A. & Minel, J.-L. (1996). SAFIR, système automatique de filtrage de textes. In *Actes, TALN'96*.

Berthoud, A-C. & Mondada, L. (1991). Stratégies et marques d'introduction et de réintroduction d'un objet dans la conversation. *Bulletin CILA*, (54), 159-179.

Berthoud, A.-C. (1996). *Paroles à propos. Approche énonciative et interactive du topic.* Paris: Ophrys.

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

Biber, D. (1989). A typology of English texts. *Linguistics*, (27), 3-43.

Biber, D. (1990). Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing*, **5**, 257-269.

Biber, D. (1992). Using computer-based text corpora to analyze the referential stragegies of spoken and written texts. In J. Svartvik, Ed. *Directions in corpus linguistics*, pp. 213-252. Berlin, New York: Mouton de Gruyter.

Biber, D. (1993a). Representativeness in corpus design. *Literary and Linguistic Computing*, **8**, 243-257.

Biber, D. (1993b). Using register-diversified corpora for general language studies. *Computational Linguistics*, **19**(2), 219-241.

Biber, D. (1993c). Co-occurrence patterns among collocations: a tool for corpus-based lexical knowledge acquisition. *Computational Linguistics*, **19**(3), 531-538.

Biber, D. (1995). *Dimensions of register variation : A cross-linguistic comparison.* Cambridge: Cambridge University Press.

Biber, D. & Finegan, E. (1993). Intra-textual variation within medical research articles. In N. Oostdijk & P. de Haan, Eds., *Corpus-based research into language. In honour of Jan Aarts*, pp. 201-221. Amsterdam/Atlanta: Rodopi.

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 of Jan Aarts*, pp. 201-221. Amsterdam/Atlanta: Rodopi.

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

Blanche-Benveniste, C. (1996). De l'utilité du corpus linguistique. *Revue Française de Linguistique Appliquée*, **I**(2), 25-42.

Borkin, A. (1984). *Problems in Form and Function.* London: Ablex.

Brennan, S.E., Friedman, M.W. & Pollard, C.J. (1987). A Centering Approach to Pronouns. In *Actes, 25th Annual Meeting of the Association for Computational Linguistics*, pp. 155-162. Stanford, CA.

Brennan, S.E. (1998). Centering as a Psychological Resource for Achieving Joint Reference in Spontaneous Discourse. In M. Walker, A. Joshi & E. Prince, Eds., *Centering Theory in Discourse*, pp. 227-249. Oxford: Clarendon Press.

Brée, D.S. & Smit, R.A. (1986). Linking Propositions. In *Actes, COLING-86*, pp. 177-180.

Bronckart, J.-P., Bain, D., Schneuwly, B., Davaud, C. & Pasquier, A. (1985). *Le fonctionnement des discours.* Neuchâtel: Delachaux et Niestlé.

Candel, D. (1994). Le discours définitoire : variations discursives chez les scientifiques. In S. Moirand, M. Ali Bouacha, J.-C. Beacco & A. Collinot, Eds., *Parcours linguistiques de discours spécialisés*, Berne, Paris, New York: Peter Lang.

Cartier, E. (1997). La définition dans les textes scientifiques et techniques : présentation d'un outil d'extraction automatique des relations définitoires. In *Actes, TIA'97 (Terminologie et Intelligence Artificielle)*, Toulouse.

Chafe, W. (1970). *Meaning and the Structure of Language.* Chicago: University of Chicago Press.

Chafe, W.L. (1976). Givenness, Contrastiveness, Definiteness, Subjects, Topics and Point of View. In C. N. Li, Ed. *Subject and Topic*, New-York: Academic Press.

Charolles, M. (1978). Introduction aux problèmes de la cohérence des textes. *Langue Française*, **38**, 7-41.

Charolles, M. (1983). Coherence as a Principle in the Interpretation of Discourse. *Text*, **3**, 71-97.

Charolles, M. (1995). Cohésion, cohérence et pertinence du discours. *Travaux de Linguistique*, **29**(112), 125-151.

Charolles, M. (1997). *L'encadrement du discours : Univers, champs, domaines et espaces*. Cahier de Recherche Linguistique 6, Université de Nancy 2.

Charolles, M., Petöfi, J.S. & Sözer, E. (1986). *Research in Text Connexity and Text Coherence. A Survey.* Hamburg: Buske.

Church, K.W. & Mercer, R.L. (1993). Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, **19**(1), 1-24.

