Analyse automatique de structures thématiques discursives - Application à la recherche d'information

Frédérik Bilhaut

UNIVERSITÉ DE CAEN / BASSE-NORMANDIE
U.F.R. de Sciences
École doctorale S.I.M.E.M.

THÈSE

présentée par

M. Frédérik Bilhaut

et soutenue

le 14 juin 2006

en vue de l'obtention du

DOCTORAT de l'UNIVERSITÉ de CAEN
Spécialité Informatique
(Arrêté du 25 avril 2002)

ANALYSE AUTOMATIQUE DE STRUCTURES THÉMATIQUES DISCURSIVES

Application à la recherche d'information

Composition du jury

M. Benoît Habert	Professeur, Université Paris X (rapporteur)
Mme. Adeline Nazarenko	Professeur, Université Paris XIII (rapporteur)
M. Michel Charolles	Professeur, Université Paris III (examinateur)
M. Philippe Laublet	Maître de Conférence, Université Paris IV (examinateur)
Mme. Marie-Paule Péry-Woodley	Professeur, Université Toulouse II (examinateur)
M. Jacques Vergne	Professeur, Université de Caen (examinateur invité)
M. Patrice Enjalbert	Professeur, Université de Caen (directeur)

Remerciements

Je souhaite tout d'abord exprimer à Patrice Enjalbert, qui a dirigé mes travaux depuis le DEA, ma plus profonde gratitude pour sa disponibilité, ses conseils clairvoyants, son soutien sans faille et la confiance qu'il a bien voulu m'accorder. Je souhaite sincèrement à tous ceux qui se lanceront dans l'aventure de la thèse d'être épaulés comme je l'ai été durant ces quatre années.

Je tiens également à remercier vivement Adeline Nazarenko et Benoît Habert pour avoir accepté de m'accorder un temps précieux en tant que rapporteurs de cette thèse, ainsi que Marie-Paule Péry-Woodley, Michel Charolles, Philippe Laublet et Jacques Vergne de me faire l'honneur de participer au jury.

Parmi les membres du GREYC avec qui j'ai eu le plaisir de travailler au cours de ces quelques années, je souhaite tout particulièrement adresser mes remerciements et toute mon amitié à Yann Mathet, Thierry Charnois, Stéphane Ferrari, Éric Faurot, Nadine Lucas, Jacques Madelaine, et Nicolas Hernandez, avec une mention particulière à Antoine Widlöcher, tant pour les nombreuses discussions partagées que pour son investissement dans nos travaux communs. Un grand merci également à Mauro Gaio dont les conseils avisés sont pour beaucoup dans ma décision d'entreprendre cette thèse.

Je souhaite aussi témoigner de ma sympathie et de ma gratitude aux membres de l'ERSS avec qui j'ai eu le plaisir de collaborer durant cette thèse, pour l'intérêt porté à nos travaux communs et leur accueil toujours chaleureux. Merci notamment à Marie-Paule Péry-Woodley, Mai Ho-Dac, Laure Sarda et Andrée Borillo, avec une mention particulière pour Marion Laignelet.

Je salue également tous ceux qu'il m'a toujours été agréable de rencontrer au hasard d'un couloir ou d'un colloque, et plus particulièrement Arnaud Soulet, Julien Lesbuegueries, Franck Dumoncel, Pierre Lousteau, Emmanuel Giguet, Christophe Turbout, Frédérique Loew Pellen, Christophe Pimm et François Rioult.

Mes plus affectueux remerciements vont évidemment à toute ma famille, et tout d'abord à mes parents qui m'ont toujours soutenu et encouragé dans tout ce que j'ai entrepris. J'adresse également un clin d'oeil tout particulier à mon frère Olivier ainsi qu'à Françoise, Hélène et Jean.

Et bien sûr, toutes mes pensées vont à ma compagne Anne, que je remercie tendrement pour sa patience et tout l'amour qu'elle me porte. Rien de ce que j'ai entrepris d'important n'aurait pu se réaliser sans son soutien indéfectible, pour lequel je lui suis infiniment reconnaissant.

Table des matières

Introduction

Introduction

Les travaux dont nous allons faire état dans ce mémoire s'inscrivent dans une dynamique de recherche conduite au GREYC par Patrice Enjalbert, centrée sur les approches dites *sémantiques* du traitement automatique du langage naturel (Enjalbert, 2005c). De façon générale, les activités de l'équipe formée autour de cette question s'attachent tout particulièrement aux méthodes qui permettent d'analyser et de représenter automatiquement des éléments de sens à partir du texte, en corrélation avec des phénomènes linguistiques bien déterminés. L'équipe nourrit également un intérêt tout particulier pour le grain « document » (Enjalbert et Gaio, 2004), et tend à favoriser des démarches expérimentales et/ou liées à des visées applicatives concrètes. Parmi les recherches récemment réalisées dans ce contexte, on citera notamment les travaux portant sur la sémantique spatiale (Mathet, 2000), le calcul de la référence (Dupont, 2003) ou encore du temps et de l'aspect (Person, 2004).

Notre premier contact avec ces problématiques prit la forme d'un travail « initiatique » portant d'une part sur la reconnaissance et l'analyse sémantique d'expressions temporelles, et d'autre part sur le développement d'outils permettant de projeter sur corpus des grammaires telles que celle développée à cette occasion. Comme nous le verrons par la suite, ces deux tâches dessinent l'approche duale que nous avons tenté de conserver tout au long des travaux ici relatés : un premier volet sera dédié l'élaboration de modèles et de procédés de traitement automatique des langues, quand un second concernera le développement d'outils facilitant leur mise en oeuvre et leur capitalisation.

Ces travaux ont pris racine au sein du projet GeoSem[1] (Enjalbert, 2005b), traitant du traitement sémantique de textes géographiques, et plus précisément de la mise en oeuvre d'analyses profondes des textes et des cartes dans les documents géographiques. Il s'agit notamment, pour ce qui est du texte, d'analyses d'ordre sémantique et discursif aboutissant *in fine* à une indexation sémantique et intra-documentaire des documents, exploitable en navigation et en recherche d'information (dorénavant RI). Ce projet possède donc des objectifs applicatifs bien précis, qui visent bien sûr l'élaboration de systèmes susceptibles de rendre des services effectifs aux utilisateurs concernés (principalement géographes), mais aussi et surtout à faire apparaître, tout en se laissant guider par des tâches spécifiques mais complexes, des perspectives de recherche bien plus générales. La démarche a semble-t-il porté ses fruits, puisqu'au delà du développement d'outils qui seront effectivement testés sur des utilisateurs, le projet a consolidé de fécondes collaborations inter-disciplinaires et fait naître dans l'équipe des préoccupations nouvelles, et notamment, pour ce qui nous concerne, celle de l'analyse automatique du discours.

Outre le traitement de la dimension temporelle, notre apport à ce projet concerne principalement le problème de la mise en relation en discours des références à des faits socio-géographiques avec leur localisation spatio-temporelle, tâche qui constitue une forme particulière d'analyse thématique visant à établir une indexation par passages sous la forme de triplets « phénomène-espace-temps ». Corrélativement, nous avons développé un prototype de moteur de recherche capable d'exploiter les résultats obtenus dans le cadre du projet et de permettre leur expérimentation *in situ*. Nous avons ainsi

[1] Programme CNRS interdisciplinaire « Société de l'information » ayant fait l'objet d'une collaboration entre le GREYC, l'ERSS (Toulouse), le LIUPPA (Pau), l'ESO (Caen), et l'EPFL (Lausanne)

été confronté à la problématique de la recherche d'information sous un angle peu commun, apportant son lot de perspectives théoriques et nous menant – du moins l'espérons-nous – en marge de certains sentiers battus.

Par suite, l'essentiel de notre travail a porté sur des problèmes de traitement automatique des langues découlant de ce projet, touchant essentiellement à l'analyse du discours. Nous nous sommes notamment intéressé à la tâche dite d'*analyse thématique*, qui vise l'étude de la structure des textes selon des critères relatifs à la répartition de leur contenu informationnel. Ces travaux concernent donc le problème de la segmentation automatique du discours ainsi que la représentation des thèmes des segments textuels que nous identifions. Nous nous sommes pour cela appuyés sur des modèles essentiellement linguistiques, en portant la plus grande attention aux apports théoriques provenant d'autres disciplines telles que la linguistique ou les sciences de l'information. Nous avons également cherché à prendre en compte des considérations touchant à la représentation des connaissances, en considérant notamment les interactions qui peuvent se produire entre l'organisation des connaissances d'un domaine et la structure du discours qui s'y rapporte.

***

Notre **première partie** consistera, après avoir posé la problématique générale de l'accès assisté à l'information, à envisager différentes approches des notions de *thème*, de *topique*, de *sujet* ou encore d'*à propos*. Il est bien sûr totalement impossible d'être exhaustif dans cette entreprise, et nous avons choisi de rendre compte de façon assez précise, parmi les travaux relevant de différentes disciplines, d'un nombre limité de propositions susceptibles de trouver une application pertinente aux problèmes qui nous occupent. Il s'agira en premier lieu de recherches touchant aux sciences de l'information et à l'ingénierie documentaire, qui nous permettront de prendre la mesure de toute la complexité d'une notion de thème définie en termes d'à propos, ainsi que de la diversité des approches que l'on peut légitimement revendiquer en la matière. Nous serons à cette occasion invités à considérer le concept d'à propos comme intrinsèquement relatif et dépendant de considérations pragmatiques voire épistémologiques. Ce panorama sera hétéroclite à dessein, et nous n'hésiterons pas à mentionner des approches relativement peu répandues comme celles qui reposent sur la formalisations logique, ou la théorie des facettes.

Nous considérerons dans un second temps différents travaux en linguistique, où la notion de thème est bien sûr abondamment débattue. Dans ce domaine, le terme recouvre des concepts parfois très différents, et nous nous concentrons sur ceux qui font intervenir la notion d'à propos, et qui sont donc susceptibles de faire écho aux propositions précédentes. Nous tenterons de les aborder par ordre de grain linguistique, de la proposition au discours, en évoquant des notions de *structure informationnelle*, de *progression thématique*, de *thème discursif*, ou encore d'*encadrement du discours*. Finalement, nous envisagerons des travaux portant explicitement sur l'analyse thématique en traitement automatique des langues. Il s'agira plus précisément de deux approches foncièrement opposées, dont la confrontation montre bien la largeur du spectre des méthodes applicables au problème : nous parlerons d'une part des procédés de la famille *text-tiling*, avant de discuter des méthodes qui se fondent plutôt sur la *modélisation linguistique*.

Après avoir dressé un bilan de ce parcours bibliographique pour en dégager un certain nombre de lignes de force, nous consacrons la **seconde partie** à la description de systèmes et de modèles de traitement automatique des langues que nous avons élaborés en nous appuyant sur elles. Hormis les travaux portant sur la *recherche d'information géographique* déjà évoqués, nous traiterons des deux axes de recherche touchant à l'analyse automatique du discours auquel nous nous sommes tout particulièrement intéressé. Le premier concerne l'*analyse automatique des cadres de discours spatiaux et temporels*. Nous nous basons ici sur l'hypothèse de l'encadrement de Michel Charolles, qui décrit des segments dits « cadres de discours », homogènes par rapport à un critère sémantique spécifié par

une expression détachée constituant un marqueur d'indexation permettant de « répartir les contenus propositionnels dans des blocs homogènes relativement à un critère spécifié par le contenu de l'introducteur », fonction dont on voit immédiatement l'intérêt dans le contexte de la RI. Nous proposons une méthode développée en collaboration avec l'ERSS, qui permet de reconnaître automatiquement les bornes de ces cadres, en proposant de premières solutions au délicat problème de l'analyse automatique de la portée de leurs introducteurs. Nous décrirons également le procédé d'évaluation que nous avons commencé à mettre en oeuvre pour tester la pertinence des résultats obtenus et surtout dégager des perspectives d'amélioration.

Le second point concernera la notion de *thème composite*, que nous avons développée dans le prolongement direct des travaux précédemment décrits, et sur laquelle nous nous appuyons pour mener l'analyse thématique d'une certaine variété de structures discursives liées à une notion relativement libérale d'univers de discours. Nous présenterons le modèle en lui-même avant de décrire la méthode d'analyse thématique automatique qui en découle. Nous envisagerons également les liens entre ce modèle et des concepts existants tels que l'encadrement du discours, la structure informationnelle ou encore la théorie de la structure rhétorique, avant d'introduire la notion d'*axe sémantique* que nous posons comme pivot entre l'organisation des connaissances d'un domaine et la structure thématique des textes qui s'y rapportent.

La **troisième et dernière partie** sera consacrée à *LinguaStream*, logiciel développé parallèlement aux travaux précédemment évoqués pour faciliter leur élaboration, avant de devenir une plate-forme générique pour le traitement automatique des langues. Elle a pour ambition de simplifier la réalisation d'expériences non triviales sur corpus, ainsi que le cycle d'évaluation / ajustements qui en découle. Sans outil adapté, le coût de développement induit par chaque nouvelle expérience devient en effet un frein considérable à l'approche expérimentale, et pour répondre à cette problématique, LinguaStream facilite la mise en oeuvre de procédés complexes tout en impliquant un investissement technique minimal. Elle se fonde pour cela sur la notion de *chaîne de traitement*, assemblage de modules d'analyse de types et de niveaux variés, et sur l'utilisation systématique des langages déclaratifs pour formaliser la connaissance linguistique.

À travers cette plate-forme qui se veut avant tout un « laboratoire virtuel » pour le TAL, nous proposons un certain nombre de principes méthodologiques applicables aux problématiques de l'annotation des documents électroniques et surtout de la constitution de procédés d'analyse complexes fondés sur la formalisation de modèles d'ordre linguistique. Nous développerons notamment la notion de *perspective d'analyse*, qui vise à associer à chaque composant d'un traitement un point de vue spécifique sur le texte, ou encore l'opportunité d'exploiter la *complémentarité des modèles d'analyse*, qui permet de faire collaborer différents formalismes préférentiellement adaptés à un ou des niveaux linguistiques particuliers.

Nous conclurons enfin en tentant de définir un certain nombre de perspectives de recherche qui ressortent de l'ensemble de ces travaux.

Première partie

Accès à l’information : de l’index au thème

Nous débuterons cette partie en posant la **problématique générale** de l'accès assisté à l'information. Nous entamerons ensuite un panorama bibliographique autour de la notion de thème, en commençant par faire état de recherches touchant aux **sciences de l'information** et à l'**ingénierie documentaire**, où nous discuterons tout d'abord de l'ambivalence du concept d'à propos, qui se confronte à la difficulté de situer *en pratique* une frontière franche entre « ce dont on parle » et « ce que l'on dit », et qui peut conduire à hésiter entre une vision du thème comme *ossature* synthétisant l'information véhiculée, et à une vision du thème comme *support* de cette information. En adoptant plutôt ce second point de vue, nous nous concentrons ensuite sur les propositions qui approchent la notion de thème comme *point de contact* entre un document (ou une base documentaire) et un (groupe d') utilisateur(s), entre l'information véhiculée et le « socle informationnel » sur lequel elle s'appuie, entre les présupposés du scripteur et les connaissances du lecteur. Nous considérerons par suite d'autres approches qui nous invitent à considérer la notion de thème et la relation d'à propos comme intrinsèquement relatives, dépendant de considérations pragmatiques voire épistémologiques, mais dont on peut néanmoins définir et étudier les propriétés de façon rigoureuse et formelle, par exemple avec des outils logiques.

Nous considérerons dans un second temps différents travaux en **linguistique**. Dans ce domaine, le terme recouvre des concepts parfois très différents, et nous nous concentrerons sur ceux qui font intervenir la notion d'à propos, et qui sont donc susceptibles de faire écho aux propositions précédentes. Nous tenterons de les aborder par ordre de grain linguistique, de la proposition au discours. Il s'agira tout d'abord de théories liées à la *structure informationnelle*, qui consiste essentiellement à étudier le rôle thématique des constituants de la proposition ou de la phrase. Nous envisagerons par la suite différentes théories portant sur des grains plus importants, notamment avec la notion de *progression thématique* au niveau inter-phrastique, puis celle de *thème discursif*. Enfin, nous considérerons des travaux qui ne sont pas explicitement liés à la notion de thème, mais qui sont d'une grande importance dans l'analyse du discours, dont la *théorie du centrage*, celle de la *structure rhétorique*, et surtout l'hypothèse de l'*encadrement du discours*.

Finalement, nous envisagerons des travaux portant explicitement sur l'analyse thématique en **traitement automatique des langues**. Il s'agira plus précisément d'un positionnement relatif aux deux grandes catégories de méthodes développées dans ce contexte. Nous évoquerons tout d'abord les approches de la famille *text-tiling*, qui s'appuient sur des méthodes quantitatives pour aboutir à une segmentation thématique linéaire. Nous évoquerons d'autre part les approches plutôt « linguistiques », et détaillerons une méthode visant à obtenir une structuration thématique hiérarchique en se fondant sur la dualité thème/rhème.

Nous conclurons ce parcours bibliographique par un bilan visant à en faire émerger un certain nombre de lignes de forces qui sous-tendent notre vision du thème comme objet *discursif*, *sémantique*, *structuré*, et lié à la structure des *connaissances* liées à un domaine.

Chapitre 1

La problématique de l'accès à l'information

Ce chapitre a été pour partie rédigé « à deux plumes » avec Patrice Enjalbert, pour publication dans (Enjalbert, 2005c).

1.1 Introduction

Le problème de la recherche d'information se présente, de prime abord, comme la recherche de documents pertinents par rapport à un besoin exprimé par un utilisateur, d'où le terme de recherche documentaire (RD). La tâche d'indexation a alors pour objectif de produire, pour chaque document, une « fiche descriptive » qui le caractérise au sein d'une base documentaire. Lors d'une recherche d'information, le contenu de ces « fiches » sera confronté à la requête de l'utilisateur, afin de déterminer le ou les documents pertinents.

Le besoin d'indexer a bien sûr précédé l'avènement des systèmes informatiques, cette tâche étant initialement réalisée manuellement par les documentalistes. Classiquement, le résultat de l'indexation est un ensemble, éventuellement pondéré, de mots-clés ou descripteurs. L'objet ainsi indexé est généralement le document dans son intégralité, en adéquation avec la nature indivisible des livres et autres documents papier qui constituent alors les bases documentaires.

L'apparition du document électronique a ouvert la voie à l'indexation automatique, tout en faisant apparaître d'immenses besoins, notamment sur le réseau Internet qui constitue une base documentaire à la fois gigantesque, mouvante, hétérogène et décentralisée. Les premières approches reproduisent le principe de l'indexation manuelle : elles visent à extraire de manière automatique, pour chaque document, une liste de descripteurs qui sera confrontée (de manière également automatique) aux requêtes des utilisateurs. Elles sont aujourd'hui encore massivement employées, tant par les moteurs de recherche disponibles sur Internet que dans les systèmes d'information associés à des fonds documentaires plus classiques comme les bibliothèques.

Ces systèmes, que nous dirons de « première génération », ont fait la preuve d'une réelle efficacité et rendent quotidiennement des services immenses. Ils présentent cependant des imperfections manifestes, perçues de tout utilisateur des moteurs de recherche du Web, et sont plus que jamais dépassés par notre capacité à produire et diffuser l'information. Il en résulte une activité de recherche particulièrement intense, qui se définit d'abord comme une amélioration des techniques d'indexation précédemment évoquées, mais également et de plus en plus, par de nouveaux objectifs susceptibles de diversifier et d'enrichir nos modes d'accès à l'information.

Plutôt que de retourner à l'utilisateur des documents entiers, on cherchera désormais à fournir à l'utilisateur un accès direct à l'information recherchée. Ainsi, certains systèmes (parfois dits de « seconde génération »), visent à sélectionner dans les documents des passages considérés comme pertinents (extraction de passages). Ces passages sont habituellement constitués d'un petit groupe de phrases ou de paragraphes, et sont censés contenir l'information recherchée par l'utilisateur. Plutôt que d'extraire effectivement ces passages de leur contexte, on cherchera généralement à orienter l'utilisateur vers des zones précises, le reste du document restant accessible si nécessaire. D'autres systèmes, parfois qualifiés de « troisième génération », visent à répondre à une question relativement précise (*Question Answering*, ou systèmes de question/réponse). Dans ce cas, il ne s'agit plus de retourner une portion de la base documentaire pertinente par rapport à un « thème général », mais bien de produire une réponse explicite, en langage naturel, à la question posée.

À ces objectifs, qui restent fidèles au paradigme de la réponse à une requête, viennent s'ajouter d'autres approches d'accès assisté à l'information. Le résumé automatique, par exemple, vise à produire automatiquement une forme condensée d'un document. L'aide à la navigation intra-documentaire, quant à elle, entend assister l'utilisateur dans sa recherche d'information au sein même du document : une aide particulièrement pertinente pour des documents longs (articles de revue, rapports administratifs, pouvant atteindre aisément plusieurs dizaines de pages...). Soulignons que si ces diverses catégories et appellations sont communément invoquées pour distinguer les différentes approches, leurs frontières demeurent très mal définies. Le modèle « idéal » d'accès assisté à l'information reste bien sûr à inventer, et il y a fort à parier qu'il tirera simultanément parti de leurs atouts respectifs.