Clark, H.H. & Haviland, S.E. (1977). Comprehension and the given-new contract. In R. O. Freedle, Ed. *Discourse Production and Comprehension*, Norwood, New Jersey: Ablex.

Clyne, M. (1981). Culture and discourse structure. *Journal of Pragmatics*, **5**, 61-66.

Cohen, R. (1984). A Computational Theory of the Function of Clue Words in Argument Understanding. In *Actes, COLING-84*, pp. 251-258.

Combettes, B. (1983). *Pour une grammaire textuelle : la progression thématique.* Bruxelles: Duculot/De Back.

Combettes, B. (1998). *Les constructions détachées en français.* Paris: Ophrys.

Condamines, A., Fabre, C. & Péry-Woodley, M.-P., Eds. (1999). *Corpus et TAL : pour une réflexion méthodologique.* TALN'99. Cargese: ATALA.

Condamines, A. & Rebeyrolle, J. (à paraître). Searching for and identifying conceptual relationships via a corpus-based approach to Terminological Knowledge Bases (CTKB): methods and results. In D. Bourigault, M.-C. L'Homme & C. Jacquemin, Eds., *Recent Advances in Computational Terminology* , Amsterdam: John Benjamins.

Corbin, P. (1980). De la production des données en linguistique introspective. In A. M. Dessaux-Berthoneau, Ed. *Théories linguistiques et traditions grammaticales*, pp. 121-179. Lyon: Presses Universitaires de Lyon.

Cornish, F. (1986). *Anaphoric Relations in English and French. A Discourse Perspective.* London, Sydney, Dover New Hampshire: Croom Helm.

Cornish, F. (1998). The Functional Grammar conception of (discourse) anaphora: a sympathetic critique. In *Actes, 8t International Conference on Functional Grammar*, Amsterdam.

Cornish, F. (1999). *Anaphora, Discourse and Understanding. Evidence from English and French.* Oxford: Clarendon Press.

Cornish, F. (à paraître). L'accessibilité cognitive des référents, le centrage d'attention, et la structuration du discours : une vue d'ensemble. *Verbum*, **22**(1),

Dachelet, R. (1994). *Sur la notion de sous-langage*. Doctorat en sciences du langage, Université de Paris VIII.

Danes, F. (1966). A three-level approach to syntax. In F. Danes *et al.*, Ed. *Travaux linguistiques de Prague*, pp. 225-240. University of Alabama Press.

Danes, F. (1974). Functional Sentence Perspective and the Organization of Text: Different types of Thematic Progression. In F. Danes, Ed. *Papers on Functional Sentence Perspective*, The Hague: Mouton.

Daniel, M.P., Nicaud, L., Prince, V. & Péry-Woodley, M.-P. (1992). Apport du style linguistique à la modélisation cognitive d'un élève. In C. Frasson, G. Gauthier & G. I. McCalla, Eds., *Intelligent Tutoring Systems*, pp. 252-259. Berlin: Springer-Verlag.

Davison, A. (1984). Syntactic markedness and the definition of sentence topic. *Language*, **60**(4), 797-846.

van Dijk, T.A. (1977). *Text and Context. Explorations in the Semantics and Pragmatics of Discourse.* London, New York: Longman.

van Dijk, T.A. (1981). *Studies in the Pragmatics of Discourse.* The Hague, Paris, New-York: Mouton.

Dik, S.C. (1997). *The Theory of Functional Grammar.* Berlin: de Gruyter.

Ducrot, O. (1983a). *Puisque* : essai de description polyphonique. *Revue Romane*, **24**, 166-185.

Ducrot, O. (1983b). Opérateurs argumentatifs et visée argumentative. *Cahiers de Linguistique Française*, **5**, 7-36.

Ducrot, O. *et al.* (1980). *Les mots du discours.* Paris: Editions de Minuit.

Enkvist, N.E. (1985). A Parametric View of Word Order. In E. Sözer, Ed. *Text Connexity, Text Coherence*, pp. 320-336. Hamburg: Helmut Buske.

Fillmore, C.J. (1992). "Corpus linguistics" or "Computer-aided armchair linguistics". In J. Svartvik, Ed. *Directions in corpus linguistics: Proceedings of Nobel Symposium 82* , pp. 35-60. Berlin, New York: Mouton de Gruyter.

Firbas, J. (1964). On defining the theme in functional sentence analysis. *Travaux Linguistiques de Prague*, **1**, 267-280.