L'élargissement de la problématique de recherche d'information (dorénavant RI), comme l'amélioration de la recherche documentaire « traditionnelle » elle-même, encouragent par ailleurs le rapprochement de méthodes provenant d'univers scientifiques et techniques auparavant distincts. Si les techniques vectorielles demeurent dominantes dans une large partie de la communauté scientifique en RD, de nouvelles méthodes se font leur place, provenant du traitement automatique des langues, de l'extraction d'information, de la représentation des connaissances, etc.

Un autre aspect à prendre en compte, qui peut être vu comme transversal aux approches évoquées plus haut, est l'émergence du Web sémantique. Derrière cette appellation se cache avant tout un idéal qui vise, notamment sur Internet, à limiter la prédominance de la forme et à associer aux ressources du réseau des informations sémantiques directement accessibles et manipulables par des systèmes automatiques. Ceci peut inclure des informations telles que le prix et le fabricant d'un article sur un site de commerce électronique, aussi bien que des métadonnées (auteurs, dates, thèmes...) et des annotations des documents. C'est bien sûr ce dernier point qui nous intéressera. On assiste ici, encore à une convergence intéressante de différentes communautés. L'initiative la plus centrale est sans doute celle du World Wide Web Consortium (W3C), qui procède à la normalisation de formats de données, notamment sous la forme du Ressource Description Framework (RDF), métamodèle pour la représentation des connaissances et l'annotation sémantique. Mais il ne s'agit là que d'un formalisme, et concernant l'annotation de documents électroniques, la réalité du « Web sémantique » dépendra probablement pour beaucoup des avancées de la recherche en TAL et en modélisation des connaissances.

Ce sont ces nouvelles orientations de la RI, croisant applications et méthodes, que nous allons ici présenter. Nous commencerons par présenter plus en détail les objectifs et méthodes de la RD « traditionnelle », ce qui tiendra lieu de point de référence pour mesurer les avancées présentées ensuite. Nous examinerons alors des approches qui, tout en demeurant fidèles au paradigme de la RD, font intervenir des méthodes de TAL et de représentation des connaissances. Nous envisagerons ensuite des extensions, telles que les systèmes de question/réponse, qui exploitent la structure linguistique des requêtes et dans lesquelles l'influence de l'extraction d'information est manifeste. Nous aborderons également la recherche de la structuration du document, avec des applications au résumé automatique, à la navigation intra-documentaire ou encore à l'extraction de passages. Finalement, nous tenterons de situer ces nouvelles tendances dans le contexte du Web sémantique.

1.2 Moteurs de recherche de « première génération »

Les systèmes d'indexation automatique que nous dirons de première génération, et qui sont encore aujourd'hui les plus utilisés, calquent le modèle de l'indexation manuelle : chaque document est indexé dans sa globalité par un ensemble de descripteurs. Les différentes méthodes existantes se différencient principalement par deux caractéristiques : la nature des descripteurs et la méthode utilisée pour les extraire.

Nature des descripteurs. Alors que l'indexation manuelle repose habituellement sur un thesaurus, cette méthode est difficilement automatisable en toute généralité. En effet, un thesaurus vise par définition à restreindre l'espace des index, en établissant une liste de termes (dits « vedettes ») en relation univoque avec les concepts retenus par l'indexeur. Il appartient donc à l'indexeur d'établir un lien entre le vocabulaire, significativement plus large, effectivement employé dans un document et les termes du thesaurus. Or, dans le cadre de l'indexation automatique, cette dernière phase nécessite une ressource lexicale exhaustive, couvrant non seulement les termes du thesaurus mais aussi le vocabulaire de tous les documents analysés. Idéalement, des systèmes complexes, d'ordre sémantique, devront intervenir, par exemple un système de désambiguïsation sémantique permettant de sélectionner le « sens adéquat » de chaque mot polysémique. Cette approche est praticable dans des contextes bien délimités (par exemple, un domaine de spécialité), pour lesquels il est envisageable de fournir au système toutes les informations nécessaires, et qui autorisent les temps de calculs nécessaires. En revanche, dans beaucoup d'autres cas, on préférera sacrifier la qualité de l'indexation au temps de calcul et à l'économie de mémoire.

Les systèmes automatiques déployés à grande échelle indexent donc directement à partir des formes trouvées dans le document lui-même, généralement des mots. La méthode la plus simple consiste à considérer ces mots sous la forme exacte où ils apparaissent, ce qui est évidemment peu efficace, ne serait-ce qu'en raison des différents facteurs de variabilité morphologique, notamment flexionnelle et dérivationnelle. Rappelons que la flexion d'une forme désigne des modifications marquant des traits grammaticaux (comme le nombre ou le cas pour un nom, le temps ou la personne pour un verbe). La dérivation lexicale permet la construction de plusieurs mots de natures potentiellement différentes à partir d'une même racine, par exemple par l'adjonction de préfixes et de suffixes. Ainsi, une racine commune est partagée par le nom « suite », l'adjectif « suivant », et les verbes « suivre » et « poursuivre ». Ces phénomènes font partie de la longue liste des mécanismes linguistiques qui conduisent à l'expression de sens proches par des formes différentes, qu'il est donc particulièrement préjudiciable de considérer indépendamment. De fait, on souhaitera généralement qu'une recherche sur le mot « mer » aboutisse à des textes contenant les mots « mers », « marin », « maritime », etc.

Pour ces raisons, différentes méthodes d'analyse morphologique ont donc été intégrées en tant que traitement préliminaire à l'indexation proprement dite. La première étape consiste souvent en l'élimination de certains mots considérés comme non significatifs pour la tâche de recherche d'information, le plus souvent parce qu'ils sont utilisés systématiquement et ne sont donc pas discriminants. Il pourra s'agir de mots dits « grammaticaux » ou « vides » (comme les déterminants), et parfois de termes considérés comme non significatifs dans un domaine particulier (comme « aéronef » dans le domaine aéronautique). Les étapes suivantes visent essentiellement à obtenir, pour chaque mot retenu, une forme plus générale que l'on espère moins dépendante d'une textualisation particulière du sens exprimé. La troncature consiste à couper le début ou la fin des mots afin d'obtenir une forme approchée de leur racine (par exemple « mang- » pour « manger », « mangé », « mangeoire »). Elle s'opère sans connaissance linguistique, à partir d'heuristiques simples (par exemple, en supprimant les n derniers caractères de chaque mot). Si l'inexactitude du résultat obtenu induit un certain bruit, cette approche est facilement opérationnalisable à grande échelle. A l'inverse, les méthodes d'extraction de racine et de lemmatisation font intervenir un processus plus complexe ainsi que des connaissances linguistiques.

On s'intéressera plus particulièrement aux secondes, qui permettent d'obtenir une forme canonique des mots, comme leur lemme. Leur intérêt pour l'indexation est évident, mais ce type d'opération peut nécessiter une analyse relativement poussée, le plus souvent basée sur des ressources lexicales importantes, bien sûr spécifiques à chaque langue. Ces ressources sont habituellement constituées d'un lexique exhaustif spécifiant, pour chaque forme, l'ensemble des lemmes possibles (dictionnaire de formes fléchies). Cette association étant dépendante du contexte grammatical, la lemmatisation est en général couplée à l'analyse des parties du discours (ou *part-of-speech tagging*) qui vise à déterminer la nature morphologique de chaque mot. Il faut toutefois noter que des méthodes efficaces, bien qu'utilisant des ressources minimes, existent aujourd'hui (Vergne et Giguet, 1998).

Méthode d'indexation et de recherche. Une fois déterminée la nature des unités linguistiques utilisées comme descripteurs, différentes méthodes d'indexation peuvent être distinguées. La plus simple est l'indexation booléenne. Elle est fondée sur l'utilisation d'un fichier inverse, indiquant pour chaque descripteur la liste des documents dans lesquels il apparaît, éventuellement accompagnée d'un nombre d'occurrences. Ce fichier permet alors de trouver quels documents contiennent les mots de la requête, et éventuellement leur fréquence ou autres informations associées, en tenant compte d'éventuels opérateurs booléens (comme la conjonction, la disjonction et la négation), par exemple : (« President » OU (« Bill » ET « Clinton »)) ET (« Monica » OU « Lewinsky »). L'estimation de la pertinence d'un document peut se faire en fonction du nombre de mots de la requête qui y apparaissent, et/ou du nombre d'occurrences de ces mots.

Le modèle vectoriel apporte un certain nombre de raffinements, en indexant chaque document par un vecteur de longueur égale au nombre total de descripteurs de la base documentaire. Il permet ainsi d'associer, pour un document donné, une valeur numérique à chaque descripteur. Cette valeur est une pondération dépendant généralement de la fréquence d'apparition du descripteur dans le document, valant 0 si le descripteur n'y apparaît pas. Une méthode élémentaire consiste considérer la fréquence « brute » (Luhn, 1957), mais alors des mots très communs, et donc peu discriminants, seront exagérément valorisés. On préférera donc des méthodes attribuant à un descripteur un poids d'autant plus important qu'il apparaît de manière « spécifique » dans le document par opposition au reste de la base (Jones, 1972), comme le très classique coefficient $tf \cdot idf$ dont nous serons amené à reparler.

L'indexation produit ainsi, pour une base documentaire contenant n documents et m descripteurs, une matrice de dimension $n \times m$ associant un poids à chaque couple document/descripteur. De la même manière, un vecteur est associé à la requête, contenant une valeur non nulle pour chaque descripteur cité. La phase de recherche fait alors intervenir un calcul vectoriel permettant d'évaluer la « distance » entre le vecteur caractérisant la requête et celui caractérisant chaque document. Là encore, diverses méthodes ont été envisagées, comme le produit scalaire ou diverses distances angulaires (cf. le coefficient de Salton (Salton, 1975)). La plupart des systèmes de recherche de première génération s'appuient sur des émanations du modèle vectoriel, et sont souvent hybrides, par exemple pour offrir à l'utilisateur la possibilité d'utiliser des opérateurs booléens. D'autres modèles existent, que nous ne détaillerons pas ici, comme le modèle probabiliste qui évalue la pertinence d'un document donné en le comparant à d'autres documents dont on connaît la pertinence relativement aux mêmes critères.

En conclusion de ce bref tour d'horizon des méthodes appliquées par les moteurs de première génération, attardons-nous un instant sur quelques particularités des systèmes de recherche aujourd'hui disponibles sur le Web. Il faut notamment avoir à l'esprit quelques ordres de grandeur auxquels ils sont confrontés, et qui contraignent très fortement la nature des méthodes susceptibles d'y être appliquées. Le moteur Google, par exemple, indexe plusieurs milliards de documents, répond à plusieurs centaines de millions de requêtes par jour (avec plusieurs milliers de requêtes par seconde en pointe), et ce, à partir d'un index de plusieurs milliards de descripteurs. La réactualisation complète de l'index prend plusieurs semaines, le moteur d'indexation traitant quelques milliers de pages par jour. Même si le dimensionnement matériel est en conséquence, le traitement d'un tel volume de données est soumis à d'importantes contraintes calculatoires. C'est pourquoi, les moteurs déployés à une telle échelle

appliquent encore aujourd'hui les méthodes décrites plus haut, qui nécessitent relativement peu de ressources en temps et en mémoire. Ils mettent toutefois en oeuvre différentes techniques permettant, sans en altérer les principes de base, d'améliorer la qualité de l'indexation. Aussi, la plupart des moteurs de ce type prennent-ils en compte les quelques méta-informations communément incorporées aux pages Web, en attribuant par exemple un poids particulier aux descripteurs trouvés dans le titre, l'adresse, l'en-tête ou autres métadonnées. La structure du maillage entre pages Web peut également être utilisée, ce qui est notamment le cas du système *PageRank* de Google : lors du calcul de pertinence, celui-ci privilégie les documents présentant le plus grand nombre de liens entrants ou sortants vers d'autres pages (par propagation des valeurs obtenues sur ces dernières selon le même principe)[1]. Enfin, il est à noter que les méthodes linguistiques élémentaires telles que décrites plus haut, comme la lemmatisation, tardent encore à apparaître dans les moteurs déployés à grande échelle, bien que certains aient annoncé l'introduction prochaine de procédés de racinisation dans leurs systèmes.

1.3 Moteurs de première génération « et demi »

Nous venons de voir que les moteurs de première génération, limités en temps de calcul et en ressources, arrêtent habituellement leur analyse à la surface du texte. Nous avons également constaté que cette approche se heurte immédiatement aux différents phénomènes linguistiques qui conduisent à la textualisation de sens proches par des formes différentes. Nous avons enfin évoqué des méthodes telles que la racinisation ou la lemmatisation, qui permettent de prendre en compte un premier niveau de variabilité morphologique. Nous allons voir maintenant comment certaines informations de sémantique lexicale peuvent améliorer considérablement la pertinence des résultats de la recherche.

Un phénomène lexical important est la synonymie, qui intervient quand des sens voisins sont exprimés par des formes n'entretenant a priori aucun lien apparent. Par exemple, les mots « bateau » et « navire » ont des sens très proches sans qu'aucune similarité morphologique ne l'indique. Et même si ces sens ne sont pas rigoureusement équivalents, on considérera généralement comme souhaitable qu'un moteur de recherche retourne les documents relatifs à l'un quand une requête contient l'autre. Mais un système informatique ne peut établir ce lien sans disposer des ressources nécessaires, par exemple sous une forme équivalente à un dictionnaire des synonymes. Cette méthode s'étend à d'autres relations lexicales, telles que l'hypo/hyperonymie ou la méronymie : ainsi, une requête portant sur des « navires » devrait récupérer des documents parlant de « bateaux » (synonyme) mais aussi de « voiliers » (hyponyme), et peut-être de « mâts » ou d'« hélices » (méronymes). On parlera d'expansion de requêtes.

Les systèmes informatiques capables de prendre en compte ce phénomène, constituent souvent des extensions aux mécanismes surfaciques envisagés plus haut, où la relation d'équivalence entre deux mots ne repose plus seulement sur leur forme, leur racine ou leur lemme, mais exploite des connaissances permettant de déceler des sens voisins. Pour représenter ces connaissances, deux types de modèles sont couramment utilisés : les graphes conceptuels et les espaces sémantiques vectoriels. Les premiers – bien connus en intelligence artificielle – sont des graphes où chaque forme connue constitue un noeud, et où les liens sémantiques sont représentés par des arcs. Ils peuvent être dédiés à un domaine de spécialité (ressource terminologique) ou bien généralistes comme WordNet (Miller *et al.*, 1990). Dans une telle structure, la distance sémantique entre deux formes peut, par exemple, s'obtenir à partir de la longueur du plus court chemin qui les lie dans le graphe, éventuellement en pondérant les différents types d'arcs.

Les espaces sémantiques vectoriels sont des espaces de grande dimensionalité, où chaque dimension correspond à un trait sémantique. Il s'agit donc d'un cadre formel proche du modèle vectoriel

[1] Il convient toutefois de noter que ces méthodes sont fort contestées, car elles peuvent être facilement détournées par des tiers pour fausser les résultats de l'indexation.

recherche documentaire « classique », mais cette fois les dimensions de l'espace ne correspondent plus à des formes (mots), mais à des valeurs sémantiques. Chaque entrée est alors représentée par un vecteur de cet espace, la valeur selon une dimension étant donnée par une valeur numérique témoignant d'un « niveau d'activation » du trait correspondant. Pour mesurer la « similarité sémantique » entre deux mots, on calcule une « distance » entre les vecteurs qui les représentent. Plus deux mots sont sémantiquement proches, plus les deux vecteurs qui les représentent pointent dans la même direction (là encore, différentes mesures de distance angulaire sont applicables). Remarquons que les ressources nécessaires à ces méthodes peuvent être constituées manuellement (Crestan *et al.*, 2004), mais que ce modèle se prête volontiers une acquisition automatisée. Par exemple, l'analyse sémantique latente (ASL) permet de construire automatiquement un espace sémantique vectoriel à partir de l'analyse statistique des cooccurrences dans un corpus suffisamment large (Landauer *et al.*, 1998). Précisons que dans ce dernier cas, l'espace obtenu n'a aucune valeur descriptive, car le système n'a pas la faculté de caractériser ni même de nommer les traits sémantiques associés aux différentes dimensions. Il ne pourra donc pas être contrôlé ni modifié, et peut être vu comme une « boîte noire » permettant de calculer une distance entre deux termes.

1.4 Limites des systèmes « traditionnels » et perspectives « nouvelles »

Une limitation fondamentale des systèmes de recherche documentaire traditionnels est qu'ils s'arrêtent, du point de vue de l'utilisateur, à la porte du document : le document est retourné comme un tout, opaque, au sein duquel il reste à rechercher la ou les informations visées. Ceci pose deux types de problèmes.

i) Ces informations peuvent être relativement précises et localisées, et on aimerait d'une manière ou d'une autre connaître les zones de texte où elles sont le plus probablement susceptibles de figurer. Le problème est crucial en ce qui concerne les documents longs (Péry-Woodley, 2005) : rapports, ouvrages, ou parfois articles de journaux.

ii) On peut aussi souhaiter prendre connaissance de l'ensemble du document, mais néanmoins être guidé dans son parcours : avoir un aperçu de son plan, des sous-thématiques abordées, les conclusions de l'auteur, etc. Il s'agit d'un travail de « lecture rapide » que nous opérons quotidiennement, mais dans lequel on pourrait souhaiter être guidé. L'organisation logique du document (en chapitres, sections, etc.), la structure des titres et sous-titres constitue certes une aide importante. Mais, il resterait encore à l'exploiter véritablement dans des procédures de recherche automatiques (autrement qu'en sur-pondérant les titres dans un calcul de score dans une indexation). Par ailleurs, l'expérience montre que cette structure reste en général d'un grain élevé sur la plan thématique, et que l'on pourrait souhaiter disposer de structures plus fines.

Tout cela nécessite clairement des traitements d'ordre discursif, s'attachant par exemple à l'organisation thématique et/ou argumentative. Une tendance relativement récente vise de fait l'analyse automatique d'éléments de la structure « interne » des documents. L'objectif est de proposer à l'utilisateur des outils de navigation intra-documentaire permettant de repérer rapidement les segments porteurs des informations spécifiques visées et/ou de prendre connaissance de l'organisation d'ensemble. On pourra alors considérer le texte comme composé de segments que l'on pourrait appeler « unités documentaires », homogènes thématiquement, possédant une forte cohésion interne, et inter-reliés de différentes manières : successions de thèmes correspondant à différentes vues sur un sujet général, thèse/développement, argumentaire, introduction et synthèse, etc. Du point de vue linguistique, nous sommes ainsi conduits vers une sémantique du discours appliquée à l'ingénierie documentaire.

1.5 Recherche d'information par requêtes structurées

Un premier type d'élargissement de la recherche documentaire « standard » consiste à passer d'une recherche par mots-clés à des requêtes structurées. Cela implique d'une part de procéder à une analyse de la structure de la requête, et d'autre part, à une analyse similaire du texte de manière à y trouver des « motifs informationnels » susceptibles d'y correspondre. On peut distinguer deux types de systèmes répondant à ce schéma, selon que le système retourne, comme en RD, un document (ou un passage) relatif au contenu de la requête, ou qu'il fournisse une réponse immédiate à une question du type « qui / quand / où / etc. ». Nous parlerons dans le premier cas de « recherche documentaire structurée », tandis que les seconds sont connus comme systèmes de question/réponse (ou Q/R pour *Question Answering*).

Un exemple de recherche documentaire structurée est le système Facile (Ciravegna *et al.*, 1999), qui permet à l'utilisateur de formuler des requêtes telles que : « les contrats passés par des institutions financières européennes dont le montant dépasse un million d'euros ». Le système utilise en amont des méthodes standard (statistiques) de classification d'articles, utilisant les mots de la requête sans tenir compte de sa structure grammaticale, pour sélectionner des articles (ici financiers) potentiellement pertinents. Mais la requête elle-même, c'est-à-dire le critère de recherche qu'elle formule, échappe visiblement à ces méthodes. En revanche, on voit clairement que l'information cherchée s'apparente au contenu d'une fiche que constituerait un système d'extraction d'information (Enjalbert, 2005a), avec une entité de type « transaction » associée à un champ « lieu » de valeur « Europe » et un champ « montant » de valeur supérieure à la valeur donnée de 1 ME. Le système détermine donc à partir de la requête le format d'une telle fiche et applique, sur les documents retenus, les méthodes de l'extraction d'information pour la remplir. En cas de succès, le document répond positivement à la requête.

Le même type de démarche est à l'oeuvre pour une recherche incluant des critères spatiaux et temporels dans le projet GeoSem, où l'on peut formuler sur une collection de documents géographiques des requêtes incluant un critère de localisation spatiale et/ou temporelle. Par exemple : « Où parle-t-on du retard scolaire dans l'Ouest de la France au cours des années 1950 ? », ou encore : « Où parle-t-on de l'évolution des politiques de sécurité maritime en Manche au cours des 20 dernières années ? ». Les critères de recherche spatiaux et temporels (« dans l'Ouest de la France », « au cours des années 1950 », etc.) ne sont évidemment pas réductibles à une liste de mots-clés, mais doivent identifier une région ou une plage temporelle, de manière à faire correspondre, par exemple, « Rennes » et « la Bretagne », ou encore « 1956 » et « les années 60 ». Une analyse syntaxico-sémantique de ces expressions peut toutefois être réalisée automatiquement, par des méthodes inspirées de la compréhension automatique, mais avec des objectifs évidemment considérablement simplifiés. En résultera une représentation symbolique des « zones » spatiales et temporelles évoquées par le texte. La requête sera traitée de la même manière, et ce sont donc deux représentations symboliques, et non des segments de textes, qui seront comparées. L'analyse est donc ici d'ordre syntagmatique et non purement lexical. En outre, la cible privilégiée étant constituée de documents longs, on recherchera à retourner des passages plutôt que des documents en tant que tels. Il s'agit là d'une problématique déjà mentionnée (extraction de passages), mettant bien sûr en jeu la recherche de la structure discursive du document.

En ce qui concerne les systèmes de question/réponse, le changement majeur tient à la forme même des requêtes, et par suite à la nature des réponses : on formulera des questions en « qui / quoi / quand / pourquoi » plutôt que des requêtes classiques du type « trouver les documents où l'on parle de X ». La figure 1.1 reproduit des exemples de telles questions. La réponse pourra être de différentes natures : il pourra s'agir d'une entité nommée (personne, lieu, date), d'une fiche de type extraction d'information (synthétisée à partir des documents et constituant une réponse effective à la question), ou encore des phrases ou segments courts extraits des textes de la base documentaire, et contenant en principe une réponse à la question. On se dirige ainsi vers « l'idéal » de l'interrogation d'une base documentaire, qui se présente alors à l'utilisateur comme une quelconque base de données. On notera que cette tâche est devenue une des rubriques importantes du programme TREC.

(1) Où est le Taj Mahal ? Quelle est la population actuelle de Tucson ? Qui était le premier secrétaire d'Etat de Nixon ? Qui a gagné le prix Nobel de Littérature en 2002 ? Que publie Knight Ridder ?

(2) Biographie résumée de X (personnage public) ? Nom, prénoms, adresse, date de naissance, formation ? Que savons-nous de la société Y ? Structures organisationnelles, lignes de produits, dirigeants ?

(3) Nommer 30 personnes ayant participé au cabinet de R. Reagan. Quels sont les acteurs du film Z ? Nommer 4 pays producteurs de diamants.

(4) Quel cépage est utilisé dans le Château Pétrus ? Combien coûte le cru 1999 ? Où le propriétaire a-t-il fait ses études ? Quel domaine possède-t-il en Californie ? Combien existe-t-il d'espèces d'araignées ? Combien sont venimeuses ? Quel pourcentage de piqûres sont fatales ?

(5) Comment faire pour copier un fichier ? Comment installer un logiciel ? Quelle commande permet de créer un répertoire ? Comment compiler un programme ?

FIG. 1.1 – Exemples de requêtes en question/réponse, d'après (Carbonell *et al.*, 2000) et (Ferret *et al.*, 2001a), pour (1) et (2), (Voorhees et Harmann, 2001) pour (3) et (4), et (Molla *et al.*, 2000) pour (5).

Une méthode assez répandue peut être décrit comme suit. En premier lieu, il convient d'analyser la question, dont on repérera deux éléments. Il s'agira d'une part du *focus*[2] (que l'on pourrait aussi appeler « pivot »), qui est la composante de la question à laquelle l'information cherchée relative, et qui devra donc impérativement figurer dans la réponse. Par exemple, dans « Qui a gagné le prix Nobel de Littérature en 2002 ? », le focus sera le prix Nobel de Littérature (éventuellement en 2002). Il s'agira d'autre part du type de la réponse attendue : « personne » pour une question en « qui », « lieu » pour une question en « où », etc. Mais aussi, « cause » ou « raison » pour une question en « pourquoi », ce qui ne va bien sûr pas sans poser une série de questions non triviales.

La seconde phase consiste en une analyse des documents de la base documentaire (qui auront pu être filtrés auparavant par des méthodes de RD classique), pour laquelle différentes stratégies peuvent être employées. On aura généralement un extracteur d'entités nommées permettant de repérer des entités correspondant au type de la question. Au-delà, les variantes apparaissent. Certains systèmes sont plus proches de la RD classique. Ainsi, Laszlo *et al.* (2000) cherchent une fenêtre de longueur fixe contenant une entité du type attendu, et s'appariant au mieux avec les autres mots de la requête (dont le focus), l'estimation se faisant par un calcul de score comme en RD.

En dehors de ce schéma courant, diverses méthodes sont bien sûr possibles. Ferret *et al.* (2001c) emploient une méthode plus structurée, en recherchant dans les textes des motifs syntaxiques (ou syntaxico-sémantiques) de la forme « GN (focus) Connecteur GN (réponse) », où le connecteur est le verbe de la question, par exemple « publier » pour « Que publie Knight Ridder ? ». On trouve par ailleurs dans (Srihari et Li, 2005) une méthode plus proche de l'extraction d'information, quand une analyse encore plus profonde est réalisée par le système Extrans (Molla *et al.*, 2000), dont la tâche est de répondre à des questions portant sur des documents très techniques comme les « man pages » d'Unix.

[2]Sans rapport explicite avec la notion de focus en linguistique, cf. chapitre 3.

1.6 Résumé automatique

Une problématique devenue classique en TAL, visant à permettre à un lecteur de prendre rapidement connaissance de « l'essentiel » d'un document, est celle du résumé automatique. Il s'agit là clairement d'une autre manière d'assister un utilisateur dans son accès à l'information, et qui peut très bien se combiner avec des procédures de recherche documentaire plus classiques, ces dernières rassemblant des documents pertinents sur un sujet donné, que le résumé automatique pourra présenter sous forme synthétique.

Sur le plan des méthodes, une première approche a, historiquement, pris racine dans les travaux en compréhension automatique. On l'appellera couramment de ce fait « approche par compréhension ». Si l'on est capable de produire une représentation conceptuelle d'un texte, alors on peut aussi imaginer en calculer par des méthodes d'IA une représentation « condensée », que l'on restituera par un système de génération de texte. Par exemple, à partir d'un texte tel que « Jean prend sa voiture et va chez le disquaire. Il choisit un CD pour Marie. », on reconnaîtra un scénario « offrir » et on pourra produire le résumé : « Jean offre un CD à Marie ». Toutefois, les difficultés posées par la compréhension automatique « lourde » ou « profonde » étant patentes (Charnois et Enjalbert, 2005), on considéra en général que cette approche est hors d'atteinte.

On se tourne aujourd'hui plutôt vers des méthodes dites « par extraction », visant à composer un résumé à partir de segments judicieusement choisis pour leur caractère représentatif du texte : phrases ou paragraphes en général. Cette « représentativité » pourra évidemment se décliner de différentes manières, et être évaluée grâce à différents types d'indices. L'approche est donc en rapport avec les techniques de segmentation et de caractérisation précédemment exposées, mais elle est « sélective » plutôt que systématique et globale.

Là encore, un premier type de méthodes se situe dans la lignée de la RD : il s'agira de calculer un score pour chaque phrase pour conserver celles dont ce score est supérieur à un certain seuil. Une première possibilité est d'utiliser un calcul de fréquence comme test de représentativité. Par exemple, on pourra prendre comme score la somme des valeurs de l'indicateur de type $tf \cdot idf$ des mots de la phrase ou du paragraphe. Une autre stratégie opère par calcul de similarité en utilisant le modèle vectoriel : un vecteur de descripteurs est associé à chaque paragraphe, et un calcul d'angle permet de calculer une distance entre paragraphes deux à deux. On pourra alors, par exemple, sélectionner les unités dont les facteurs de similarité avec les autres est la plus grande (selon différents modes de calcul concrets possibles).

D'autres méthodes repèrent des formes linguistiques particulières, signalant des passages du texte à mettre en valeur dans l'optique d'un résumé : par exemple, des annonces thématiques ou des expressions conclusives (Paice, 1981; Déclès et Minel, 2000). De simples critères positionnels peuvent également être exploités : phrases en début de paragraphe, première occurrence d'une entité nommée, etc. Des marques d'intégration évoquées ci-dessus peuvent aussi intervenir dans le repérage de certaines articulations entre segments, et ainsi améliorer la cohésion du résumé. Remarquons encore que les deux types de méthodes (numériques et linguistiques) peuvent être combinées (Saggion et Lapalme, 2002).

Mais on peut aussi étendre la notion de résumé, en gardant l'idée de restitution condensée de l'information portée par un texte. Une approche originale, qualifiée par ses auteurs de « fortement linguistique », est proposée par (Boguarev et Kennedy, 1997). Elle peut être également présentée comme « référentielle » dans la mesure où une de ses caractéristiques est d'être centrée sur la découverte d'entités saillantes. Au lieu de chercher à repérer des phrases représentatives du texte, on applique des méthodes d'extraction de terminologie pour trouver des termes (groupes nominaux) représentatifs ; un calcul d'anaphore, en mesurant le nombre de reprises de ces termes, permet de repérer les plus saillants, c'est-à-dire ceux qui dénotent les entités les plus représentatives. Les ruptures dans le degré

de saillance signalent un changement de thématique et permettent de déclencher une segmentation du texte. On proposera alors comme caractérisation du contenu du texte (dénomination remplaçant ici celle de résumé), les termes désignant les entités les plus saillantes, accompagnés d'extraits les contenant. Nous avons donc là un exemple « typique » d'exploitation d'un type particulier de structure discursive, les chaînes de coréférence, pour une tâche concrète apparentée au résumé.

1.7 Segmentation et structuration thématique

Cette tâche vise à déterminer une segmentation thématique *a priori*, indépendamment de toute requête. Il s'agit plus précisément de découper le texte en une succession de segments thématiquement homogènes, de caractériser ces segments en termes de contenu, et éventuellement de calculer certaines formes d'organisation les reliant. Différentes méthodes d'analyse thématique automatique seront plus amplement discutées dans un prochain chapitre, et nous ne le détaillerons donc pas ici. Rappelons simplement les deux « familles » que l'on identifiera généralement pour catégoriser les différentes approches proposées :

- Les méthodes habituellement qualifiées de quantitatives ou numériques se fondent plus ou moins explicitement sur la notion de cohésion lexicale, en exploitant la répétition des mots comme indicateur d'homogénéité thématique. Il s'agit notamment des travaux se plaçant dans la lignée de (Youmans, 1991) et (Hearst, 1994), qui procèdent à une segmentation linéaire du texte, c'est-à-dire en segments adjacents et non enchâssés. Nous qualifierons dorénavant de *text-tiling* cette famille d'approches, du nom de la méthode de Hearst.
- Les méthodes souvent qualifiées de « linguistiques » exploitent marqueurs, indices, et plus généralement formes linguistiques et dispositionnelles, porteurs d'indications de la structure thématique et plus généralement discursive. Un exemple est donné par la notion de cadre thématique , possédant un introducteur tel que « en ce qui concerne X », « à propos de X », « considérant X », etc., et dont on peut voir qu'il introduit en général plusieurs phrases, ainsi présentées comme « relatives au thème X ».

Nous verrons dans le chapitre 4 des exemples concrets de méthodes appartenant à ces deux catégories, tout en établissant notre propre position par rapport à elles.

1.8 Le Web sémantique

Aucun processus informatique ne peut appréhender « le monde » sans se baser sur un modèle prédéterminé, en se bornant au traitement des données qui s'y conforment. Or, la quasi-totalité des données disponibles sur le Web sont formatées selon des modèles dont la portée est très limitée, et dont les spécifications sont d'ailleurs assez peu scrupuleusement observées. Ces modèles portent principalement sur la représentation des caractères (ASCII, Unicode, etc.) et la mise en forme du texte (HTML, bientôt remplacé par XHTML et autres vocabulaires XML). De ce fait, l'immense masse d'information accessible sur le Web ne l'est que pour des humains, et reste paradoxalement opaque pour les processus informatiques, qui n'accèdent directement qu'à la surface des documents.

La préoccupation de « sémantiser » le Web, c'est-à-dire de se donner les moyens de rendre son contenu facilement accessible à des systèmes automatiques, n'est pas nouvelle, précédant largement le récent engouement autour de l'appellation de « Web sémantique ». Ainsi, les initiatives de standardisation menées notamment par le W3C ont depuis longtemps suggéré l'intégration de données « satellites » aux documents web, via la normalisation de formats de métadonnées (Dublin-Core), et la promotion de mécanismes permettant de distinguer la structure logique d'un document de son rendu (visuel ou autre). A ce premier niveau, il s'agit donc de spécifier des informations, que l'on peut dire

sémantiques, sur le document lui-même.

L'objectif de ce que l'on appelle le « Web sémantique » est plus ambitieux, puisqu'il s'agit de spécifier des informations sémantiques potentiellement sur toute ressource du Web. Une ressource peut évidemment être un document, mais aussi tout objet (ou personne) susceptible d'être décrit, présenté. Il pourra, par exemple, s'agir d'un article catalogué sur un site de commerce en ligne, auquel on associera une fiche descriptive contenant prix, description, disponibilité, etc. Cette fiche étant décrite sous un format normalisé, les informations qu'elle contient, exprimées dans un langage formel bien défini selon un modèle clairement identifié, ne seront plus seulement disponibles au lecteur humain sous leur forme textuelle, mais aussi accessibles aux traitements informatiques.

Le formalisme proposé par le W3C est appelé *Resource Description Framework* (RDF). Il s'agit d'un modèle de représentation de données sous forme de graphe, où des objets abstraits, appelés « ressources » et identifiés par des URI (*Unified Resource Identifier*), sont connectés par des relations (dites « prédicats »). Un graphe RDF est communément représenté par l'ensemble de ses arcs (dits « triplets »), chaque triplet spécifiant un sujet (une ressource quelconque), un prédicat, et un objet (une autre ressource, ou bien une valeur littérale). Ainsi, le triplet (Jean, ami-de, Pierre) spécifie que la ressource « Jean » est un ami de la ressource « Pierre », et le triplet (Jean, âge, 30) indique que « Jean » est âgé de 30 ans. Bien évidemment, ce modèle n'apporte aucune nouveauté conceptuelle, puisqu'un graphe RDF n'est rien d'autre qu'un certain type de Réseau Sémantique, formalisme classique et utilisé depuis longtemps en IA. L'intérêt de RDF est surtout d'être une norme flexible, ouverte et rigoureusement spécifiée, permettant d'envisager une interopérabilité à grande échelle.

Toutefois, le formalisme RDF ne permet pas à lui seul d'envisager cette « interopérabilité sémantique » : il reste à s'entendre, pour une application donnée, sur un modèle décrivant la réalité que l'on cherche à manipuler, en s'appuyant sur le cadre générique fourni par RDF. Notons en premier lieu que le W3C propose pour cela les langages RDF Schema et OWL (*Web Ontology Language*)[3], qui permettent de spécifier formellement les classes et les relations à utiliser dans un graphe RDF. On notera également que de nombreuses initiatives visent à définir des modèles applicatifs pour le Web sémantique, comme les récentes versions du *Dublin Core* ou de RSS, ou encore, plus anecdotiques, DOAP (description de projets open-source) ou FOAF (description de liens interpersonnels). Concernant la représentation du temps, on trouve déjà une dizaine de modèles différents[4]...

Que peut-on donc attendre du Web sémantique ? On peut pronostiquer sans grand risque que ces technologies perceront à court terme dans des domaines tels que le commerce électronique. Par exemple, pour peu qu'un vocabulaire RDF de description de catalogues de vente en ligne se généralise, il serait très simple pour les vendeurs d'exporter leur catalogue depuis leurs bases de données vers ce format. Il deviendrait alors relativement aisé pour un moteur de recherche de recueillir ces données pour fournir des services de recherche évolués.

Dans le cadre de la recherche d'information, le problème est plus complexe, et ne serait que très partiellement résolu par l'avènement de techniques que nous venons d'évoquer. En effet, bien que l'on dispose de ces formalismes normalisés, de nombreuses questions importantes demeurent :

i) A quel type de ressource, au sens RDF, doit-on associer des informations sémantiques ? Nous avons vu que le développement des méthodes de RI tend vers une approche intra-documentaire, et que l'on ne peut plus se contenter de décrire les documents dans leur ensemble. Mais alors, à quelles unités textuelles est-il pertinent d'associer des informations sémantiques ?

ii) A quels modèles ces informations devront-elles se conformer ? Au-delà de la complexité de la modélisation elle-même, l'interopérabilité sémantique pose le problème de l'instauration de modèles

[3]Nous ne rentrerons pas ici dans les détails de ces langages, qui sont clairement spécifiés et expliqués sur le site web du W3C (http ://www.w3c.org). Notons simplement que OWL est un raffinement de RDF Schema, explicitement dédié à la spécification d'ontologies.

[4]Voir par exemple, le site : `http://www.daml.org/ontologies`.

partagés. Or, le développement et la standardisation de ceux-ci constituent une gageure tant sociale que théorique. Nous avons par exemple évoqué la prolifération de modèles de représentation de la temporalité : la diversité des préoccupations particulières conduit naturellement à l'élaboration de modélisations totalement différentes de réalités pourtant analogues.

iii) Comment produire ces informations ? Chacun peut constater que les possibilités d'adjoindre des méta-informations aux documents, bien qu'existant depuis bientôt une décennie, ont été faiblement exploitées. En effet, des mécanismes tels que le « Dublin Core » permettent depuis longtemps de spécifier dans les pages Web des informations simples telles que l'auteur du document, son titre ou un ensemble de mots-clés. Relativement à ce que l'on peut imaginer dans le cadre du Web sémantique, le coût de production de ces informations est extrêmement faible, et pourtant elles sont encore très rarement spécifiées, si ce n'est dans un but détourné[5]. En outre, dans le cas d'annotations fines (par exemple, une représentation formelle du sens de chaque expression temporelle du texte, une fiche à la MUC...), on ne peut plus envisager de procéder manuellement, et les traitements sémantiques automatiques s'imposent, avec toutes les difficultés qui en découlent.

Il convient donc de considérer que le Web sémantique constitue avant tout un défi pour les communautés du TAL et de la RI, tout en ouvrant la porte à une collaboration plus étroite entre ces deux disciplines. Un défi, puisqu'il motive le développement d'une approche à la fois sémantique et discursive du document, encore peu explorée, et qu'il amène à reconsidérer les problématiques de modélisation à la lumière des applications concrètes qui pourront être proposées aux utilisateurs. Il encourage également le développement de méthodes applicables en temps raisonnable à des masses documentaires conséquentes, et c'est ici que le TAL peut rejoindre la RI de façon effective. En effet, nous avons vu plus haut que les techniques utilisées pour la RI sur Internet se doivent de privilégier le temps de calcul sur la finesse de l'indexation, et ne peuvent donc mettre en oeuvre que des traitements légers, principalement numériques. Or, l'adoption généralisée d'un format de données et de modèles communs permettrait d'envisager un système distribué, où l'analyse fine des documents serait réalisée sur le site même qui les héberge, et donc sur une masse documentaire limitée. Ainsi, les robots des moteurs de recherche pourraient-ils récolter des annotations sémantiques fines, en complément de l'indexation plein-texte classique[6].

[5]En particulier, les mots-clés sont souvent utilisés pour « piéger » les robots des moteurs de recherche, au point que certains d'entre eux ont été amenés à les ignorer.

[6]On peut certes douter de la faisabilité « sociale » d'un tel système, qui poserait notamment des problèmes de confiance dans les annotations générées et de rentabilité du service ainsi offert par les fournisseurs d'information. Mais, nous ne discuterons pas ici cette question complexe.

Chapitre 2

Notions de thème en ingéniérie documentaire et en sciences de l'information

Nous allons nous intéresser dans cette partie à différents travaux que l'on pourrait situer à la croisée de l'ingénierie documentaire et des sciences de l'information, et qui s'attachent explicitement aux notions de topique, de thème, d'à propos ou encore de pertinence. Précisons que l'ensemble des contributions évoquées ici est hétérogène à dessein, l'objectif étant de dresser un panorama des différents points de vue possibles sur ces questions.

Bien que leur intérêt pratique soit peut-être moins avéré que celui des méthodes « classiques » de la RI, ces travaux nous concernent tout particulièrement dans la mesure où ils s'attachent à la fois au « grain » documentaire tout en questionnant abondamment la notion d'à propos (ou d'*aboutness*). Ce dernier point constitue un apport considérable relativement à beaucoup de travaux en RI, qui font bien souvent l'économie de ce questionnement : comme nous l'avons vu dans le chapitre précédent, le principe de la représentation d'un document par un simple ensemble de descripteurs (en général des « mots ») numériquement représentatifs a tendance à perdurer. La question semble pourtant primordiale dans le contexte de l'accès assisté à l'information, mais demeure extrêmement complexe et ouverte, alors même que la notion de « thème » est couramment manipulée, de façon intuitive, par chacun d'entre nous.

Car si deux observateurs d'un même texte parviennent à s'entendre sur un même « thème » ou « sujet », ce sera généralement à la suite d'un processus dialogique incluant périphrases et reformulations, généralisations ou restrictions conceptuelles, le parcours non linéaire du texte, etc. Et même à l'issue d'une telle négociation, la formulation du thème obtenu pourra rester « fuyante », et fortement dépendante du contexte. Et surtout, l'explicitation objective du processus qui sous-tend l'identification de ce thème reste extrêmement difficile à donner. Devant la complexité de ces tâches, il semble donc nécessaire, pour dépasser les limites des systèmes actuels, d'interroger la nature même du thème ou l'à propos d'un texte, et d'envisager de nouvelles façons de le représenter formellement. Les travaux présentés dans cette section apportent de premiers éléments de réponse à ces questions. Par ailleurs, en s'intéressant expressément au grain documentaire, ils arpentent un terrain relativement peu étudié par la linguistique qui, comme nous le verrons dans le prochain chapitre, s'intéresse généralement à des niveaux de grain beaucoup plus petits.