Firbas, J. (1972). On the Interplay of Prosodic and Non-Prosodic Means of Functional Sentence Perspective. In U. Fried, Ed. *The Prague School of Linguistics and Language Teaching*, pp. 77-94. London: Oxford University Press.

Firbas, J. (1986). Thoughts on Functional Sentence Perspective, intonation and emotiveness. *Brno Studies in English*, **16**, 11-48.

Flowerdew, J.L. (1992). Salience in the performance of one speech act: the case of definitions. *Discourse Processes*, **15**, 165-181.

Fradin, B. (1990). Approche des constructions à détachement. Inventaire. *Revue Romane*, **52**(1), 3-34.

Francis, W.N. & Kucera, H. (1982). *Frequency analysis of English usage: lexicon and grammar.* Boston: Houghton Mifflin.

Givón, T. (1979). *On Understanding Grammar.* New-York, San Francisco, London: Academic Press.

Givón, T. (1983). Topic continuity in discourse: An introduction. In T. Givón, Ed. *Topic continuity in discourse*, Amsterdam/Philadelphia: John Benjamins.

Goody, J. (1977). *The Domestication of the Savage Mind.* Cambridge: Cambridge University Press.

Gordon, P.C., Grosz, B.J. & Gilliom, L.A. (1993). Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, **17**, 311-347.

Grice, H.P. (1975). Logic and conversation. In P. Cole & J. Morgan, Eds., *Syntax and Semantics 3: Speech Acts*, New York: Academic Press.

Grimes, J.E. (1975). *The Thread of Discourse.* The Hague: Mouton.

Grishman, R. & Kittredge, R., Eds. (1986). *Analyzing language in restricted domains. Sublanguage description and processing.* Hillsdale, N.J.: Laurence Erlbaum.

Grosz, B.J. (1977). The representation and use of focus in a system for understanding dialogs. In *Actes, 5th International Joint Conference on Artificial Intelligence*, Cambridge, MA.

Grosz, B.J. & Sidner, C.L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, **12**(3), 175-204.

Grosz, B.J., Joshi, A. & Weinstein, S. (1995). Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, **21**(2), 203-225.

Grosz, B.J. & Ziv, Y. (1998). Centering, global focus, and right-dislocation. In M. Walker, A. Joshi & E. Prince, Eds., *Centering Theory in Discourse*, pp. 293-308. Oxford: Clarendon Press.

Gundel, J.K. (1985). Shared knowledge' and topicality. *Journal of Pragmatics*, **9**, 83-107.

Gundel, J.K. (à paraître). Statut cognitif et formes des anaphoriques indirects. *Verbum*, **22**(1).

Gundel, J.K. (1988). Universals of topic-comment structure. In M. Hammond, E. Moravcsik & J. Wirth, Eds., *Studies in Linguistic Typology*, pp. 209-239. Amsterdam/Philadelphia: John Benjamins.

Gundel, J.K., Hedberg, N. & Zacharski, R. (1993). Cognitive Status and the Form of Referring Expressions. *Language*, **69**, 274-307.

Gundel, J.K. (1998). Centering Theory and the Givenness Hierarchy: Towards a Synthesis. In M. Walker, A. Joshi & E. Prince, Eds., *Centering Theory in Discourse*, pp. 183-198. Oxford: Clarendon Press.

Habert, B. (1985). Etudes des formes "spécifiques" et typologie des énoncés (les résolutions générales des congrès de la CFTC-CFDT de 1945 à 1979). *Mots*, **11**, 127-154.

Habert, B. (1995). Introduction. *TAL*, **36**(1-2), 3-6. Traitements probabilistes et corpus, Benoît Habert, (resp.).

Habert, B. (1998). *Un corpus clé pour le français actuel.* http://www.biomath.jussieu.fr/CLEF/

Habert, B. & Jacquemin, C. (1993). Noms composés, termes, dénominations complexes : problématiques linguistiques et traitements automatiques. *TAL*, **34**(2), 5-42.

Habert, B., Barbaud, P., Dupuis, F. & Jacquemin, C. (1995). Simplifier des arbres d'analyse pour dégager les comportements syntactico-sémantiques des formes d'un corpus. *Cahiers de Grammaire*, (20), 1-32.

Habert, B. & Salem, A. (1995). L'utilisation de catégories multiples pour l'analyse quantitative de données textuelles. *TAL*, **36**(1-2), 249-276.