2.1 Le problème de l'à propos dans l'analyse du document

Le titre de cette section est emprunté à W. J. Hutchins, et plus précisément à l'une de ses contribution souvent considérée comme pionnière dans le domaine (Hutchins, 1977). Il s'agit d'une approche que l'on pourrait qualifier de linguistico-documentaire, et qui croise différents principes linguistiques que nous seront amenés à évoquer dans des sections ultérieures. Du point de vue qui nous intéresse ici, l'intérêt majeur de son approche réside dans le souci constant de restituer ces notions dans le contexte des systèmes d'information et de la recherche documentaire.

2.1.1 À propos et progressions thématiques

Hutchins reprend dans un premier temps la notion de *progression thématique* qui se base sur l'opposition habituelle entre *thème* et *rhème*. Nous ne rentrerons pas ici dans le détail de ces notions dans le mesure où elles seront détaillées par la suite (cf. chapitre 3). Rappelons simplement que ce modèle vise à rendre compte du phénomène d'accumulation de l'information au fil du texte : un segment textuel dit thématique fait référence à « des éléments liés d'une façon ou d'une autre au texte qui précède ou à des propriétés de l'environnement dans lequel se place le discours » ; un segment dit rhématique introduit quant à lui une « information nouvelle pour le lecteur ou le locuteur, ou qui ne découle pas directement de ce qui a été déjà dit ou écrit » (Hutchins, 1977, p. 19). En termes plus intuitifs, on considère généralement que le thème introduit ce dont on s'apprête à discourir, alors que le rhème exprime ce qu'on souhaite en dire (« ce dont on parle » *vs.* « ce qu'on en dit »). Précisons que si ces notions sont souvent appliquées au niveau de la phrase, elles ne sont en rien spécifiques à ce niveau d'analyse, et peuvent être appliquées, comme nous allons le voir, à différents niveaux.

À partir de ces notions, on peut envisager différents modes de progression thématique en considérant l'articulation thème/rhème au fil du texte. Notons que cette notion de progression thématique est là encore indépendante du grain considéré, différents mécanismes discursifs pouvant intervenir pour marquer cette structure. Au niveau phrastique, en particulier, l'articulation reposera fortement (mais pas uniquement) sur les chaînes référentielles, comme on peut l'observer dans les exemples suivants (proposés par Hutchins) :

> **(1)** The boy was reading a book. It was about armadillos. They are found in South America.
>
> **(2)** The boy was reading a book. He had been given it for his birthday. He was ten years old.
>
> **(3)** All substances can be divided into two classes : elementary substances and compounds. An elementary substance is a substance which [...]. A compound is a substance which [...].

Ces exemples font apparaître trois types de progression, explicités dans la figure 2.1. Dans le premier cas, il s'agit d'une progression dite *linéaire*, où chaque élément thématique fait référence au rhème précédent. Dans le second cas, il s'agit d'une progression dite *parallèle*, où le thème reste constant. Le troisième exemple est un cas de configuration hybride, que l'on rencontrera plus fréquemment dans les textes « réels » : la première phrase introduit un rhème double (« elementary substances » et « compounds »), dont chacun est repris successivement (progression parallèle) sous la forme d'une progression linéaire. Notons dès à présent qu'une typologie plus détaillée des progressions thématiques sera abordée en section 3.2.2.

Concernant la portée de ce modèle à d'autres niveaux de grains, Hutchins remarque que dans toute progression thématique telle que celles que nous venons de présenter, « la première phrase constitue le

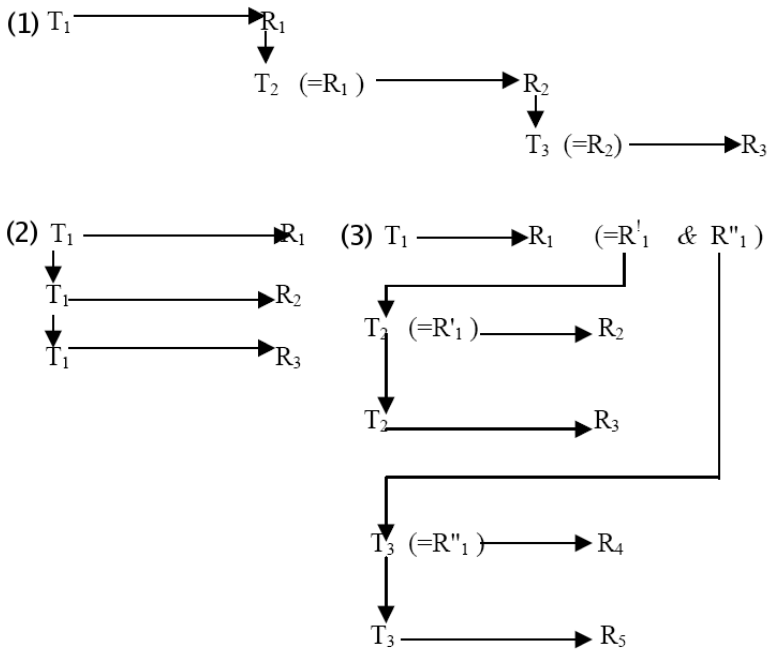


FIG. 2.1 – Exemples de progressions thématiques (pp. 20-21).

point de départ ou le fondement des phrases suivantes. En ce sens, elle peut être considérée comme le thème du paragraphe, les phrases suivantes constituant le rhème » (p. 21). Il rappelle, à ce sujet, le terme de « topic sentence » introduit par (Christensen, 1967) pour désigner la phrase initiale d'un paragraphe. De la même façon, il observe que les premiers éléments d'un texte sont représentatifs de son thème, en indiquant ce dont il sera question dans la suite (au moins pour les textes expositifs). Hutchins propose d'étendre encore ce principe à un niveau extra-documentaire, en postulant une analogie entre l'accumulation d'information qui s'opère dans un texte et le mécanisme de citation qui relie les textes scientifiques : de même qu'une séquence de phrases forme une progression thématique, la littérature scientifique produit, étape par étape, de nouvelles informations en s'appuyant sur des données issues des publications antérieures. L'auteur fournit l'exemple de la figure 2.2, illustrant cette analogie entre la relation thème/rhème et la relation cité/citant.

Si elle peut paraître anecdotique, cette analogie nous mène en fait au coeur de l'argumentaire de Hutchins : elle illustre le fait que la propriété essentielle qu'il attribue au thème n'est pas tant sa position « en initiale » d'un bloc textuel que la fonction qui consiste à invoquer les connaissances nécessaires à la compréhension de ce bloc. Ce point de vue ne lui est bien sûr pas spécifique, puisque la relation « donné »/« nouveau » ainsi que sa corrélation avec les positions initiales est abondamment décrite par la littérature linguistique (cf. section 3.1). Mais ces notions y sont généralement envisagées à des grains discursifs relativement fins, en fonction de l'évolution de l'état attentionnel du lecteur au fil du texte. Pour sa part, Hutchins situe ces notions dans la perspective particulière des systèmes d'information, au niveau documentaire.

Il nous apporte ainsi des éléments essentiels quant aux liens que nous souhaitons établir entre thème et index dans le contexte de la recherche documentaire, en posant la question suivante : si l'on considère que la représentation de l'à propos que l'on cherche à donner pour un document a pour vocation de faciliter l'accès à celui-ci par les utilisateurs d'un système d'information, quelles informations cette représentation doit-elle fournir ? En tout état de cause, il est clair que cette représentation devra tenir compte du texte considéré comme un tout cohérent, dont l'à propos ne se résume pas à la somme des thèmes de ses constituants. Hutchins rappelle à ce sujet la distinction introduite par (Fair-

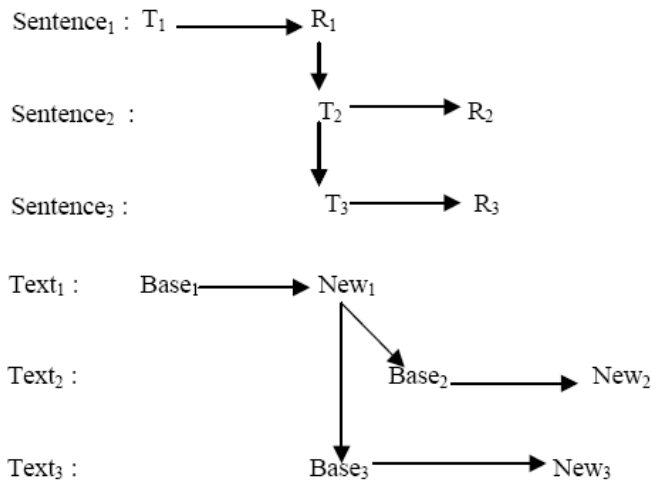


FIG. 2.2 – Analogie entre progression thématique et relations bibliographiques dans la littérature scientifique (p. 30).

thorne, 1969) entre à propos « extensionnel » et « intensionnel »[1], le premier désignant les thèmes des différents constituants du texte, et le second le thème du texte pris dans sa globalité.

Partant du principe que la définition de l'à propos d'un document doit permettre de servir les besoins d'un ensemble d'utilisateurs dont chaque individu dispose de connaissances préalables qui lui sont propres, l'auteur évoque deux approches différentes, l'une en termes de représentation complète de la structure sémantique du texte (approche par résumé), l'autre en termes de connaissances du lecteur telles que présupposées par le scripteur (approche thématique).

2.1.2 Approche par résumé

Dans le premier cas, il s'agirait d'une forme de résumé obtenue à partir d'opérations de généralisation et de réduction appliquées à la micro-structure sémantique, dont les résultats seraient agencés selon la macro-structure du texte pour former une représentation globale de son sens. Les termes de micro- et macro-structure sont ici empruntés à Van Dijk (1972, 1977), et nous aurons l'occasion de revenir plus précisément sur ces notions dans la section 3.3.1. Selon les termes de Hutchins, ces structures correspondent, à différents niveaux, à « des réseaux de propositions[2] reliées par des connecteurs, dont les constituants apparaissent comme arguments d'un certain nombre de propositions, ces occurrences étant inter-reliées par divers moyens anaphoriques » (p. 26).

Plus précisément, la micro-structure d'un texte correspond au « réseau sémantique représentant les propositions qui le composent, leurs relations et leur rôle au sein de la progression thématique globale » (*ibid.*). Sans rentrer dans plus de détails, précisons simplement que Hutchins explicite cette structure à partir des progressions thématiques et des « connecteurs de phrases », reprenant la classification rhétorico-sémantique de ces connecteurs due à (Longacre, 1970), qui n'est pas sans rappeler la classification des relations établie plus tard par la Rhetorical Structure Theory (Mann et Thompson, 1987). En d'autres termes, on peut dire que la micro-structure d'un texte est un objet sémantique complexe, composé des contenus propositionnels des énoncés qui le composent, ces contenus étant liés par

[1] La graphie utilisée dans diverses publications (dont (Hutchins, 1977)) est « intentional » (que nous traduirions par « intentionnel »), mais celle-ci est selon toute vraisemblance erronée, car il s'agit bien ici de la différence entre « extension » et « intension ».

[2] Le terme de »proposition« doit ici être conçu dans son sens sémantique (un prédicat), et non syntaxique.

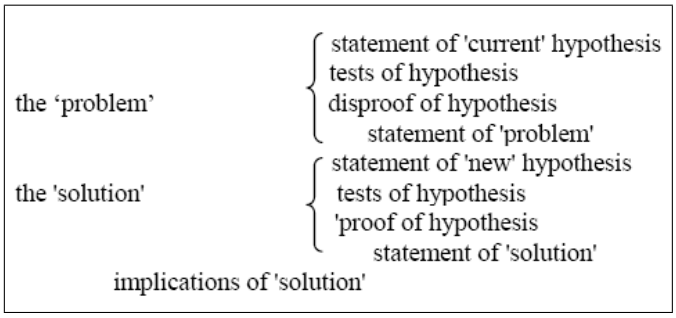


FIG. 2.3 – Exemple de macro-structure trouvée fréquemment dans les articles scientifiques (Hutchins, 1977, p. 25). Les différents items correspondent à des épisodes, dont certains forment d'autres épisodes d'ordre supérieur.

différentes relations issues des différents mécanismes de connexion inter-phrastique.

La macro-structure constitue quant à elle « un réseau sémantique représentant les propositions des épisodes ainsi que les relations qu'elles entretiennent entre elles et au sein de la progression du texte » (*ibid.*). Le terme d'« épisode » correspond ici à une unité logique appartenant à la structure « type » du genre textuel considéré, par analogie avec la notion du même nom dans les structures narratives. En effet, de même qu'une régularité structurelle a pu être observée dans les textes narratifs (par exemple, chez (Longacre, 1974) : ouverture, cadre, déclencheur, développement, apogée, suspense final, clôture), Hutchins suggère la possibilité de définir des régularités analogues pour d'autres genres, donnant par exemple le schéma de la figure 2.3 pour ce qui est des publications scientifiques. Chaque épisode peut lui-même être représenté par une proposition (toujours au sens logique), qui représente de façon synthétique la fraction de la micro-structure appartenant à cet épisode. Ces différentes propositions forment à leur tour la macro-structure, où elles sont liées par des relations propres à la structure globale du texte.

La figure 2.4 illustre notre interprétation du principe du « résumé » qui découle de ce modèle. La notation « G/R » désigne les opérations de généralisation et de réduction évoquées plus haut, qui consistent respectivement à caractériser un épisode par une seule proposition (notée P_i), et à éliminer d'un réseau sémantique les éléments considérés comme inessentiels. On notera, non sans un certain amusement, que l'auteur évoque la possibilité de mécaniser ce processus, alors qu'il est encore aujourd'hui considéré comme inaccessible au traitement automatique (les systèmes actuels de résumé automatique procèdent généralement par extraction et non par reformulation). Terminons sur ce point avec un exemple, donné par Hutchins, de ce que pourrait donner le résumé d'un argumentaire scientifique, une fois reformulé (p. 26) :

> The view that X is true has been shown to be invalidated because of A, B and C. Therefore it is proposed that Y is true. And if Y is true, then Z follows.

2.1.3 Approche thématique

La seconde approche proposée par Hutchins repose sur l'application au niveau documentaire de la dualité donné/nouveau, conformément aux principes évoqués plus haut. Dans cette perspective, la notion d'à propos correspond bien au thème par opposition au rhème, en tant que « point de contact fourni au lecteur par le scripteur, permettant de relier son propos soit à un certain contexte ou environnement, soit à un discours ou texte antérieur » (p. 29). Cette définition en termes de « point de

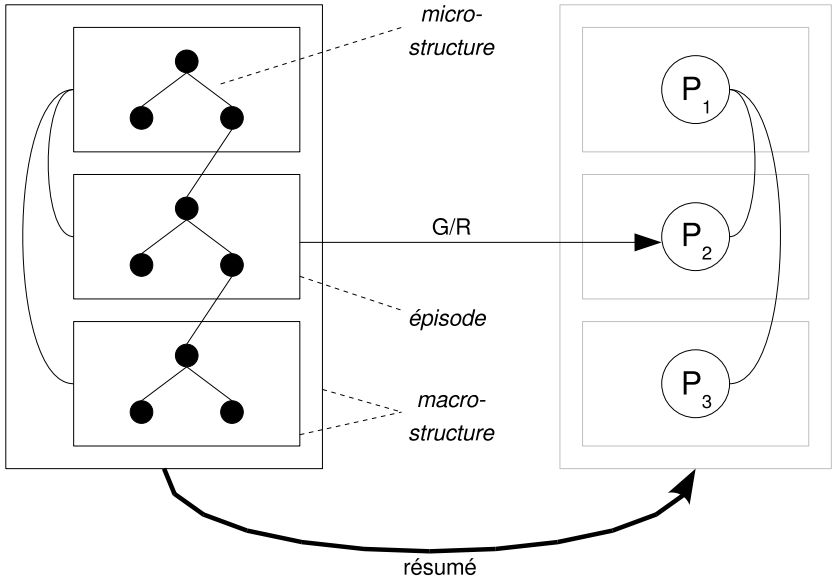


FIG. 2.4 – Processus de résumé par généralisation/réduction, où le contenu micro-structurel de chaque épisode est réduit sous la forme d'une proposition qui prend part à son tour à une structure qui est le reflet de l'organisation globale – en épisodes – du texte.

contact » fait directement intervenir les connaissances supposées du lecteur : « l'à propos d'un document doit être recherché dans ces sections initiales où l'auteur introduit les composants principaux de la macro-structure et établit des points de contact avec ce qu'il suppose être l'état des connaissances[3] des lecteurs potentiels » (*ibid.*).

Ce rapprochement de la notion de thème avec celle de « donné » (ou « connu ») considérée en dehors du texte lui-même a bien sûr une portée particulière dans le contexte de la recherche documentaire. Hutchins rappelle que, par définition, quiconque consulte un système d'information à la recherche d'un document ne peut pas exprimer avec précision ce que devrait être le contenu de ce document : « il ne peut pas spécifier quelle information 'nouvelle' devrait être exprimée par un document pertinent », mais seulement « formuler ses besoins relativement à ce qu'il sait déjà, l'état présent de ses connaissances » (p. 30). En d'autres termes, Hutchins conçoit la pertinence d'un document relativement à la requête d'un utilisateur en fonction du degré de correspondance existant entre la portion thématique de ce document et le « réseau sémantique propre » de cet utilisateur : « ce qu'il cherche, c'est un document initié par une base de connaissances avec laquelle il peut établir des points de contact, un document qui présuppose un état de connaissances présentant des affinités avec le sien » (*ibid.*).

Pour ces raisons, Huthins considère qu'une définition de l'à propos en termes de connu/nouveau devrait être préférée à une approche en termes de résumé du contenu sémantique global, considérant que « l'objectif de l'indexation est de fournir aux lecteurs des points de contact, les guidant depuis ce qu'ils savent vers ce qu'ils souhaitent apprendre », et que le résumé ne permet pas de « prendre en compte [cette] structure communicationnelle élémentaire des textes » (p. 31).

Il est toutefois évident que cette approche se heurte immédiatement à la variété des « états de connaissances » des différents utilisateurs du système d'information. Face à ce problème, Hutchins insiste sur l'importance que revêt « l'environnement d'indexation » : il s'agit pour l'indexeur de prendre en compte un « niveau moyen », valide pour la majorité des utilisateurs, ce qui est plus ou moins réalisable selon que l'on s'adresse à une communauté réduite dont on pourrait caractériser un « savoir

[3] »States of knowledge« dans le texte original.

commun », ou au contraire un large public au sujet duquel il est difficile de faire de telles hypothèses. Il faut également considérer l'influence des autres documents de la collection indexée : « comme pour tout lecteur, l'indexeur ne peut juger des aspects 'nouveaux' d'un texte que sur la base de ses connaissances des autres textes », et doit « chercher à intégrer ce que le document apporte au sein d'un cadre[4] établi » (p. 30).

2.1.4 Conclusion

En conclusion, remarquons tout d'abord que les vues de Hutchins n'étaient pas spécifiquement orientées vers la question de l'indexation automatique, alors très récente : même si le problème de l'automatisation est discuté à plusieurs reprises, il est clair que le terme d'« indexeur » utilisé par l'auteur semble le plus souvent désigner l'opérateur humain. Toutefois, sa contribution paraît aujourd'hui encore susceptible d'alimenter à bien des égards la réflexion dans ce domaine. D'une part, la problématique posée reste tout à fait d'actualité, puisque la communauté de la RI interroge rarement les notions mêmes d'index, de thème ou de pertinence, alors que la question est loin d'être résolue. D'autre part, les éléments de réponse apportés par Hutchins suggèrent plusieurs chemins dont l'exploration dans le cadre de l'accès assisté à l'information nous paraît très prometteuse.

En premier lieu, le choix de poser la dualité donné/nouveau comme une notion totalement indépendante du grain, en la considérant comme valable depuis la phrase jusqu'à la collection de documents, nous semble parfaitement pertinent dans le cadre de la RI. Cette même dualité étant par ailleurs finement décrite par la linguistique aux niveaux phrastiques et discursifs, nous n'hésiterons pas à nous y adosser pour établir des ponts entre certains modèles purement linguistiques et notre approche plus « documentaire ».

D'autre part, la prise en compte explicite d'un certain « contexte d'indexation » constitue en soi un large champ d'investigation. Tout d'abord, même si elle est inapplicable dans des contextes « ouverts » tels qu'Internet, l'idée d'exploiter les caractéristiques d'une classe d'utilisateurs peut se révéler tout à fait pertinente dans le cas où le système s'adresse à une communauté restreinte. En particulier, pour mieux distinguer dans les documents les connaissances présupposées des connaissances « nouvelles », nous pouvons envisager de fournir à un système d'indexation une certaine représentation des connaissances typiques du lectorat considéré. Pour reprendre les termes de Hutchins, il pourrait s'agir d'une certaine forme de « réseau sémantique » composé de concepts susceptibles de constituer des « points de contacts » entre une base documentaire et une communauté d'utilisateurs.

L'idée de prendre en compte la totalité de la collection documentaire pour indexer chaque document est tout aussi réalisable si cette collection est de taille raisonnable. En particulier, il est intéressant de considérer la possibilité d'extraire automatiquement d'une collection de documents une certaine forme de réseau sémantique qui serait propre à une communauté, et qui, en tant que telles, pourrait être utilisée dans la tâche d'indexation elle-même.

Cette perspective nous permet d'envisager une indexation capable de prendre en compte le réseau sémantique propre à *chaque* utilisateur, de façon à obtenir une représentation de l'à propos des documents qui lui soit spécifique. Grâce à des outils permettant à chaque utilisateur de faire évoluer cette représentation des connaissances à partir desquelles il recherche de l'information, cette indexation « ad hoc » pourrait même être spécifique à chaque séance de travail, en fonctions d'objectifs informationnels ponctuels. Ces réflexions sont à la base d'un certain nombre de méthodes et de systèmes que nous développerons dans la partie II. En particulier, nous mettrons en pratique dans le chapitre 8 l'idée de prendre en compte dans l'analyse d'un texte la structuration des connaissances propres à un domaine. Un procédé d'apprentissage sur corpus permettant d'extraire ces connaissances sera également développé dans la section 8.4.4.

[4] »Framework« dans le texte original.

2.2 Approches fondées sur des modèles logiques

Nous allons évoquer dans cette partie une approche du problème de la recherche d'information, née à la fin des années 1980, qui a pour particularité de se baser sur des modèles logiques. Dans cette perspective, la recherche d'information est vue comme un processus de raisonnement logique, permettant d'évaluer l'à propos d'un élément d'information relativement à un autre. Les propriétés de la relation d'à propos sont décrites par un jeu d'opérateurs logiques adéquats, accompagnés d'un ensemble de postulats propre à chaque modélisation du processus de recherche d'information.

Dans cette perspective, de nombreux jeux d'opérateurs et de postulats ont été envisagés, au sein de différents modèles logiques (propositionnelle pure, modale, floue, etc.). Une vue d'ensemble de ces diverses approches étant donnée dans (Lalmas, 1998), nous nous contenterons ici de faire l'écho des travaux de Huibers et de Bruza, qui comptent parmi les contributeurs importants de ce courant. Pour cela, nous nous baserons principalement sur (Huibers, 1996) et (Bruza *et al.*, 2000), dans l'objectif de donner un aperçu concret des principes qui peuvent être développés dans cette perspective. On notera que les contributions dont nous faisons état ici ont pour particularité d'énoncer des principes conçus comme indépendants de tout modèle de RI particulier (booléen, vectoriel, etc.), ce qui les rend en principe assez largement applicables.

Les auteurs se basent sur la notion de « vecteur d'information » (VI), qui désigne habituellement un document ou une requête (l'évaluation de l'à propos portant habituellement sur une requête relativement à un document). Il est important de préciser que le terme « vecteur » est ici employé pour traduire le terme anglais « carrier », sans lien direct avec les modèles vectoriels à la Salton (cf. chapitre 1). Un « vecteur d'information minimal » désigne un élément indivisible d'information, en général un mot-clef, et correspondrait dans une terminologie plus classique à un « concept ». Les auteurs avancent que la relation d'à propos est en premier lieu applicable à ces vecteurs minimaux d'information, considérant par exemple qu'il existe une relation d'à propos entre les mots-clefs « football » et « sport ».

A partir de ces vecteurs minimaux, des objets plus complexes peuvent être formés à l'aide de l'opérateur dit de composition d'information, noté $\oplus$. Etant donnés deux vecteurs A et B, le résultat de leur composition est un nouveau VI correspondant à la « rencontre » de A et B. Les auteurs font l'hypothèse que cette relation est idempotente ($A \oplus A = A$), commutative ($A \oplus B = B \oplus A$), et associative ($(A \oplus B) \oplus C = A \oplus (B \oplus C)$). Il remarquent également que tous les VI ne sont pas composables, observant par exemple que la composition de « flying » et « crocodile » ne produit pas un VI acceptable, contrairement à la composition de « swimming » et « crocodile ». Cette relation, dite de « préclusion », sera notée $A \perp B$.

La relation dite d'inclusion permet d'exprimer le fait que « A contient au moins la même information que B ». Deux nuances de cette relation sont distinguées dans la littérature, l'une dite « de surface » ($\supseteq$), l'autre « forte » ($\mapsto$). Une inclusion de surface $A \supseteq C$ signifie que A est formé par la composition de C avec un autre VI, c'est-à-dire s'il existe B tq. $A = B \oplus C$. On dira par exemple que $avion \supseteq aerien$ s'il on admet que $avion = vehicule \oplus aerien$

La relation d'inclusion forte correspond au cas où l'inclusion se produit au niveau « sémantique », par exemple : $saumon \mapsto poisson$. Selon toute vraisemblance, les auteurs entendent par « niveau sémantique » un niveau inférieur au VI minimal, c'est-à-dire dont on ne peut rendre compte par composition de plusieurs VI. Cela qui n'exclut bien sûr pas qu'il existe un modèle sémantique capable de rendre compte plus précisément d'une relation d'inclusion forte. Par exemple, si on considère que le VI minimal correspond au mot, le modèle componentiel permettra de rendre compte des relations d'inclusion forte. Finalement, la distinction entre inclusions faible et forte paraît essentiellement justifiée par la nécessité *pratique* de distinguer le grain du VI minimal du grain sémantique minimal dont on est capable de rendre compte, mais ne nous semble pas relever de phénomènes fondamentalement différents. Les auteurs définissent d'ailleurs une relation d'inclusion plus générale (notée $\rightarrow$), formée

de l'union de ces deux relations : $A \rightarrow B$ signifie qu'il existe une relation d'inclusion entre A et B, qu'elle soit surfacique ou forte.

Etant donnés ces différents opérateurs, un certain nombre de postulats sont posés (pour tout A et B appartenant à l'ensemble $\mathbb{V}$ des vecteurs d'information) :

- Réflexivité (R) : $A \rightarrow A$.
- Transitivité (T) : $A \rightarrow B$ et $B \rightarrow C$ impliquent $A \rightarrow C$.
- Anti-symétrie (AS) : $A \neq B$ et $A \rightarrow B$ impliquent $\neg (B \rightarrow A)$
- Composition-Inclusion (CC) : $A \oplus B \rightarrow A$ et $A \oplus B \rightarrow B$.
- Absorption (AB) : si $A \rightarrow B$ alors $A \oplus B = A$.
- Inclusion non conflictuelle (NCC) : si $A \mapsto B$ alors $\neg (A \perp B)$
- Préclusion-Inclusion (CP) : si $A \mapsto B$ et $B \perp C$ alors $A \perp C$.

La notion d'à propos en tant que telle est modélisée par une relation binaire entre VI, habituellement notée $\models$. Intuitivement, $A \models B$ signifiera que « A est à propos de B ». Différentes variantes de cette relation ont été envisagées, dont nous verrons des exemples plus loin. Précisons tout d'abord qu'elle n'a pas de lien réel avec l'opérateur logique habituel, même si certains travaux lui ont initialement prêté des propriétés similaires.

A partir du formalisme que nous venons de présenter, (Huibers, 1996) a développé une notion de *système de preuve* appliquée à l'à propos. Tout comme son pendant en logique traditionnelle, ce système opère par inférences à partir d'un ensemble d'axiomes et de règles. La notation utilisée pour décrire ces inférences est celle habituellement utilisée pour représenter les preuves logiques. Par exemple, la notation suivante représente une règle permettant d'inférer B à partir des prémisses A_i :

$$\frac{A_1 \quad ... \quad A_n}{B}$$

2.2.1 A propos par recouvrement

Nous allons dans un premier temps envisager différentes règles d'inférence liées à une définition particulière de l'à propos discutée dans (Bruza *et al.*, 2000). Cette définition est basée sur la notion de *recouvrement*, et répond à une intuition voulant que si deux VI se recouvrent, alors ils sont considérés comme étant l'un à propos de l'autre (p. 5). Les auteurs mettent cette approche en parallèle avec le modèle vectoriel classique, qui évalue le degré de recouvrement un document et la requête à partir d'une mesure angulaire entre les vecteurs qui les représentent[5]. Cette relation d'à propos « par recouvrement » sera notée $\models_o$ et définie comme suit :

$$\forall A, B \in \mathbb{V}, A \models_o B \Leftrightarrow \exists C \in \mathbb{V} | (A \rightarrow C \wedge B \rightarrow C)$$

Dans cette perspective, l'assertion $A \models_o B$, c'est-à-dire « A est à propos de B », est vérifiée s'il existe un vecteur d'information C inclus à la fois dans A et B. Précisons que cette définition, si elle en est évidemment proche, n'est pas équivalente à une mesure « saltonienne » de similarité puisqu'elle donne une simple valeur booléenne là où une mesure angulaire fournit une grandeur exprimant un *degré de similarité*.

Partant de cette définition et des axiomes présentés plus haut, les auteurs énumèrent un certain nombre de propriétés de la relation d'à propos sous la forme de règles d'inférence, donc voici quelques-unes :

Inclusion (C) : $\frac{A \rightarrow B}{A \models_o B}$

[5]Le terme de « vecteur » doit cette fois être considéré dans le sens mathématique du terme (cf. infra).

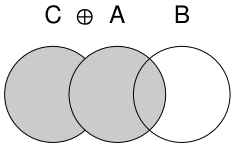
Symétrie (S) : $\frac{A \models_o B}{B \models_o A}$

Monotonie compositionnelle à gauche (LM) : $\frac{A \models_o B}{A \oplus C \models_o B}$

Monotonie compositionnelle à droite (RM) : $\frac{A \models_o B}{A \models_o B \oplus C}$

Conjonction (And) : $\frac{A \models_o B \quad A \models_o C}{A \models_o B \oplus C}$

La justification de ces règles à partir des propriétés des opérateurs primitifs est immédiate. Par exemple, pour la monotonie compositionnelle gauche : par définition, $A \models_o B$ implique qu'il existe D tel que $A \rightarrow D \wedge B \rightarrow D$; par ailleurs la propriété (CC) nous donne $A \oplus C \rightarrow A$, dont nous tirons, par transitivité (T), $A \oplus C \rightarrow D$; finalement, de ce dernier fait et de $B \rightarrow D$ il découle par définition que $A \oplus C \models_o B$. En d'autres termes, il est clair que si A recouvre B, le fait qu'il recouvre aussi C n'affecte pas ce recouvrement :



Nous ne ferons pas état ici de l'ensemble des propriétés proposées par les auteurs, et le lecteur intéressé trouvera dans (Bruza *et al.*, 2000) d'autres propriétés accompagnées de leurs justifications, ainsi que la preuve du fait qu'un sous-ensemble composé des propriétés de réflexivité et de monotonie compositionnelle gauche et droite suffit à caractériser la relation d'à propos par recouvrement.

On retiendra de cette partie une définition relativement « naïve » de la relation d'à propos, que nous serons par la suite amenés à comparer avec d'autres modèles. Cela nous permet également de nous familiariser avec les concepts propres à l'approche logique, que nous pouvons résumer ainsi : étant donnés l'ensemble $\mathbb{V}$ des vecteurs d'information et un ensemble d'opérateurs « primitifs » dotés d'un certain nombre de propriétés, la définition d'une relation d'à propos, et par suite d'un modèle de RI, est donnée par un ensemble de postulats portant sur cette relation. Dans le cas de l'à propos par recouvrement, tous les postulats découlent directement de la définition formelle de la relation. Dans d'autres cas, comme celui que nous allons envisager dans la section suivante, ce sont à l'inverse les postulats posés qui constituent une définition de la relation considérée.

2.2.2 A propos de « sens commun »

Dans la même contribution, les auteurs proposent un modèle plus élaboré de l'à propos, qui serait plus proche des principes habituellement admis par des agents humains. Ce modèle définit une nouvelle relation d'à propos notée $\models$, par le biais là encore d'un certain nombre de postulats prenant la forme de règles d'inférence. Comme dans la section précédente, nous allons là encore nous limiter à certains postulats que nous considérons comme représentatifs de la démarche des auteurs.

En premier lieu, si cette relation conserve la propriété de réflexivité qui était déjà attribuée à $\models_o$, les auteurs postulent que contrairement à cette dernière, elle est fondamentalement asymétrique. Cela répond effectivement à l'intuition voulant que si $football \models sport$, il semble peu pertinent d'admettre la relation inverse $sport \models football$.

Un second postulat, dit de cohérence, permet de s'assurer qu'un VI est « compatible » avec son à propos. Cette notion de compatibilité découle directement de la relation de préclusion introduite plus haut :

$$\frac{A \models B}{\neg (A \perp B)} \quad \text{(AC)}$$

Le postulat suivant reprend la relation d'inclusion en ajoutant une contrainte sur la relation primitive du même nom. Les auteurs se basent pour cela (Brooks, 1995), qui montre que la perception de pertinence entre deux vecteurs d'information est inversement proportionnelle à leur distance sémantique[6]. Pour représenter ce phénomène, les auteurs utilisent la notation $\mapsto_n$ pour désigner une relation d'inclusion forte (cf. infra) établie par au plus n étapes de transitivité. En considérant le nombre 2 comme une limite acceptable quand à la relation d'à propos, on considérera que $A \models B$ si $A \mapsto_2 B$. Ce principe est étendu par les auteurs à la relation d'inclusion générale ($\rightarrow$), pour obtenir le postulat ci-dessous :

$$\frac{A \rightarrow_2 B}{A \models B} \quad \text{(C)}$$

Un autre postulat est une adaptation de la monotonie compositionnelle gauche, dite monotonie « prudente ». Elle est en effet plus restrictive que la règle donnée pour l'à propos par recouvrement, qui autorisait la déduction de $A \oplus C \models_o B$ en partant de $A \models_o B$ seulement. Cette version est plus prudente dans le sens où elle ajoute une nouvelle prémisse à la règle, en demandant que A soit lui-même à propos de C. Dans ce cas, leur composition revient à ajouter à A une information « compatible », ce qui réduit les risques de perte de précision lorsque l'on considère que le résultat de cette composition est elle-même en relation d'à propos avec B. La règle obtenue est représentée ci-dessous, une forme équivalente étant donnée pour la monotonie componisitionnelle droite. Notons que le type de monotonie supportée par une relation d'à propos a un impact important sur les propriétés générales d'une relation d'à propos, ce qui sera discuté dans la section suivante.

$$\frac{A \models B \quad A \models C}{A \oplus C \models B} \quad \text{(CLM)}$$

Dans ce modèle, les auteurs introduisent la relation de « non-à-propos », notée $\not\models$, permettant d'exprimer le fait qu'un VI n'est pas à propos d'un autre VI[7]. De façon générale, l'intérêt pratique de cette relation réside d'une part dans le fait que le non-à-propos peut parfois s'avérer plus facile à déterminer que l'à propos, et d'autre part dans la nécessité de raisonner directement sur la notion de non-à-propos dans certains contextes particuliers comme le filtrage d'informations.

Les auteurs proposent diverses règles concluant au non-à-propos. Par exemple, la règle suivante transpose la relation de préclusion (P) au niveau de l'à propos : si A et B sont sémantiquement incompatibles, on considérera que A n'est pas en relation d'à propos avec B.

$$\frac{A \perp B}{A \not\models B} \quad \text{(P)}$$

En tant que prémisse, la relation de non-à-propos apparaît dans diverses règles exprimant ses rapports avec la relation d'à propos. Différents jeux de règles sont envisageables selon que l'on adopte une attitude dite « optimiste » ou « pessimiste ». Dans le premier cas, on favorisera l'existence d'une relation d'à propos :

$$\frac{A \not\models B \quad C \models B}{A \oplus C \models B} \quad \text{(OL)} \qquad \frac{A \not\models B \quad A \models C}{A \models B \oplus C} \quad \text{(OR)}$$

Inversement, une attitude pessimiste privilégiera l'existence d'une relation de non-à-propos :

[6] Nous avons d'ailleurs utilisé ce même principe pour exploiter le modèle componentiel dans le cadre de la segmentation thématique, voir annexe B.

[7] Précisons que selon les auteurs, des sens légèrement différents seront attribués à cette relation qui sera, selon les cas, équivalente ou distincte de la négation de la relation d'à propos. Voir par exemple le concept d'*anti-aboutness* chez (Huibers, 1996).

$$\frac{A \not\models B}{A \oplus C \not\models B}\ (\text{PL}) \qquad \frac{A \not\models C \quad B \models C}{A \not\models B}\ (\text{PM}) \qquad \frac{A \not\models B}{A \not\models B \oplus C}\ (\text{PR})$$

Il est clair qu'un même système ne pourra admettre simultanément des postulats optimistes et pessimistes, qui sont mutuellement exclusifs. Ce point particulier illustre de façon concrète l'incidence du choix des propriétés d'une relation d'à propos sur les fonctionnalités à attendre d'un système de RI fondé sur elle. Par exemple, le choix de postulats optimistes ou pessimistes déterminera la réponse qui sera donnée à la question suivante : en considérant que « danser » est à propos de « s'amuser » et n'est pas à propos de « rester immobile », doit-on considérer que « danser » est à propos de « s'amuser $\oplus$ rester immobile » ?

2.2.3 Apports de l'approche logique

Nous en savons maintenant assez pour mesurer l'apport des approches discutées ici. En premier lieu, elles fournissent une méthode générale permettant d'évaluer l'à propos d'un VI relativement à un autre à partir de propriétés énoncées formellement. Un rapprochement immédiat peut être fait avec le modèle de RI booléen (cf. chapitre 1), où le test de pertinence d'un document d relativement à une requête q reviendrait à chercher une preuve de q en partant de d, à partir d'un jeu R de règles d'inférence. En cas de succès, on dira que q est inférable (ou dérivable) à partir de d, et on écrira $d \vdash_R q$. Mais la portée de l'approche logique est beaucoup plus large, puisqu'on pourra l'exploiter plus généralement pour chercher à démontrer (ou infirmer) de façon formelle la validité d'une relation d'à propos entre deux vecteurs d'information, étant donné un jeu donné de règles et d'axiomes. Il existe des algorithmes capables de fournir automatiquement ce type de preuve (Lee et Chang, 1973), qui permettraient d'inclure de type de raisonnement au sein d'un système de recherche d'information fonctionnel.

Mais aussi et surtout, cette approche fournit un cadre formel permettant de discuter des propriétés de telle ou telle relation d'à propos tout en bénéficiant d'une certaine rigueur apportée par l'outillage logique. Cela permet permet par exemple d'étudier les effets de l'introduction d'un axiome ou d'une règle de dérivation au sein d'un système, indépendamment du modèle sous-jacent. Considérons par exemple les axiomes suivants, où t, p, b, f et j désignent des VI minimaux[8] :

- « Tweety (t) is a pinguin (p) » : $t \mapsto p$ et donc $t \models p$;
- « Tweety is a bird (b) » : $t \mapsto b$ et donc $t \models b$;
- « Jack (j) is a bird » : $j \mapsto b$ et donc $j \models b$;
- « Jack flies (f) » : $j \models f$;
- « Pinguins do not fly » : $p \perp f$.

Dans ces conditions, et en se basant sur les postulats évoqués jusqu'ici, nous pouvons produire la dérivation suivante, concluant que « Tweety » n'est pas à propos de « Jack » :

$$\dfrac{\dfrac{\dfrac{\dfrac{\mathrm{t} \rightarrow \mathrm{p} \quad \mathrm{p} \perp \mathrm{f}}{\mathrm{t} \perp \mathrm{f}}\,(CP) \quad \mathrm{t} \models \mathrm{b}}{\mathrm{t} \not\models \mathrm{f}}\,(P)}{\mathrm{t} \not\models \mathrm{f} \oplus \mathrm{b}}\,(PR) \quad \dfrac{\mathrm{j} \models \mathrm{f} \quad \mathrm{j} \models \mathrm{b}}{\mathrm{j} \models \mathrm{f} \oplus \mathrm{b}}\,(And)}{\mathrm{t} \not\models \mathrm{j}}\,(PM)$$

Remplaçons maintenant les postulats « pessimistes » évoqués plus haut par le suivant, qui est une nouvelle variante des différents postulats de monotonie déjà envisagés :

[8]Cet exemple reprend une situation désuète mais traditionnellement utilisée dans les écrits de logique. Nous aurons toutefois l'occasion de voir plus loin un exemple concret d'application de la notion de preuve logique à la recherche d'information.

$$\frac{A \models B \quad \neg (A \perp C)}{A \oplus C \models B} \text{ (QRM)}$$

Admettons également, en plus des axiomes précédents, que Tweety et Jack n'ont pas de raison de s'exclure mutuellement, par exemple parce qu'ils appartiennent à la même personne. Nous avons alors l'axiome $\neg (A \perp C)$, ce qui permet d'appliquer la règle QRM donnée ci-dessus et de produire les inférences suivantes, qui concluent cette fois à une relation d'à propos entre « Tweety » et « Jack » (l'opérateur de préclusion barré désigne la négation de celui-ci) :

$$\frac{\dfrac{\mathrm{t} \mid\!= \mathrm{b} \quad \mathrm{t} \not\perp \mathrm{j}}{\mathrm{t} \mid\!= \mathrm{b} \oplus \mathrm{j}}\,(QRM)}{\mathrm{t} \mid\!= \mathrm{j}}\,(Absorption)$$

Nous observons ainsi que selon que l'on choisisse des postulats pessimistes (ici PR et PM) ou la monotonie QRM, nous obtenons des raisonnements contradictoires à partir des mêmes axiomes. Cela nous renseigne sur l'incompatibilité de ces deux jeux de postulats, et permet de préférer l'un ou l'autre en connaissance de cause, selon les besoins particuliers de chaque système de recherche d'information.