Habert, B., Nazarenko, A. & Salem, A. (1997). *Les linguistiques de corpus.* Paris: Armand Colin.

Habert, B., Fabre, C. & Issac, F. (1998). *De l'écrit au numérique. Constituer, normaliser et exploiter les corpus électroniques.* Paris: Masson.

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

Hahn, U. & Strube, M. (1997). Centering in-the-large: Computing referential discourse segments. In *Actes, 35th Annual Meeting of the ACL*, pp. 104-111..

Halliday, M.A.K. (1967a). Notes on Transitivity and Theme in English. Part 1. *Journal of Linguistics*, **3**(1), 37-81.

Halliday, M.A.K. (1967b). Notes on Transitivity and Theme in English. Part 2. *Journal of Linguistics*, **3**(2), 199-244.

Halliday, M.A.K. (1968). Notes on Transitivity and Theme in English. Part 3. *Journal of Linguistics*, **4**(2), 179-215.

Halliday, M.A.K. (1980). Text semantics and clause grammar: some patterns of realization. In J. E. Copeland & P. Davis, Eds., *The 7th LACUS Forum*, Columbia, S.C.: Hornbeam Press.

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

Halliday, M.A.K. (1991). Corpus studies and probabilistic grammar. In K. Aijmer & B. Altenberg, Eds., *English corpus linguistics. Studies in honour of Jan Svartvik*, pp. 30-43. London: Longman.

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

Harris, Z.S. (1968). *Mathematical Structures of Language.* New York: Wiley & Sons.

Harris, Z.S. (1982). *A grammar of English on mathematical principles.* New York: Wiley-Interscience.

Harris, Z.S. (1988). *Language and information.* New York: Columbia University Press.

Harris, Z.S. (1991). *A theory of language and information; A mathematical approach.* Oxford: Clarendon Press.

Harris, Z., Gottfried, M., Ryckman, T., Mattick Jr, P., Daladier, A., Harris, T.N., *et al.* (1989). *The form of information in science. Analysis of an immunology sublanguage.* Dordrecht: Kluwer Acdemic Publishers.

Hathout, N. (1996). Pour la construction d'une base de connaissances lexicographiques à partir du Trésor de la Langue Française : Les marqueurs superficiels dans les définitions spécialisées. *Cahiers de Lexicologie*, **68**(1), 137-173.

Hearst, M.A. (1998). Automated discovery of WordNet relations. In C. Fellbaum, Ed. *WordNet: an electronic lexical database* , pp. 131-151. Boston: MIT Press.

Hinds, J. (1983). Contrastive Rhetoric: Japanese and English. *Text*, **3**, 183-196.

Hitzeman, J. & Poesio, M. (1998). Long distance pronominalisation and global focus. In *Actes, COLING-ACL'98*, Montreal, ACL.

Hobbs, J.R. (1985). *On the Coherence and Structure of Discourse*. CSLI-85-37, Center for the Study of Language and Information.

Hoey, M. (1983). *On the Surface of Discourse.* London: Allen & Unwin.

Hovy, E. (1990). Parsimonious and profligate approaches to the question of discourse structure relations. In *Actes, 5th International Workshop on Language Generation*, Pittsburgh, PA.

Hunt, K.W. (1965). *Grammatical structures written at three grade levels*. NCTE Research Report 3, NCTE.

Hunt, K.W. (1970). *Syntactic maturity in schoolchildren and adults.* Chicago: University of Chicago Press.

Hurewitz, F. (1998). A quantitative look at discourse coherence. In M. Walker, A. Joshi & E. Prince, Eds., *Centering Theory in Discourse*, pp. 273-292. Oxford: Clarendon Press.

Illouz, G. (1999). Méta-étiqueteur adaptatif : vers une utilisation pragmatique des ressources linguistiques. In *Actes, TALN'99 (Traitement Automatique des Langues Naturelles)*, pp. 185-194. Cargese, ATALA.

Illouz, G., Habert, B., Fleury, S., Folch, H., Heiden, S. & Lafon, P. (1999). Maîtriser les déluges de données hétérogènes. In A. Condamines, C. Fabre & M.-P. Péry-Woodley, Eds., *Actes, Corpus et TAL : pour une réflexion méthodologique. TALN'99 (Traitement Automatique des Langues Naturelles)*, pp. 37-46. Cargese.

Jacquemin, C. & Bush, C. (2000). Combining lexical and formatting cues for named entity acquisition from the Web. In *Actes, COLING'2000* (soumis).

Kameyama, M. (1998). Intrasentential Centering: A Case Study. In M. Walker, A. Joshi & E. Prince, Eds., *Centering Theory in Discourse*, pp. 89-114. Oxford: Clarendon Press.

Kass, R. (1989). Student Modeling in Intelligent Tutoring Systems. In A. Kobsa & W. Wahlster, Eds., *User Models in Dialog Systems*, Berlin, Heidelberg, New York: Springer Verlag.

Keenan, E.L. (1985). Passive in the world's languages. In T. Shopen, Ed. *Language Typology and Syntactic Description*, Cambridge: Cambridge University Press.

Kieras, D.E. (1981). The role of major referents and sentence topics in the construction of passage macrostructure. *Discourse Processes*, (4), 1-15.

Kobsa, A. & Wahlster, W. (1988). Preface. *Computational Linguistics. Special Issue on User Modeling*, **14**(3), 1-4.

Lakoff, R. (1984). The Pragmatics of Subordination. In *Actes, 10th Annual Meeting of the Berkeley Linguistics Society*, pp. 481-492. Berkeley, CA, Berkeley Linguistics Society.

Lambrecht, K. (1981). *Topic, antitopic and verb-agreement in non-standard French.* Amsterdam/Philadelphia: John Benjamins.

Lambrecht, K. (1982). *Discourse Pragmatics.* University of California, Berkeley.

Lambrecht, K. (1987). On the status of SVO sentences in French discourse. In R. Tomlin, Ed. *Coherence and grounding in discourse*, pp. 217-262. Amsterdam/Philadelphia: John Benjamins.

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

Lambrecht, K. (1998). Sur la relation formelle et fonctionnelle entre topiques et vocatifs. *Langues*, **1**(1), 34-45.

Lambrecht, K. (à paraître). Dislocation. In M. Haspelmath, Ed. *Language Typology and Language Universals*, Berlin, New York: Walter de Gruyter.

Landelle, M. (1988). *Analyse syntaxique de l'expression de la segmentation : Approche linguistique pour un traitement informatique des structures textuelles*. DEA de linguistique, Université de Toulouse 2.

Lautamatti, L. (1978). Some Observations on Cohesion and Coherence in Simplified Texts. In J. O. Östman, Ed. *Cohesion and Semantics*, pp. 165-181. Åbo, Finland: Research Institute of the Åbo Akademi Foundation.

Lautamatti, L. (1987). Observations on the development of the topic of simplified discourse. In U. Connor & R. B. Kaplan, Eds., *Writing across Languages: Analysis of L2 Text*, pp. 87-114. Reading, MA: Addison-Wesley.

Leech, G. (1991). The state of the art in corpus linguistics. In K. Aijmer & B. Altenberg, Eds., *English corpus linguistics,* pp. 8-29. London: Longman.

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

Longacre, R.E. (1976). *An Anatomy of Speech Notions.* Lisse: The Peter de Ridder Press.

Longacre, R.E. (1979). The paragraph as a grammatical unit. In T. Givón, Ed. *Discourse and Syntax*, pp. 115-134. New York, London: Academic Press.

Luc, C. (1998a). Contraintes sur l'architecture textuelle. *Document Numérique*, **2**(2), 203-219.

Luc, C. (1998b). Types de contraintes architecturales sur la composition d'objets textuels. In *Actes, CIDE'98 (Colloque International sur le Document Électronique)*, pp. 15-30. Rabat, Maroc.

Luc, C., Mojahid, M. & Virbel, J. (1999). Connaissances structurelles et modèles nécessaires à la génération de textes formatés. In *Actes, GAT'99 (Génération Automatique de Textes)*, pp. 157-170. Grenoble.

Luc, C., Mojahid, M., Virbel, J., Garcia-Debanc, C. & Péry-Woodley, M.-P. (1999). A linguistic approach to some parameters of layout: A study of enumerations. In R. Power & D. Scott, Eds., *Actes, AAAI 1999 Fall Symposium: "Using Layout for the Generation, Understanding or Retrieval of Documents"*, pp. 20-29. North Falmouth, Massachussetts.

Luc, C., Mojahid, M., Péry-Woodley, M.-P. & Virbel, J. (2000). Les énumérations : structures visuelles, syntaxiques et rhétoriques. In *Actes, CIDE'2000 (Colloque International sur le Document Électronique)*, Lyon, France.