Dans une démarche inverse, la modélisation logique permet de comparer les propriétés des relations d'à propos qui sous-tendent différents modèles de RI existants. Considérons par exemple les relations suivantes :

- Soit $\models_B$ la relation d'à propos en vigueur dans le modèle booléen. On admet dans ce cas l'équivalence suivante : $D \models_B Q \Leftrightarrow D \vdash Q$ (D et Q sont alors des formules de logique propositionnelle).
- Soit $\models_{UV}$ la relation d'à propos en vigueur dans un modèle « saltonnien » non seuillé, fondé sur une quantification vectorielle de l'à propos. On admet alors l'équivalence suivante : $D \models_{UV} Q \Leftrightarrow cos(D, Q) > 0$ (dans ce cas D et Q sont des vecteurs au sens saltionien).
- Soit $\models_{TV}$ la relation d'à propos en vigueur dans un modèle « saltonnien » avec seuil. On admet alors l'équivalence suivante : $D \models_{TV} Q \Leftrightarrow cos(D, Q) > \delta$ (où δ est un seuil donné par une valeur arbitraire strictement supérieure à zéro).

Les outils logiques dont nous disposons nous permettent de caractériser leurs différences entre ces trois relations. Ainsi, il est démontré dans (Bruza *et al.*, 2000) que la relation $\models_B$ supporte les propriétés R et LM (réflexivité et monotonie gauche), alors que $\models_{UV}$ supporte R, LM et RM (idem plus monotonie droite), et que $\models_{TV}$ ne supporte que la réflexivité (R). Comme le soulignent les auteurs, ces différences peuvent se traduire par des différences sensibles quant aux services rendus aux utilisateurs. Le degré de support des propriétés de monotonie, en particulier, aura une incidence notable sur la précision des résultats.

Considérons par exemple le syntagme « surfer à Hawaï », qui pourrait se traduire par le vecteur d'information $surfer \oplus Hawaii$. Les propriétés élémentaires de la relation d'à propos nous permettent d'en déduire que ce syntagme est (entre autres) à propos de « surfer » : $surfer \oplus Hawaii \models surfer$. Or, si l'on admet la monotonie droite (RM), nous pouvons inférer : $surfer \oplus Hawaii \models surfer \oplus Australie$, ce dernier fait signifiant que « surfer à Hawaii » est à propos de « surfer en Australie », avec les conséquences que l'on imagine en termes de recherche d'informations, c'est-à-dire une précision moindre. Inversement, un système vectoriel seuillé, qui ne supporte aucune propriété de monotonie, fournira un meilleur taux de précision. En effet, même si les deux vecteurs partagent un même sous-élément (ici « surfer »), la divergence entre les autres éléments (« Hawaii » et « Australie ») peut suffire à ramener la distance angulaire sous le seuil fixé. Précisons toutefois que cela ne constitue pas une preuve de la supériorité du modèle vectoriel, qui nécessite d'une part de déterminer le seuil expérimentalement, et dont le gain en précision peut se traduire par ailleurs par une diminution du taux rappel.

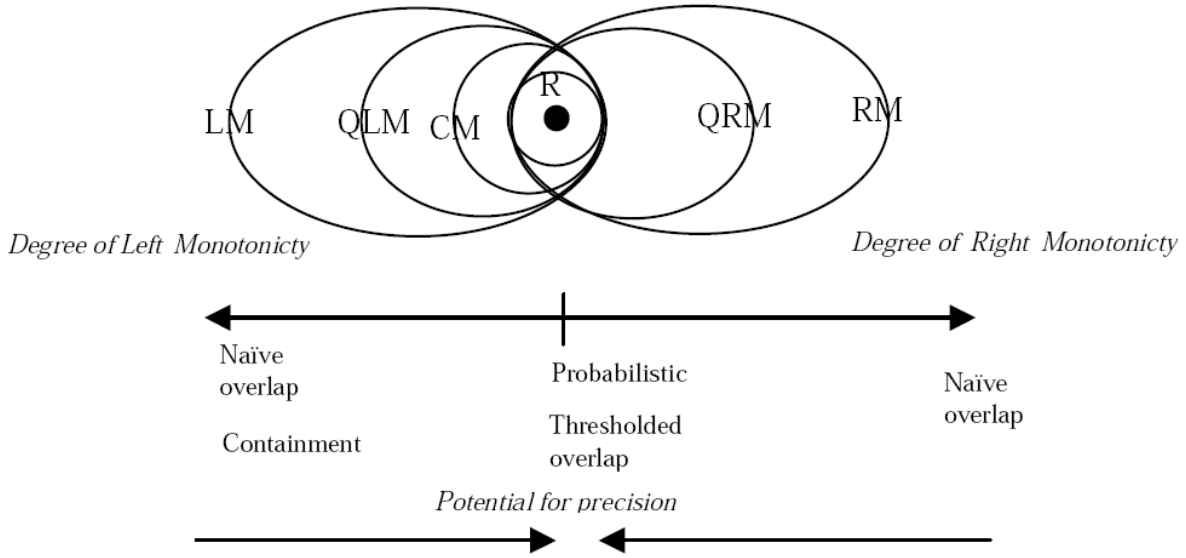


FIG. 2.5 – Spectre des relations de monotonie (Bruza *et al.*, 2000).

Finalement, selon les différents degrés de monotonie envisagés dans (Bruza *et al.*, 2000) (dont la plupart ont été abordés ici), les auteurs établissent une typologie des différents modèles de RI. Cette typologie, qui traduit le fait que la précision du modèle est inversement proportionnelle à la « permissivité » des relations de monotonie qu'il supporte, est reproduite dans la figure 2.5.

Un autre aspect intéressant, bien que plus anecdotique, du modèle logique présenté dans cette partie réside dans ses liens éventuels avec des modèles du discours tels qu'envisagés dans la section 2.1. En particulier, les auteurs revendiquent un lien avec les notions de micro- et macro-structures telles qu'envisagées dans (Hutchins, 1977) : les éléments de la macro-structure seraient ordonnés selon la relation d'inclusion de surface ($\supseteq$), alors que les éléments de la micro-structure seraient liés par la relation d'inclusion forte ($\mapsto$). Partant de ce principe, ils avancent que l'à propos extensionnel (au sens de (Fairthorne, 1969) repris dans (Hutchins, 1977)) pourrait être modélisé par « la clôture[9] d'un système de preuve construit à partir de la micro-structure », et l'à propos intensionnel serait modélisé par « la clôture d'un système de preuve construit à partir de la macro-structure du texte ». Dans ce cas précis toutefois, les auteurs soulèvent une vaste problématique sans donner plus de détails, ce point demeurant accessoire dans les travaux de ce courant. En tout état de cause, il nous est permis de douter de la portée du parallèle ici établi, et notamment de la possibilité de rendre compte de la complexité et de la variété des relations sémantiques en jeu dans le discours avec un jeu d'opérateurs aussi restreint, et sans faire plus d'hypothèses sur la modélisation du sens.

2.2.4 Conclusion

Bien que très abstraits et éloignés des réalités linguistique du discours, les travaux que nous venons de présenter nous semblent apporter une contribution intéressante à la problématique de l'à propos. On pourra certes s'étonner d'un décalage certain entre la rigidité des modèles logiques et le caractère « mou » et fuyant des concepts liés à la notion d'à propos. Il est vrai qu'en toute généralité, la notion d'à propos est par nature difficile à saisir, toujours sujette à interprétation, et que l'on peut considérer que c'est là simplifier à outrance que de la représenter par un simple opérateur binaire. On pourrait de même

[9]Telle que définie dans (Bruza *et al.*, 2000) p. 4.

observer que la validité des règles d'inférence qui lui sont attribuées est difficile à évaluer en toute généralité, sans faire plus d'hypothèses sur les propriétés sémantiques des vecteurs d'information. De fait, pour chacune des règles de dérivation évoquées plus haut, il sera probablement toujours possible de trouver un contre-exemple pour l'infirmer.

Il faut toutefois garder à l'esprit que le formalisme logique présenté ici ne prétend pas à l'universalité. Il ne s'agit aucunement de définir *une* théorie de l'à propos, mais seulement de fournir un cadre formel permettant de formuler *des* théories de l'à propos, dont on peut étudier objectivement les propriétés. Le jeu d'opérateurs pourra si nécessaire être étendu pour prendre en compte les différents phénomènes sémantiques que l'on estime pertinents dans un cadre donné. On pourrait par exemple affiner la relation d'inclusion forte en introduisant de nouveaux opérateurs pour prendre en compte différentes relations lexicales « classiques » que sont l'hypéronymie, la méronymie, etc. (voir par exemple (Nie, 2001)). En résumé, le formalisme logique nous permet de modéliser une relation d'à propos, et d'étudier les conséquences de tel ou tel choix de modélisation.

L'intérêt pratique de cette démarche est évident : tout système automatique de recherche d'information doit procéder à ce type de choix, en décidant *a priori* d'un ensemble de règles qui permettront de statuer sur la pertinence d'un document relativement à une requête. Pourtant, alors qu'elles existent dans tout système d'information, ces règles ne sont pas toujours rendues explicites en tant que telles. Elles sont plutôt la conséquence du modèle général du système, et ne sont donc pas pleinement maîtrisées. C'est là l'intérêt principal que nous reconnaissons à l'approche présentée ici dans le cadre de la RI : à l'aide d'un formalisme centré sur la notion d'à propos, on peut expliciter formellement les propriétés de différents modèles de recherche d'information, les comparer objectivement et caractériser finement chaque modèle. Inversement, un système de RI fondé la notion de système de preuve logique permettrait de rendre totalement explicites des propriétés de l'à propos sur lesquelles il se fonde. On pourrait alors, selon chaque besoin particulier, « paramétrer » le système de façon déclarative en ajustant les règles d'inférences utilisées par le moteur d'inférence.

Finalement, abordons un dernier point qui constitue, de notre point de vue, à la fois une force et une faiblesse de cette approche. Il est intéressant de constater qu'en apparence, très peu d'hypothèses sont faites sur la nature même d'un vecteur d'information. De fait, la notion d'à propos est uniquement envisagée comme une relation : la question de l'à propos d'un texte n'est pas posée dans l'absolu, mais seulement *relativement à* un autre vecteur d'information (qui sera généralement une requête). Cette approche purement relationnelle est séduisante, puisqu'elle conduit à formuler le moins d'hypothèses possibles sur la nature de ces vecteurs d'information et de la façon dont on peut les obtenir à partir d'un objet textuel.

Mais la réalité des systèmes d'information nous ramène nécessairement à la notion d'index, et à la nécessité de produire, à un moment ou un autre, une certaine représentation d'un document qui sera effectivement confrontée à une représentation équivalente de la requête. On ne peut évidement pas reprocher à l'approche logique de ne faire aucune hypothèse sur les procédés d'analyse des documents ou des requêtes qui permettent d'obtenir un vecteur d'information, car là n'est pas son propos. Mais il faut cependant observer que le formalisme choisi reste fortement influencé par la démarche classique d'indexation par mots-clefs.

Car si les auteurs ne considèrent le mot-clef que comme un cas particulier de vecteur d'information minimal, il nous paraît clair que les opérateurs choisis sont préférentiellement adaptés à ce cas particulier. L'opérateur de composition ($\oplus$), notamment, nous paraît clairement orienté dans le sens de structures « plates », composées de listes de mots-clefs. L'utilisation d'objets plus complexes, comme les structures sémantiques que nous utiliserons dans la partie II, rendrait en effet nécessaire une modélisation plus fine de la composition. Il en va de même, comme nous l'évoquions ci-dessus, de la relation d'inclusion forte qui est fortement dépendante de la modélisation sémantique sous-jacente aux vecteurs d'information minimaux. Et dans la mesure où les différents modèles proposés s'appuient sur

des règles d'inférence elles-mêmes fondées sur les propriétés de ces opérateurs primitifs, il est évident qu'elles devraient être reconsidérées dans le cas où ces opérateurs seraient modifiés. Ce point ne remet aucunement en cause les avantages du cadre théorique qui semble applicable de façon assez générale, mais laisse à penser que son application dans des contextes plus distants du paradigme des mots-clefs nécessiterait probablement de développer d'une théorie de l'à propos (c'est-à-dire un jeu d'opérateurs et de règles de dérivation) sensiblement différente de celles évoquées ici.

2.3 Approche bibliothéconomique de l'à propos

Les sciences de l'information définissent un champ d'investigation extrêmement large, alimenté par une littérature abondante, et couvrent dans leur conception moderne la plupart des concepts évoqués dans cette partie. Cette section a trait à une conception plus « classique » de ce domaine, plus particulièrement liée aux problématiques associées à la bibliothéconomie, c'est-à-dire aux pratiques des documentalistes en tant que juges de l'à propos d'un document face à un ensemble d'utilisateurs. Il s'agit là encore d'un aperçu très parcellaire d'une discipline que nous ne pouvons bien-évidemment pas couvrir ici, mais qui nous permet d'évoquer un certain nombre de principes qui peuvent se révéler éclairants dans le contexte de la RI conçue processus informatique.

2.3.1 Subjectivité de l'à propos et approche probabiliste

Maron (1960, 1977) envisage la notion d'à propos relativement à la notion de processus interprétatif. Plus précisément, ses travaux s'appuient sur l'idée que la notion d'à propos ne devrait pas être conçue de façon absolue, mais relativement aux propriétés des différents processus interprétatifs pouvant conduire à son élaboration. Maron remarque en effet que la démarche qui conduit un individu à considérer qu'un texte est à propos de quelque-chose est intimement liée à son expérience personnelle, et semble difficile à transmettre : « nous sommes tous capables de penser, de comprendre et de savoir à propos de quoi est un texte, mais nous ne pouvons pas dire exactement ce qui nous y conduit, et nous ne pouvons certainement pas indiquer à quelqu'un d'autre comment il devrait procéder pour le faire » (p. 40). Une façon de prendre ce problème en considération est de considérer l'à propos comme un objet intrinsèquement relatif : « nous devrions peut-être considérer que l'à propos peut être interprétée de différents points de vue, dont chacun fait apparaître différentes interprétations du sens de l'à propos » (*ibid.*).

Nous avons vu à plusieurs reprises (notamment dans la section 2.2) que le choix des propriétés que l'on souhaite attribuer à la relation d'à propos est une question centrale en RI. En d'autres termes, cette question consiste à déterminer quel jugement de pertinence il est préférable d'appliquer, relativement à un problème donné. Face à ce problème, Maron définit trois types de relations d'à propos (p. 41) :

- L'à propos dit « subjectif », ou *S-about*, est « une relation entre un document et l'expérience individuelle de ses lecteurs ».
- L'à propos dit « objectif », ou *O-about*, correspond à une relation entre un document et un utilisateur « du point de vue d'un observateur tiers ». Il s'agit par exemple de l'ensemble des descripteurs qui seront, du point de vue de l'indexeur, utilisés par les utilisateurs pour obtenir un document.
- L'à propos de recherche d'information, ou *R-about* pour « retrieval about », « n'est pas relatif au comportement d'individus isolés, mais plutôt aux habitudes de recherche d'information d'une classe d'individus », c'est-à-dire les utilisateurs du système, telles que décrites par un observateur tiers.

C'est bien sûr cette dernière relation qui présente, selon l'auteur, le plus grand intérêt pour la recherche d'information. Nous pouvons la reformuler comme suit : si D est un document et T un des-

cripteur, l'assertion « T est R-about D » signifie qu'une proportion suffisante d'utilisateurs formulant une requête en utilisant le descripteur T est satisfaite par le document D. Il s'agit selon l'auteur d'un compromis entre le S-about, centré sur l'utilisateur, et le O-about, centré sur le système. Le choix d'une relation du type R-about permet de tenir compte des besoins d'une classe d'utilisateurs sans pour autant chercher à satisfaire chaque souhait individuel. Il s'agit donc de raisonner en termes de *probabilité de satisfaction*.

C'est en partant de ces considérations que Maron a élaboré un modèle de recherche d'information qui est à l'origine du courant probabiliste déjà évoqué dans la section 1.2. Rappelons que cette approche consiste à évaluer la « probabilité de pertinence » de chaque document relativement à une requête. Le système peut alors présenter à l'utilisateur ces documents selon cette probabilité, sans jamais avoir à prendre une décision ferme sur la pertinence d'un document. Pour cela, la forme « classique » de l'approche probabiliste consiste à exploiter une certaine forme d'évaluation (ou *feedback*) de la part des utilisateurs de la pertinence des documents retournés relativement à leur requête. En termes naïfs, on peut dire que le système « apprend » à reconnaître un document pertinent relativement à une requête, s'adaptant ainsi à un utilisateur ou une classe d'utilisateurs.

Les approches probabilistes modernes tendent plutôt à s'affranchir des appréciations subjectives des utilisateurs pour procéder à l'évaluation de la probabilité de pertinence des documents, et s'éloignent peut-être ainsi de l'objectif initialement fixé. Nous pouvons toutefois retenir de l'approche de Maron la nécessité pour tout modèle de RI de se positionner entre deux pôles correspondant à un centrage sur le système ou sur l'utilisateur. Il est clair que ces considérations ne sont en rien spécifiques à l'approche probabiliste, qui constitue un moyen parmi d'autres de prendre en compte les besoins spécifiques des utilisateurs, et qui n'est peut-être pas la plus efficace. La méthode du *feedback* a en effet la particularité d'intervenir très tardivement dans le processus de recherche d'information, puisqu'elle permet d'affiner le jugement de pertinence sans remettre en cause l'indexation. Mais on peut aussi envisager d'autres modes d'adaptation aux utilisateurs qui interviendraient en amont de l'indexation. Nous rejoindrions ainsi encore une fois les prescriptions de Hutchins (cf. section 2.1), qui suggère de prendre en considération le réseau sémantique propre à une classe d'utilisateurs dans la définition même de l'à propos d'un document.

2.3.2 L'approche matérialiste de Hjørland

Nous allons dans cette section faire état de certaines vues sur la notion de « sujet » que l'on doit à Birger Hjørland, théoricien important des sciences de l'information. Précisons tout d'abord que cette notion de sujet (*subject* dans le texte) est pour cet auteur exactement équivalente à la notion d'à propos (*aboutness*). Hjørland considère en effet que l'introduction par Hutchins de cette dernière notion pour remplacer celle de sujet, jugé trop floue, relève essentiellement d'un changement de terminologie et non d'une réelle clarification conceptuelle (Hjørland, 2001).

Pour introduire la difficulté inhérente à la définition du sujet d'un ouvrage, Hjørland cite à plusieurs reprises (Hjørland, 1992; Hjørland, 2001) une liste établie par Wilson (1968), explicitant différentes méthodes susceptibles d'être employées y parvenir :

1. identifier les objectifs visés par l'auteur en écrivant le document ;
2. comparer les relations de dominance et de subordination entre les différents éléments du schéma général perçu à la lecture du document ;
3. regrouper et dénombrer les usages de concepts et de références du document ;
4. élaborer un ensemble de règles permettant de distinguer les éléments essentiels du document (par opposition aux éléments superflus).

Wilson a montré par l'expérience qu'aucune de ces méthodes ne permet d'obtenir une représentation satisfaisante du sujet d'un document, et conclut à l'indétermination de la notion même de sujet. S'il partage l'avis de Wilson sur l'insuffisance des méthodes énumérées ci-dessus, Hjørland récuse toutefois l'idée selon laquelle les sujets sont indéfinissables. Il avance que la détermination du sujet d'un document est possible à condition d'adopter une attitude « matérialiste », ce point de vue étant argumenté dans (Hjørland, 1992), qui débute par la critique d'un certain nombre d'autres postures.

La première d'entre elles est qualifiée d'*idéaliste-subjective*. Dans l'acception philosophique du terme, l'attitude idéaliste est selon Hjørland principalement caractérisée par le fait qu'elle considère les processus mentaux (les idées) comme prédominants relativement à la réalité du monde, par opposition aux attitudes dites réalistes ou matérialistes. Par suite, l'approche idéaliste-subjective conduit selon l'auteur à considérer les concepts et les sujets comme l'expression des perceptions ou des opinions d'une ou plusieurs personnes. Dans cette perspective, « la clef du concept de sujet réside dans l'étude des processus mentaux de certaines personnes, par exemple les auteurs ou les utilisateurs des documents » (p. 173). Hjørland rejette cette option en considérant que le point de vue de l'auteur, du lecteur, ou même du documentaliste ne peuvent constituer une vue objective du sujet d'un document.

La seconde posture critiquée par Hjørland est dite *idéaliste-objective*. Cette approche considère comme objective une définition du sujet qui serait partagée par au moins deux individus.[10] Plus généralement, l'idéalisme objectif considère les concepts comme des entités abstraites ayant une existence propre, c'est-à-dire universels et indépendants de la conscience humaine. Dans le cadre qui nous intéresse ici, cette approche consiste donc à considérer que les sujets existent de façon autonome, et que les documents concrets ne font que « partager les 'idées' exprimées dans un certain sujet » (p. 177). À titre d'exemple, Hjørland mentionne la classification à facettes de Raganathan (dont nous reparlerons plus loin), qui considérait qu'il était possible de définir une « syntaxe absolue » des sujets, indépendante de tout point de vue, document ou domaine. Cette approche est là encore rejetée, l'auteur considérant qu'il est indispensable de considérer la notion de sujet comme dépendante à la fois des propriétés de chaque document et du contexte de son utilisation (p. 179).

Ces derniers points font partie de la posture qualifiée de *pragmatique*, qui reprend la distinction posée par (Soergel, 1985) entre une indexation orientée vers le contenu ou vers l'utilisateur : l'indexation orientée *contenu* est conçue comme dépendant exclusivement des propriétés que l'on peut attribuer à un document, alors que l'indexation orientée *requête* (ou *utilisateur*, ou encore *besoins*) établit une relation dite *instrumentale* entre les propriétés d'un document et les besoins (réels ou supposés) d'un certain groupe d'utilisateurs. Hjørland illustre ainsi cette distinction (p. 180) : alors qu'une approche orientée contenu se limitera à observer qu'un document contient la formule chimique de l'acide sulfurique, une approche orientée requête pourra par exemple procéder au raisonnement suivant : l'acide sulfurique est corrosif, et les graveurs utilisent des agents corrosifs, donc une catégorisation pertinente du document pourrait être « agents chimiques utiles pour la gravure ». S'il est évident que les approches de ce type ne peuvent s'appliquer qu'à des systèmes d'information spécialisés, elles présentent l'intérêt majeur de rejeter la notion de sujet comme inhérente aux documents, en la considérant au contraire comme liée également à une pratique.

Cette posture pragmatique fonde l'approche finalement proposée par Hjørland. Il adopte alors un point de vue *matérialiste*, considérant que « les choses existent objectivement et englobent des propriétés objectives » (p. 181). L'auteur s'intéresse donc à ces *propriétés objectives* des documents, évidemment distinctes du point de vue subjectif de leurs auteurs[11], parmi lesquelles il retient (p. 182) :

– L'aspect de la réalité qu'il reflète (son à propos, au sens restreint du terme).

[10] La notion d'intersubjectivité paraît très proche des concepts ici évoqués, mais n'est pas mentionnée par l'auteur.

[11] Pour éclairer ce point, Hjørland donne l'exemple suivant : si un auteur avance que « l'intelligence d'une personne est proportionnelle à la taille de son cerveau », il s'agit d'un jugement subjectif (et faux). Mais le fait que le document exprime cette opinion n'en est pas moins un fait objectif.

– La façon dont cette réalité est relatée : si les assertions sont-elles vraies ou fausses, superficielles ou essentielles, etc.
– Propriétés émergeant d'une utilisation particulière du document, comme ses qualités pédagogiques ou scientifiques.
– La fréquence et la distribution des mots, le vocabulaire utilisé.
– Etc.

L'identification de ces propriétés constitue selon Hjørland le premier niveau d'analyse d'un document, qui permet d'obtenir ce qu'il nomme des prédicats de premier degré. Le second niveau de l'analyse vise la définition du sujet en tant que tel, qui est alors conçu comme une fonction appliquée à ces propriétés élémentaires. En d'autres termes, il s'agit de déterminer quelles sont celles qui sont susceptibles de participer à la définition du sujet d'un document, et de quelle façon elles y participent, ce qui revient à établir des méta-prédicats, ou prédicats de second degré[12].

Partant du principe que les prédicats du premier degré sont relativement faciles à établir (du moins pour un observateur humain), Hjørland avance que la difficulté de la définition de la notion de sujet réside essentiellement dans la détermination de méta-prédicats que l'on doit leur appliquer pour l'obtenir. Conformément à la posture pragmatique évoquée plus haut, il considère que ce choix est dépendant du contexte et doit viser « l'optimisation de la perception potentielle du document » (p. 185). Nous arrivons ainsi à l'argument principal de l'auteur, qui est que cet objectif doit être poursuivi d'un point de vue *épistémologique* : « les sujets en eux-mêmes doivent être définis comme les potentiels épistémologiques des documents » (*ibid.*).

L'auteur remarque en effet que les documents alimentent les processus cognitifs humains au même titre que les individus eux-mêmes, les objets, les faits, les processus, etc. Incidemment, rendre un document « visible » pour les utilisateurs d'un système d'information revient à l'intégrer de la meilleure façon possible dans leur démarche d'acquisition de connaissances, et donc de considérer avant tout « leur rôle potentiel dans les développements futurs de la connaissance » (p. 187). Plus concrètement, cette démarche consiste à évaluer un document selon son positionnement relativement à l'état actuel des connaissances, la nature des hypothèses qu'il pose et de ses conclusions, le jugement d'une communauté sur ses apports, les liens qu'il établit avec d'autres disciplines, etc. Finalement, Hjørland avance qu'une description « pure » d'un document, c'est-à-dire sans connexion avec la globalité du processus d'évolution des connaissances au sein duquel il s'inscrit, ne permet d'accéder qu'à des « propriétés superficielles », et qu'une approche épistémologique de l'indexation est donc nécessaire pour « faciliter le développement des connaissances en direction de la substance des choses ».

Une première remarque que l'on peut formuler à propos de cette proposition de Hjørland concerne l'absence de toute référence à la typologie de Maron que nous avons évoquée plus haut (Maron, 1977). On pourra pourtant observer une proximité manifeste entre la posture idéaliste-subjective et la notion S-about, la posture idéaliste-objective et la notion de O-about, et enfin entre la posture pragmatique et la notion de R-about. On notera en revanche que l'approche matérialiste/épistémologique finalement revendiquée par Hjørland ne trouve pas d'équivalent direct dans la typologie de Maron, à moins de la considérer comme une approche particulière du R-about. Si l'on prend en compte le fait que les travaux de Maron sont plus proches de la problématique proprement informatique que ceux de Hjørland, cela s'explique par le fait que l'emploi de critères d'indexation fondés sur un raisonnement épistémologique (comme d'ailleurs sur tout autre raisonnement du même ordre de « complexité ») est totalement hors de la portée de toute l'analyse automatique.

Nous nous arrêterons de fait, si l'on considère les apports de cette contribution de Hjørland aux problèmes de RI nous occupent ici, au niveau pragmatique. On remarquera en particulier que les arguments développées par l'auteur en sa faveur corroborent l'idée, déjà rencontrée chez Hutchins, selon

[12]Un exemple de méta-prédicat, donné par l'auteur, est la réflexivité, qui s'applique à un prédicat d'ordre inférieur pour établir le fait qu'il dispose de la dite propriété.

laquelle la prise en compte de l'organisation des connaissances spécifiques à une classe d'utilisateurs peut améliorer sensiblement les services rendus par un système d'information. Or ce type de procédés n'est pas hors de la portée de l'indexation automatique. Revenons par exemple sur le cas évoqué plus haut au sujet de l'acide sulfurique : l'automatisation de ce type de raisonnement, par exemple sous forme logique (cf. section 2.2), est envisageable pour peu que les connaissances nécessaires soient fournies au système. Mais ce n'est là qu'une possibilité parmi tant d'autres, et nous verrons dans la section suivante avec une version assouplie de la classification par facettes, un autre moyen de prendre en compte les spécificités d'un domaine dans la tâche d'indexation.

2.3.3 La théorie des facettes

La théorie des facettes est liée à un modèle de classification documentaire introduit par Ranganathan, dit *Colon Classification*[13] (Ranganathan, 1963; Gopinath, 1976). L'apparition de ce modèle, en 1924, constituait un apport considérable dans le sens où elle constituait la première alternative au modèle hiérarchique qui prévalait jusqu'alors. Dans ce dernier modèle, la définition du sujet d'un document devait nécessairement se positionner au sein d'une hiérarchie pré-établie, rendant difficile l'expression d'un sujet qui serait le fruit d'un croisement de concepts qui ne serait pas prévu. Le modèle de la Colon Classification (dorénavant CC) introduit quant à lui une certaine notion d'orthogonalité, en proposant de représenter le sujet d'un document selon différentes « facettes » pré-établies.

Afin d'illustrer les différences entre ces deux approches, reprenons l'exemple donné dans (Maniez, 1999), qui décrit le sujet « la prévention des maladies virales du riz » selon les deux modèles. Dans le modèle hiérarchique de la *Dewey Decimal Classification* (DCC), ce sujet serait représenté par l'indice « 633.189.8 », obtenu à partir des dérivations suivantes :

- 633 : « céréales »
- 633.18 : « riz »
- 633.189 : « maladies du riz »
- 633.189.8 : « maladies du riz d'origine virale »

Dans ce modèle, il serait impossible d'ajouter la notion d'« éradication » à cet indice, qui n'est pas prévue comme une sous-ramification du dernier noeud obtenu. Dans le modèle CC, l'indice obtenu serait « EJ,381:421:5 », et représente ce qu'il est convenu d'appeler un « sujet composé », comprenant quatre indices élémentaires traduisant chacun une facette. On remarquera que dans ce cas il est aisé d'utiliser l'une des facettes pour ajouter la notion d'éradication :

- Facette principale : EJ (« agriculture »)
- Facette « Personnalité » : 381 (« riz »)
- Facette « Matière » : 421 (« maladie virale »)
- Facette « Énergie » : 5 (« éradication »)

Le modèle CC oppose donc aux modèles hiérarchiques, dont l'ensemble des sujets possibles est fermé, un modèle où la combinatoire produite par le croisement des facettes est considérable. Ranganathan considérait toutefois sa liste de facettes comme immuable et universelle : selon la dernière édition de la CC (1989), « tout élément d'un sujet appartient nécessairement à l'une des cinq - et seulement cinq - catégories fondamentales ». Cette liste est la suivante (hormis la facette dite « principale »), souvent désignée par l'acronyme « PEMST » :

- **P**ersonnalité : s'applique à l'objet principal du sujet.
- **É**nergie : s'applique à l'opération principale du sujet.
- **M**atière : propriété ou une substance du sujet.
- **S**pace (Espace) : localisation spatiale du sujet.

[13]Cette appellation est due au rôle particulier que joue la ponctuation dite « colon » en anglais (deux points) dans la notation associée.

sujet	*verbe principal*	*complément d'objet*	*complément de manière*
le chien	mange	un os	goulûment
elle	ferme	les volets	avec douceur
lire	dissipe	l'ennui	inconsciemment

facette principale	*personnalité*	*matière*	*énergie*
EJ = agriculture	382 = riz	421 = maladie virale	5 = éradication
L = médecine	45 = poumon	430 = tuberculose	6 = traitement
T = enseignement	522 = aveugles	234 = jeu de rôles	7 = éducation

FIG. 2.6 – Analogie entre les dualités facette/isolat et syntagmatique/paradigmatique (Maniez, 1999).

– **T**emps : localisation temporelle du sujet.

Ranganathan pensait avoir ainsi découvert la « syntaxe absolue » de l'énonciation des sujets. Notons que le terme de « syntaxe » paraît ici adéquat dans l'acception qui l'oppose au paradigme. En effet, comme le remarque (Maniez, 1999), la relation qui lie la formule PEMST à l'ensemble des concepts susceptibles de définir chaque facette (« isolats » dans le vocabulaire consacré) n'est pas sans rapport avec la relation entre axe syntagmatique et axe paradigmatique, comme l'illustrent les tableaux de la figure 2.6, à la différence que dans le cas du modèle CC la formule PEMST désignerait à la fois l'équivalent des catégories syntaxiques et des catégories conceptuelles.

Le caractère absolu de cette « syntaxe » de l'à propos est en revanche discutable : on peine à imaginer qu'un jeu si limité de facettes, même très générales, suffise à représenter efficacement le sujet de tout document, quel que soit son domaine. Cette hypothèse sera par exemple critiquée dans (Maniez, 1999), ainsi que par Hjørland qui la considère comme appartenant à la démarche « idéaliste-objective » (Hjørland, 1992, p. 178) qu'il récuse (cf. supra).

Le modèle de la CC n'en a pas moins constitué un apport conséquent en son temps, et fait depuis l'objet de nombreuses études, critiques et améliorations, aboutissant par exemple à la notion de « thesaurus à facettes ». La plupart de ces évolutions ont levé la contrainte d'universalité des facettes, qui constitue clairement la faiblesse principale du modèle original. Il en va ainsi des travaux du *Classification Research Group* (Vickery, 1963), qui préconise la définition de « schémas spéciaux » adaptés à des domaines de spécialité. Dans cette optique, l'ensemble des facettes n'est plus fixé *a priori* mais au contraire le résultat d'un « examen détaillé de la documentation relevant du domaine à traiter ». Vickery propose même une méthode permettant d'établir les facettes propres à un domaine, qui consiste à extraire une quantité raisonnable de termes de la littérature de ce domaine, et de suivre à partir d'eux les relations d'hyperonymie (par exemple à l'aide d'un dictionnaire) qui permettront d'arriver à des termes suffisamment généraux (mais pas plus) pour couvrir l'ensemble initial. A titre d'exemple, l'auteur cite différentes facettes valables dans le domaine de la pédologie[14] :

– types de sol (ex : « limon rouge ») ;
– actions sur les sols (ex : « érosion par le vent ») ;
– processus (ex : minéralisation) ;
– etc.

De fait, le processus visant à établir les facettes propres à un domaine constitue en soi une tâche intéressante en termes d'ingénierie des connaissances, qui gagnera sans doute, comme toutes les tâches de cet ordre, à prendre racine dans l'étude (automatique ou non) de corpus. D'autre part, les facettes obtenues nous semblent constituer un guide effectivement très intéressant pour la description du contenu informationnel d'un document, mais aussi, ce qui nous intéressera tout particulièrement, de segments

[14] Science étudiant les sols.

intra-documentaires. On imagine en effet très bien comment, étant données un ensemble de facettes, certaines d'entre elles pourraient être utilisées pour décrire l'ensemble du document, quand d'autres prendraient plusieurs valeurs au fil du discours. Pour reprendre l'exemple précédent, on pourra par exemple trouver un document portant dans son ensemble sur un seul type de sol, dont le traitement sera décliné selon différents types de modification des sols (érosion, décomposition, etc.). Plus généralement, il est clair que des unités documentaires de grains divers pourront être décrits de la sorte, pour finalement former une structure hiérarchique de type thème / sous-thème. Ce sont des structures analogues que nous chercherons plus loin à analyser, mais en nous concentrant sur un niveau discursif plus fin.

Chapitre 3

Notions de thème dans la théorie linguistique

Nous allons maintenant nous intéresser aux notions de thème telles qu'elles sont discutées par des théories plus proprement linguistiques. Une fois encore, nous ne prétendrons bien sûr aucunement à l'exhaustivité, tout en espérant dresser un panorama suffisamment représentatif de l'état de l'art en la matière. On remarquera immédiatement que dans ce domaine, le terme même recouvre des concepts parfois très différents, et nous essaierons de nous concentrer sur ceux qui font intervenir la notion d'à propos, et qui sont ainsi susceptibles de faire écho aux propositions du chapitre précédent.

Nous tenterons de les aborder par ordre de grain linguistique, de la proposition au discours. Il s'agira tout d'abord de théories liées à la *structure informationnelle*, qui consiste essentiellement à étudier le rôle thématique des constituants de la proposition ou de la phrase. Nous envisageons par la suite différentes théories portant sur des grains plus importants, notamment avec la notion de *progression thématique* au niveau inter-phrastique, ou celle de *thème discursif*. Enfin, nous considérons des travaux qui ne sont pas explicitement liés à la notion de thème, mais qui sont d'une grande importance dans l'analyse du discours, dont la *théorie du centrage*, celle de la *structure rhétorique*, et surtout l'hypothèse de l'*encadrement du discours*.

3.1 Approches centrées sur la phrase

Une grande variété de théories ont cherché à rendre compte, de façon plus ou moins liée à sa syntaxe, de la façon dont la phrase participe à la construction de l'information dans un texte. La dualité thème/rhème, telle que théorisée dans les années 1920 par l'Ecole de Prague (en particulier par Mathesius), distingue ainsi le *thème*, ce dont on parle, l'élément connu, du *rhème* (ou propos ou commentaire), qui constitue l'apport d'information, ce que l'on dit du thème. Ainsi, dans la phrase « Jean est allé au cinéma », le thème est « Jean » et le rhème est constitué du reste de la phrase : c'est au sujet de Jean, personnage *a priori* connu, que la phrase apporte une information nouvelle, cette information étant qu'il est allé au cinéma.

La linguistique bien évidemment fait évoluer considérablement ce modèle depuis son introduction, à travers de multiples travaux auxquels on fait généralement référence par les termes de *structure informationnelle* ou *communicative*. Nous allons tenter dans cette section d'en faire un bref tour d'horizon, en axant notre étude sur deux dualités qui nous seront particulièrement utiles par la suite :

– Le première concerne les modèles du type « thème/rhème » proprement dits. Cette dénomination comptant parmi les plus anciennes et les plus répandues, nous avons choisi de l'utiliser pour intituler cette première partie, même si les différentes théories que nous évoquerons utilisent

souvent des termes différents.
– La seconde, semble-t-il beaucoup moins étudiée, s'insère au sein même de la notion de thème : il s'agit de la dualité « sujet/topique ». L'intitulé de cette seconde partie reprend les termes de Chafe (1976), mais nous trouverons là aussi différentes terminologies selon les auteurs.

3.1.1 La dualité thème/rhème

Selon les théories, différents critères ont été donnés pour caractériser cette dualité. Ceux-ci peuvent être catégorisés ainsi :

1. Critères portant sur des propriétés cognitives des référents, et plus particulièrement sur leur *identifiabilité* et leur *accessibilité* (notamment chez Chafe).
2. Critères portant sur la notion d'à propos, le thème étant considéré comme « ce sur quoi porte la proposition » (notamment chez Lambrecht).
3. Critères portant sur sa *position* dans la phrase, critère donnant un statut préférentiellement thématique à la position initiale, au moins pour certaines langues (par exemple chez Halliday).
4. Critères portant sur la notion de *saillance*, du point de vue cognitif, prosodique, typographique, etc. Notons que la position initiale peut également être vue comme un critère de saillance.

Selon ces différentes approches, et selon les auteurs, des termes différents ont été employés pour désigner les pôles thématiques et rhématiques. On trouvera notamment :

Pôle "thématique" : thème, topique, sujet, focus[1]

Pôle "rhèmatique" : rhème, commentaire, objet, focus[2]

On notera que ces deux « pôles » sont généralement considérés comme mutuellement exclusifs. Cela n'est toutefois pas toujours le cas, comme par exemple chez Dik (1981) qui considère que topique et focus peuvent parfois se confondre. Il faut noter également que certains auteurs réfutent une division strictement binaire, comme Firbas (1964) qui envisagera plutôt une « échelle du dynamisme communicatif », où chaque élément aura un rôle plus ou moins important dans la progression de l'information.

Pour notre part, nous nous concentrerons tout particulièrement sur le pôle thématique, plus immédiatement lié à la notion d'à propos. Mais on ne peut bien sûr pas envisager l'un ou l'autre de ces pôles de façon totalement indépendante, et il ne s'agit donc là que d'un choix d'éclairage qui nous permettra, le cas échéant, de simplifier notre parcours dans ce domaine extrêmement foisonnant. Parmi les critères énumérés plus haut, nous insisterons essentiellement dans cette partie sur les points 1 (notion d'à propos), 2 (donné/nouveau et identifiabilité), et 3 (position dans la phrase). Nous reviendrons plus tard sur le point 4 (notion de saillance) à travers la théorie du centrage d'attention (cf. section 3.4.2).

Identifiabilité et accessibilité des référents

Différentes approches de ce problème sont envisageables, selon le statut que l'on accorde aux connaissances « extérieures » du lecteur. La première, prépondérante en linguistique et développée notamment par Chafe (1976), et Lambrecht (1994), évalue la « nouveauté » d'un référent du discours en fonction de l'état de conscience du lecteur au fil de la perception d'un texte. Dans cette perspective, les connaissances prêtées au lecteur ne sont pas déterminantes : la nouveauté d'un référent dépend principalement du fait qu'il soit présent ou non à l'esprit du récepteur de l'énoncé au moment même où il le perçoit :

[1] Focus au sens cognitif.
[2] Focus au sens informationnel.

> Knowing something and thinking of something are different mental states. In order for an addressee to be able to process the presuppositions evoked by an utterance it is not only necessary that she be aware of the relevant set of presupposed propositions but that she have access to these propositions and to the elements of which they are composed. In other words, as Chafe as repeatedly emphasized, the conveying of information in natural language not only involves *knowledge*, but also *consciousness*. (Lambrecht, 1994, p. 93)

Tel que conçu par Chafe et Lambrecht, ce principe repose sur les notions d'*identifiabilité* et d'*activation*. Un référent sera considéré comme identifiable dès lors qu'il en existe une représentation partagée dans l'esprit du locuteur et de l'allocutaire au moment de l'énoncé. Inversement, un référent sera considéré comme non identifiable tant que cette représentation n'existe que chez le locuteur. La notion d'identifiabilité est parfois considérée comme liée à la qualité définie ou indéfinie (*definiteness*) d'un syntagme nominal, cette propriété grammaticale pouvant constituer un marqueur formel d'identifiabilité, soit sous la formes d'articles, soit par le biais de l'ordonnancement des mots dans les langues qui n'en disposent pas.