Maier, E. & Hovy, E. (1993). Organising discourse structure relations using metafunctions. In H. Horacek & M. Zock, Eds., *New Concepts in Natural Language Generation*, pp. 69-86. London: Pinter.

Mann, W.C. & Thompson, S.A. (1986). Relational propositions in discourse. *Discourse Processes*, **9**(1), 57-90.

Mann, W.C. & Thompson, S.A. (1987). Rhetorical structure theory: a theory of text organization. In L. Polanyi, Ed. *The Structure of Discourse*, Norwood, N.J.: Ablex.

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

Mann, W.C. & Thompson, S.A., Eds. (1992). *Discourse description. Diverse linguistic analyses of a fund-raising text.* Pragmatics and Beyond. Amsterdam, Philadelphia: John Benjamins.

Mann, W.C., Matthiessen, C. & Thompson, S.A. (1989). *Rhetorical structure theory and text analysis*. ISI/RR-89-242, Information Sciences Institute.

Martin, J.R. (1983). Conjunction: the Logic of English Text. In J. R. Martin, Ed. *Papers in Text Linguistics, Vol. 45*, pp. 1-72. Hamburg: Helmut Buske Verlag.

Martin, R. (1983). *La logique du sens.* Paris: PUF.

Martin, R. (1990). La définition "naturelle". In J. Chaurand & F. Mazière, Eds., *La définition*, pp. 86-96. Paris: Larousse.

Matthiessen, C. & Thompson, S.A. (1988). The structure of discourse and "subordination". In J. Haiman & S. A. Thompson, Eds., *Clause Combining in Grammar and Discourse*, Amsterdam: Benjamins.

McKeown, K. (1985). *Text Generation: Using discourse strategies and focus constraints to generate natural language text.* Cambridge UK: Cambridge University Press.

McNaught, J. (1993). User needs for textual corpora in natural language processing. *Literary & Linguistics Computing*, **8**(4), 227-234.

Montgomery, C.A. & Glover, B.C. (1986). A sublanguage for reporting and analysis of space events. In R. Grishman & R. Kittredge, Eds., *Analysing language in restricted domains. Sublanguage description and processing*, pp. 129-162. Hillsdale, N.J.: Lawrence Erlbaum.

Morin, E. (1999). Using lexico-syntactic patterns to extract semantic relations between terms form a technical corpus. In P. Sandrini, Ed. *Actes, 5th International Congress on Terminology and Knowledge Engineering (TKE'99)*, pp. 268-278. Innsbrück, Austria.

Nicaud, L. & Prince, V. (1990). TEDDI : An ITS for definitions learning. In *Actes, PRICAI'90*.

Nunberg, G. (1990). *The Linguistics of Punctuation.* Menlo Park: Center for the Study of Language and Information.

O'Brien, T. (1995). Rhetorical Structure Analysis and the case of the inaccurate incoherent source-hopper. *Applied Linguistics*, **16**(4), 442-482,

Ong, W.J. (1982). *Orality and Literacy: The Technologizing of the Word.* London: Methuen.

Pascual, E. (1991). *Représentation de l'architecture textuelle et génération de texte*. Doctorat d'informatique, Université Paul Sabatier, Toulouse.

Pascual, E. & Péry-Woodley, M.-P. (1995). La définition dans le texte. In J.-L. Nespoulous & J. Virbel, Eds., *Textes de type consigne – Perception, action, cognition*, pp. 65-88. Toulouse: PRESCOT.

Pascual, E. & Péry-Woodley, M.-P. (1997a). Définition et action dans les textes procéduraux. In E. Pascual, J.-L. Nespoulous & J. Virbel, Eds., *Le texte procédural : langage, action et cognition*, pp. 223-248. Toulouse: PRESCOT.

Pascual, E. & Péry-Woodley, M.-P. (1997b). Modèles de texte pour la définition. In *Actes, Premières Journées Scientifiques et Techniques du Réseau francophone de l'Ingénierie de la Langue de l'AUPELF-UREF*, pp. 137-146. Avignon, AUPELF-UREF.

Pascual, E. & Péry-Woodley, M.-P. (1997c). Modélisation des définitions dans les textes à consignes. In J. Virbel, J.-M. Cellier & J.-L. Nespoulous, Eds., *Cognition, Discours procédural, Action*, pp. 37-55. Toulouse: PRESCOT.

Passonneau, R.J. (1998). Interaction of Discourse Structure with Explicitness of Discourse Anaphoric Noun Phrases. In M. Walker, A. Joshi & E. Prince, Eds., *Centering Theory in Discourse*, pp. 327-358. Oxford: Clarendon Press.

Péry-Woodley, M.-P. (1989). *Textual designs: Signalling coherence in first and second language academic writing*. Doctorat (Ph.D.) de linguistique, Université de Lancaster. (édité en 1991 comme Notes et Documents LIMSI 91-1, CNRS/Université Paris XI)

Péry-Woodley, M.-P. (1990a). Contrasting discourses: contrastive analysis and a discourse approach to writing. *Language Teaching*, **23**(3), 143-151.

Péry-Woodley, M.-P. (1990b). De la langue aux discours: recherches sur l'analyse des textes d'apprenants. In B. Schneuwly & J.-P. Bronckart, Eds., *Diversifier l'enseignement du français écrit (IVe Colloque International de Didactique du Français Langue Maternelle)*, pp. 329-335. Neuchâtel, Paris: Delachaux & Niestlé.

Péry-Woodley, M.-P. (1991a). French and English passives in the construction of text. *Journal of French Language Studies*, **1**(1), 55-70.

Péry-Woodley, M.-P. (1991b). Writing in L1 and L2: analysing and evaluating learners' texts. *Language Teaching*, **24**(2), 69-83.

Péry-Woodley, M.-P. (1993a). *Les écrits dans l'apprentissage. Clés pour analyser les productions des apprenants.* F Références. Paris: Hachette.

Péry-Woodley, M.-P. (1993b). *Textual clues for user modelling in an intelligent tutoring system*. Notes et Documents LIMSI 93-21, CNRS/Université Paris XI.

Péry-Woodley, M.-P. (1994). Une pragmatique à fleur de texte: marques superficielles des opérations de mise en texte. In S. Moirand, A. Ali Bouacha, J.-C. Beacco & A. Collinot, Eds., *Parcours linguistiques de discours spécialisés*, pp. 337-348. Berne: Peter Lang.

Péry-Woodley, M.-P. (1995). Quels corpus pour quels traitements automatiques? *TAL*, **36**(1-2), 213-232.

Péry-Woodley, M.-P. (1998). Signalling in written text: a corpus-based approach. In M. Stede, L. Wanner & E. Hovy, Eds., *Actes, COLING 98 (Workshop on Discourse Relations and Discourse Markers)*, pp. 79-85. Montreal, ACL.

Péry-Woodley, M.-P. (à paraître). *Cadrer* ou *centrer* son discours ? Introducteurs de cadres et centrage. *Verbum*, **22**(1).

Péry-Woodley, M.-P. (1996b). Syntaxe et discours : l'expression du but. In S. T. Paribakht, H. Seguin, M.-C. Tréville & R. Williamson, Eds., *L'expression orale et écrite : recherche, enseignement, technologie*, pp. 20-25. Ottawa: University of Ottawa.

Péry-Woodley, M.-P. & Rebeyrolle, J. (1998). Domain and genre in sublanguage text: Definitional microtexts in three corpora. In A. Rubio, N. Gallardo, R. Castro & A. Tejada, Eds., *Actes, First International Conference on Language Resources and Evaluation*, pp. 997-992. Granada, ELRA.

Prince, E.F. (1981). Toward a Taxonomy of Given-New Information. In P. Cole, Ed. *Radical Pragmatics*, pp. 223-255. New-York: Academic Press.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language.* London & New York: Longman.

Rebeyrolle, J. & Péry-Woodley, M.-P. (1998). Repérage d'objets textuels fonctionnels pour le filtrage d'information: le cas de la définition. In *Actes, Rencontre Internationale sur l'Extraction, le Filtrage et le Résumé Automatiques*, pp. 19-30. Sfax, Tunisie.

Rebeyrolle, J. (à paraître). Utilisation de contextes définitoires pour l'acquisition de connaissances à partir de textes. In *Actes, Ingénierie des Connaissances (IC'2000)*, Toulouse.

Redeker, G. (1990). Ideational and pragmatic markers of discourse structure. *Journal of Pragmatics*, (14), 367-381.

Reichman, R. (1984). Extended person-machine interface. *Artificial Intelligence*, **22**, 157-218.

Reinhart, T. (1980). Conditions for Text Coherence. *Poetics Today*, **1**, 161-180.

Régent, O. (1985). A comparative approach to the learning of specialized written discourse. In P. Riley, Ed. *Discourse and Learning*, London, New York: Longman.

Régent, O. (1992). Pratiques de communication en médecine: contextes anglais et français. *Langages*, **26**(105), 66-75.

Riegel, M. & Tamba, I. (1987). Définition directe et indirecte dans le langage ordinaire : les énoncés définitoires copulatifs. *Langue Française*, (73), 29-53.

Riegel, M. (1990). La définition, acte du langage ordinaire - De la forme aux interprétations. In J. Chaurand & F. Mazière, Eds., *La définition*, pp. 97-110. Paris: Larousse.

Rodrigues Faria Coracini, M-J. (1992). L'hétérogénéité dans le discours scientifique français et brésilien: un effet persuasif. *Langages*, (105), 76-86.

Roulet, E., Auchlin, A., Moeschler, J., Rubattel, C. & Schelling, M. (1985). *L'articulation du discours en français contemporain.* Berne: Peter Lang.

Ryckman, T. (1990). De la structure d'une langue aux structures de l'information dans le discours et dans les sous-langages scientifiques. *Langages*, (99), 21-28.

Sager, N. (1986). Sublanguage: Linguistic phenomenon, computational tool. In R. Grishman & R. Kittredge, Eds., *Analyzing language in restricted domains. Sublanguage description and processing*, pp. 1-18. Hillsdale, N.J.: Laurence Erlbaum.

Sager, N., Friedman, C. & Lyman, M.S., Eds. (1987). *Medical language processing. Computer management of narrative data.* Reading, MA.: Addison-Wesley.

Schiffrin, D. (1981). Tense variation in narrative. *Language*, **57**, 45-62.

Schiffrin, D. (1987). *Discourse markers.* Cambridge: Cambridge University Press.

Shopen, T. & Williams, J.M. (1981). *Styles and Variables in English.* Cambridge, MA: Winthrop.

Sinclair, J., Ed. (1987). *Looking Up: An Account of the Cobuild Project in Lexical Computing.* London: Collins.

Sinclair, J., Hanks, P., Fox, G., Moon, R. & Stock, P., Eds. (1987). *Collins COBUILD English Language Dictionary.* Glasgow: Collins.

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Sleeman, D. & Brown, J.S. (1982). *Intelligent Tutoring Systems.* New York: Academic Press.

Smith, N. (1997). Improving a Tagger. In R. Garside, G. Leech & A. McEnery, Eds., *Corpus Annotation: Linguistic Information from Computer Text Corpora*, pp. 137-150. London: Addison Wesley.

Stenström, A.-B. (1986). What does *really* really do? In G. Tottie & I. Bäcklund, Eds., *English in speech and writing: a symposium*, pp. 149-164. Stockholm: Almqvist and Wiksell.

Strawson, P. (1971). *Logico-linguistic Papers.* London: Methuen.

Sueur, J.-P. (1982). Pour une grammaire du discours. *Mots*, (5), 143-185.

Swales, J. (1990). *Genre analysis: English in academic and research settings.* Cambridge: Cambridge University Press.

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

Virbel, J. (1985). Langage et métalangage dans le texte du point de vue de l'édition en informtique textuelle. *Cahiers de Grammaire*, **10**, 5-72.

Virbel, J. (1989). The Contribution of Linguistic Knowledge to the Interpretation of Text Structures. In J. André, V. Quint & R. K. Furuta, Eds., *Structured Documents*, pp. 161-181. Cambridge: CUP.

Virbel, J. (1999). *Structures textuelles - planches, fascicule 1 : Énumérations*. Rapport IRIT, Toulouse.

Walker, M.A. (à paraître). Vers un modèle de l'interaction du centrage avec la structure globale du discours. *Verbum*, **22**(1),

Walker, M.A., Iida, M. & Cote, S. (1994). Japanese Discourse and the Process of Centering. *Computational Linguistics*, **20**(2), 193-232.

Walker, M., Joshi, A. & Prince, E. (1998). Centering in Naturally Occurring Discourse: An Overview. In M. Walker, A. Joshi & E. Prince, Eds., *Centering Theory in Discourse*, pp. 1-28. Oxford: Clarendon Press.

Werth, P. (1984). *Focus, Coherence and Emphasis.* Dover, New Hampshire: Croom Helm.

Index