Toutefois, le statut défini d'un syntagme ne peut être considéré comme un indice sûr, dans la mesure où de nombreuses exceptions peuvent être relevées, et où ce type de marque n'existe pas dans toutes les langues. Plus généralement, il est impossible de s'appuyer uniquement sur des marques formelles pour juger de l'identifiabilité d'un référent, et il est donc nécessaire de recourir à des critères d'ordre pragmatique. Le plus évident concerne les termes qui ne peuvent désigner, dans un certain contexte d'élocution, qu'un seul référent. Cela peut par exemple être le cas d'un nom propre, ou encore de syntagmes tels que « le soleil » ou « le président des États-Unis ». D'autres syntagmes qui n'ont pas cette qualité intrinsèque peuvent également l'acquérir dans le cas où un référent dispose d'un statut privilégié dans l'univers du locuteur et l'allocutaire, par exemple « les enfants » ou « la voiture ». Un autre critère concerne les déictiques, qui permettent de désigner un référent appartenant à l'environnement sensible des interlocuteurs (« cette horrible peinture »), ou lié de façon indéfectible à l'un d'entre eux (« ta jambe gauche »). Enfin, dans le cas d'un lien anaphorique, l'identifiabilité du référent est acquise par définition, puisqu'elle est nécessaire à l'anaphore elle-même est présuppose que le référent a déjà été mentionné dans le discours.

Quand un référent est non-identifiable, il apparaîtra comme totalement nouveau (*brand new*) pour le récepteur de l'énoncé. Pour préciser cette situation, Lambrecht reprend la notion d'ancrage due à Prince (1981) : le référent sera dit *ancré* s'il est mentionné relativement à un autre référent qui, lui, est identifiable. C'est par exemple le cas pour « un personne avec qui je travaille » : la personne en question n'est pas identifiable, mais est connectée au référent identifiable qu'est le locuteur. Inversement, un référent non identifiable mentionné sans lien avec un autre objet du discours sera dit *non-ancré*.

D'autre part, la notion d'activation s'appuie sur l'idée, déjà évoquée dans la citation précédente, que la mémoire d'un individu peut contenir une quantité très importante de connaissances et d'informations, dont seulement une petite partie peut faire l'objet de notre attention à un instant donné. Ainsi, tous les référents identifiables ne sont pas simultanément présents à l'esprit des interlocuteurs. Cette hypothèse rejoint les notions de conscience immédiate et de mémoire à court ou long terme, et fonde l'idée qu'un référent peut être, à un point donné du discours, *activé* ou non. Le fait qu'un référent soit activé, c'est-à-dire qu'il soit effectivement présent à l'esprit d'un locuteur, se traduit à l'oral par des marques prosodiques (baisse de l'intonation), et en général par la possibilité de reprise pronominale (pourvu qu'aucune ambiguïté ni contrainte stylistique ne l'empêche). Inversement, la verbalisation d'un référent inactivé se traduit par une accentuation au niveau prosodique, ainsi que par l'obligation d'avoir recours à une lexicalisation complète.

Un référent peut également se trouver dans un état intermédiaire, et sera alors dit *semi-actif* ou simplement *accessible*. Il appartient alors à ce que Chafe qualifie de « conscience périphérique, d'arrière-plan ». Dans ce cas, on peut encore distinguer trois possibilités, selon la nature du phénomène qui est

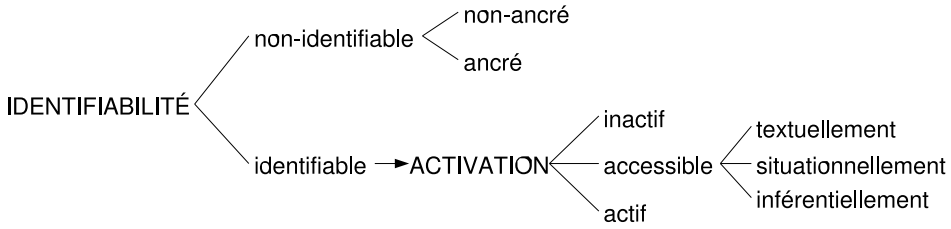


FIG. 3.1 – Identifiabilité et accessibilité des référents (Lambrecht, 1994, p. 109).

à la source de cette semi-activité. S'il s'agit d'un référent préalablement activé mais qui s'est trouvé désactivé au fil du discours, il sera dit *textuellement accessible*. S'il est partiellement accessible par le biais d'inférences à partir d'un autre référent activé ou accessible, il s'agira d'une *accessibilité inférentielle*. Enfin, si le référent est rendu accessible par une certaine proéminence dans le contexte extra-linguistique, il sera dit *situationnellement accessible*.

La figure 3.1 résume l'ensemble des états dans lequel peut se trouver un référent, états qui jouent un rôle important dans la théorie de la structure informationnelle. Nous suivrons dans la section suivante d'autres développements de Lambrecht, qui considère que l'identifiabilité d'un référent n'est qu'un critère parmi d'autres, et doit être prise en compte en corrélation avec d'autres phénomènes.

Mais avant de clore cette section, il nous semble important de mentionner d'autres approches de la dualité donné-nouveau, comme (Clark et Haviland, 1977), qui ne sont pas directement liées au moment précis de la réception d'un énoncé, mais qui considèrent plutôt le texte comme un *tout* confronté aux *connaissances* du lecteur. Cette approche permet de tenir compte du fait que certains référents sont supposés par le locuteur comme connus de l'allocutaire, indépendamment des référents particuliers qui sont cognitivement actifs à un instant donné.

Cette approche se réfère ainsi à la notion de *connaissance partagée*, en l'occurrence entre le locuteur et l'allocutaire, qui correspondent aux connaissances sur lesquelles s'appuient les informations nouvelles, et sans lesquelles l'interprétation du discours sera difficile voire même impossible. Précisons que ce terme de « connaissance partagée » ne doit pas être pris au sens premier : il ne s'agit bien évidemment pas de « comparer » les connaissances de différents individus, mais de chercher à déterminer quelles sont les connaissances de l'allocutaire telles que *supposées* par le locuteur. L'adéquation entre ces connaissances supposées et les connaissances effectives de l'allocutaire sont ici secondaires, *a fortiori* quand celui-ci n'est pas en interaction directe avec le locutaire (discours écrit ou en public par exemple). Pour cette raison, on préférera sans-doute utiliser le terme de « familiarité supposée » introduit par Prince (1981), qui semble désigner plus justement ce phénomène.

Même si elle semble minoriaire en linguistique, cette approche nous semble également digne d'intérêt du point de vue de la recherche documentaire, et se rapproche d'ailleurs des propos de Hutchins (cf. section 2.1). En effet, en considérant que les connaissances du lecteur telles que présupposées par le scripteur ont une incidence sur la structure du discours (hypothèse que nous défendrons ici, notamment dans le chapitre 8), il paraît indiqué les faire intervenir dans le processus de recherche d'information. Dans ce contexte, il paraît indiqué de considérer l'adéquation entre les connaissances de l'utilisateur, que l'on peut raisonnablement chercher à modéliser, et celles qui semblent présupposées dans un document donné. En revanche, l'approche purement cognitive qui consiste à étudier le statut d'activation instantané des référents au fil du discours semble beaucoup plus difficilement opérationnalisable, même si elle décrit probablement plus finement la réalité du discours.

Topique, focus et à propos

Dans (Lambrecht, 1994), l'auteur remarque que les critères d'identifiabilité et d'activation ne sont pas suffisants pour rendre compte de la structure informationnelle d'un énoncé. Il s'appuie pour cela sur les exemples suivants (pp. 110 et 113) :

> **(1)** I heard something terrible last night. Remember Mark, the guy we went hiking with, who's gay ? His lover just died of AIDS.
>
> **(2)** I heard something terrible last night. Remember Mark, the guy we went hiking with, who's gay ? I ran into his lover yesterday, and he told me he has AIDS.

Lorsque le lecteur rencontre le syntagme « his lover », le degré d'activation de son référent est le même dans les deux cas : il n'est pas mentionné dans le discours qui précède, mais pourra être identifié par inférence à partir d'un autre référent du contexte (Mark). En revanche, Lambrecht remarque que la structure syntaxique au sein de laquelle apparaît ce syntagme suggère dans le cas (1) que le référent est déjà accessible dans le discours, alors que dans le cas (2) elle suggère au contraire qu'il était auparavant inactif : « la différence entre (1) et (2) est que (1), par sa structure, enjoint le lecteur à agir comme si le référent du SN était déjà pragmatiquement disponible, contrairement à (2) » (p. 114). Dans ce dernier cas, la proposition « I ran into his lover yesterday » peut en effet être considérée comme *présentationnelle* (au même titre que « Remember Mark ? »), dans le sens où sa fonction est d'introduire un référent au sein du discours plutôt que d'apporter une information à propos du sujet « I ».

Lambrecht en conclut que les paramètres d'identifiabilité et d'activation agissent conjointement avec la structure syntaxique de la phrase au sein d'un système plus général, qualifié de structure *topique/focus*. Dans cette perspective, la notion de topique est définie relativement à la notion d'à propos : « le topique d'une phrase est la chose *à propos* de laquelle est la proposition exprimée par la phrase » (p. 118). Pour éclairer cette définition, l'auteur reprend une caractérisation de l'à propos due à Strawson :

> Statements, or the pieces of discourse to which they belong, have subjects, not only in the relatevely precise sens of logic and grammar, but in a vaguer sens with which I shall associate the words "topic" and "about". [...] Stating is not a gratuitous and random human activity. We do not, except in social desperation, direct isolated and unconnected piecs of information at each other, but on the contrary intend in general to give or add information about what is a matter of standing current interest or concern. (Strawson, 1964, p. 97)

La notion d'à propos dans la définition du topique est ainsi liée au *principe de pertinence* dû à ce même auteur (*Principle of Relevance*) : « un proposition liée à un topique donné sera considérée comme informative seulement si elle véhicule une information pertinente relativement à ce topique », et « de même qu'il existe différents degrés de pertinence, un élément d'une proposition pourra remplir à différents degrés les conditions requises pour accéder au statut de topique » (Lambrecht, 1994, p. 118).

Lambrecht remarque que si elle rejoint la notion de sujet dans son sens classique (qui remonte à Aristote), la notion de topique n'est pas liée de façon univoque à la notion de sujet au sens de la syntaxe moderne, puisque dans de nombreux cas le topique et le sujet grammatical d'un même énoncé ne coïncideront pas. D'autre part, il ne considère pas le topique comme correspondant systématiquement aux constituants initiaux d'une phrase, hypothèse que nous envisagerons dans la section suivante. Au contraire, une définition en termes d'à propos implique une approche purement pragmatique de la dualité topique/focus. Concernant la notion de topique, Lambrecht donne ainsi la définition suivante, qui reprend les différents éléments relatés jusqu'ici :

> A referent is interpreted as the topic of a proposition if *in a given discourse* the proposition is construed as being *about* this referent, i.e. as expression information which is *relevant to* and which increases the addressee's *knowledge* of this referent. [...] Topic is a *pragmatically construed sentence relation*. (Lambrecht, 1994, p. 127)

Cette définition du topique en termes d'à propos et de « pertinence contextuelle » est selon l'auteur liée à la notion de *présupposition pragmatique* : puisque le topique constitue la « préoccuppation courante » des interlocuteurs (c'est-à-dire au sujet de laquelle on souhaitera apporter de l'information nouvelle), ce référent doit nécessairement déjà faire partie du discours, ou appartenir au contexte. Selon les termes de Lambrecht, on dira alors que « le référent appartient à la présupposition pragmatique » (p. 150). Cette notion de présupposition est centrale dans la définition du focus, et l'on peut d'une certaine façon la considérer comme articulation entre les deux notions :

> The focus of the proposition expressed by a sentence in a given utterance context is seen as the element of information whereby the presupposition and the assertion *differ* from each other. The focus is that portion of a proposition which cannot be take for granted at the time of speech. It is the *unpredictable* or pragmatically *non-recoverable* element in an utterance. The focus is what makes an utterance into an assertion. (Lambrecht, 1994, p. 207)

Lambrecht utilise ici les termes *unpredicable* et *non-recoverable* en lieu et place du terme plus habituel de « nouveau » afin de ne pas laisser entendre qu'un focus doit nécessairement être mentionné pour la première fois dans le discours. Par exemple, dans le dialogue suivant, le référent « restaurant » est clairement activé dans la question (et n'est donc plus « nouveau »), mais n'en constitue pas moins un focus dans la réponse (p. 211) :

(Q) Où êtes-vous allés hier soir, au cinéma ou au restaurant ?
(R) Nous sommes allés **au restaurant**.

Cette approche pragmatique implique que ni les relations syntaxiques, ni l'ordre des mots, ni les critères d'identifiabilité/accessibilité ne suffisent à eux seuls à rendre compte de la structure informationnelle d'un énoncé. Il est donc nécessaire de prendre en compte les intentions communicationnelles du locuteur et l'état d'esprit de l'allocutaire pour déterminer relativement à quel référent une proposition apporte une information pertinente. Considèrons par exemple l'énoncé « les enfants vont à l'école ». Selon Lambrecht, il est impossible de statuer sur le statut topical d'un quelconque référent de cette proposition sans prendre en compte son contexte. Pour lever cette ambiguïté, il est nécessaire de considérer l'énoncé en discours, ce que l'on peut « simuler » en le considérant en tant que réponse à différentes questions :

a. Qu'on fait les enfants ce matin ? → Les enfants sont allés **à l'école**.
b. Qui est allé à l'école ce matin ? → **Les enfants** sont allés à l'école.
c. Que s'est-il passé ce matin ? → **Les enfants** sont allés **à l'école**.

Dans le premier cas, le référent du syntagme sujet « les enfants » est bien le topique de la phrase, puisqu'il correspond à son propos, l'énoncé dans son ensemble ayant pour objectif d'apporter une information nouvelle à son sujet. En revanche, dans le cas (b), la proposition a pour but de fournir le référent sollicité par le pronom interrogatif de la question qui précède. De ce fait, « les enfants » constitue ici le *focus*, ce qui sera marqué à l'oral par la prosodie. Dans le cas (c), enfin, la proposition dans son ensemble à pour objectif de répondre au pronom interrogatif de la question, ce qui implique qu'aucun constituant ne puisse accéder au statut d'expression topicale, puisqu'aucun de ses référents n'est mentionné dans la question (les référents partagés par les interlocuteurs appartiennent ici au contexte).

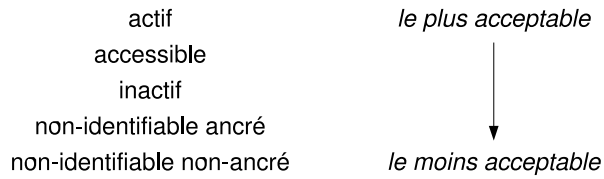


FIG. 3.2 – Échelle d'acceptabilité des topiques (Lambrecht, 1994, p. 165).

Comme nous l'avons déjà évoqué plus haut, Lambrecht considère la notion de topicalité comme intrinsèquement distincte des notions d'identifiabilité/activation vues en section 3.1.1 : les seconds constituent une *propriété* d'un référent, alors que la première désigne une *relation* entre un référent et un proposition. Il ne s'agit cependant pas de concepts totalement indépendants, et s'il considère que ni l'indentifiabilité ni l'accessibilité ne sont pas des conditions suffisantes à l'accès au statut de topique (comme nous l'avons vu dans l'exemple du restaurant, un référent activé peut aussi bien apparaître comme focus), Lambrecht reconnaît qu'ils en sont une condition nécessaire :

> A degree of activeness or at least accessibility is a necessary condition for a referent to be interpreted as having a high degree of topicality. But [...] activeness or accessibility are not sufficient conditions for topic function of a referent in a proposition. (Lambrecht, 1994, p. 164)

Plus précisément, Lambrecht avance que le *degré d'acceptabilité* d'une phrase dépendra du degré d'accessibilité de son topique. Ainsi, une phrase dont le topique serait insuffisamment accessible paraîtra mal formée, et on choisira préférentiellement un référent activé pour former un topique. Entre ces deux pôles se trouvera une échelle d'acceptabilité qui ordonne, comme le montre la figure 3.2, les différents statuts possibles d'un référent.

Influence de la position dans la phrase

Si aucun auteur ne soutient aujourd'hui que la notion de thème ou de topique est *exclusivement* liée à la position des constituants de la phrase, cette dernière est généralement considérée comme un critère plus ou moins important. Halliday (1976), notamment, compte parmi les auteurs qui accordent un statut tout particulier aux constituants initiaux de la phrase. Il est toutefois important de remarquer que Halliday *ne définit pas* le thème par la position initiale de la phrase. Comme nous allons le voir, sa définition du thème est purement *fonctionnelle*. En revanche, il considère que dans le cas particulier de l'anglais (et par extension d'un certain nombre de langues indo-européennes), cette fonction est bien *réalisée* par la position initiale. Il faut également noter que Halliday distingue clairement la *structure thématique* (i.e. en thème/rhème) de la *structure informationnelle* (ie. en donné/nouveau) telle que nous l'avons évoquée jusqu'ici.

Dans (Halliday, 1994), l'auteur reprend les termes de l'école de Prague pour donner une première définition de la structure thématique de la proposition vue comme un *message* :

> In English, as in many other languages, the clause is organized as a message by having a special status assigned to one part of it. One element in the clause is enunciated as the theme ; this then combines with the remainder so that the two parts together constitute a message. [...] Following the terminology of the Prague school of linguists, we shall use the term Theme as the label for this function. The Theme is the element which serves as the point of departure of the message, the part in which the clause is concerned. The remainder of the message, the part in which the Theme is developed, is called in Prague

Thème	Rhème
The duke	has given my aunt that teapot.
My aunt	has been given that teapot by the duke.
Once upon a time	there were three bears.
Very carefully	she put him back on his feet again.
For want of a nail	the shoe was lost.
As for my aunt,	the duke has given her that teapot.
What the duke gave to my haunt	was that teapot.
On the ground or in the air	small creatures live and breathe.
The first chair of G. L. in this country	was established in London.
This responsability	we accept fully.

FIG. 3.3 – La structure thème-rhème (Halliday, 1994, pp. 38–40).

> school terminology the Rheme. As a message structure, therefore, a clause consists of a Theme accompanied by a Rheme (p. 37).

Halliday ne considère pas que cette fonction est universellement remplie par les constituants initiaux de la phrase, mais soutient toutefois qu'en anglais, le thème apparaît *systématiquement* dans cette position, ou plutôt que c'est la position initiale qui constitue dans cette langue la marque de thématisation :

> In some languages which have a pattern of this kind [i.e. theme-rheme], the theme is announced by means of a particle : in Japanese, for example, there is a special postposition *-wa*, which signifies that whatever immediately precedes is thematic. In other languages, of which English is one, the theme is indicated by position in a clause. In speaking or writing English we signal that an item has thematic status by putting it first. No other signal is necessary, although it is not unusual in spoken English for the theme to be marked off also by the intonation pattern (*ibid.*).

Comme le montre la figure 3.3, cette hypothèse implique que la fonction thématique puisse être attribuée à des syntagmes de diverses natures (adverbiaux, prépositionnels) et éventuellement composés, et non pas seulement à des syntagmes nominaux. Dans les énoncés déclaratifs, cette position initiale correspondra le plus souvent au sujet, et le thème est alors *non-marqué*. Dans le cas où il apparaît sous une forme non sujet, comme une forme détachée en initiale, le thème sera considéré comme *marqué*. Selon Halliday la forme la plus marquée est celle où l'objet apparaît en position initiale, comme dans « nature I loved » ou le dernier exemple de la figure 3.3. On notera que la situation est légèrement différente dans le cas des énoncés interrogatifs et impératifs (et plus particulièrement les interrogatives générales[3]), ce que nous ne détaillerons pas ici.

Pour rendre compte des différents types de constituants qui peuvent revêtir une fonction thématique ainsi que de leurs rôles respectifs, Halliday s'appuie sur les métafonctions discursives qui constituent le fondement de son approche fonctionnelle (Halliday et Hasan, 1976; Halliday, 1994). Ces métafonctions sont les suivantes[4] :

- La métafonction *idéationnelle* (*experiential* dans le texte) a trait à l'expression du contenu, de l'information véhiculée.
- La métafonction *interpersonnelle* se rapporte à la dimension « sociale » du discours, c'est-à-dire à l'attitude, l'attention ou les intentions des interlocuteurs en situation.

[3] Interrogatives attendant une réponse en oui/non, par opposition aux interrogatives dites partielles.

[4] Nous reprenons ici les traductions françaises de (mppw00).

On the other hand,	maybe	on a week-day	it would be less crowded.
conjonctif	modal	topical	
textuel	interpersonnel	idéationnel	
thème			rhème

If	winter	comes	can	spring	be far behind
thème (1)			rhème (1)		
structurel	topical		verbal	topical	
textuel	idéationnel		interpersonnel	idéationnel	
thème (2)		rhème (2)	thème (3)		rhème (3)

FIG. 3.4 – Exemples de thèmes multiples (Halliday, 1994).

– La métafonction *textuelle* est relative à la construction des textes, par exemple aux éléments responsables de la cohésion.

À partir de ces métafonctions, Halliday définit différents types de composantes thématiques (Halliday, 1994, p.53) :

– Un *thème textuel* sera réalisé par des connecteurs de discours dits *continuatifs* (« yes », « no », « well », « oh », « now »), *structurels* (« and », « but », « then »), ou *conjonctifs* (« for instance », « in addition », « nevertheless »).
– Un *thème interpersonnel* sera réalisé par des éléments modaux (« probably », « in my opinion », « in general ») ou vocatifs (expression utilisée pour désigner l'interlocuteur ou attirer son attention, par ex. un nom propre).
– Un *thème idéationnel* correspond à la plupart des exemples vus plus haut, où le thème est un sujet, un complément ou un adverbe détaché.

Le fait que ces différents types de constituants apparaissent souvent au sein d'une même phrase conduit Halliday à considérer la notion de *thème multiple*. Il considère que l'ordre *typique* d'apparition de ces différents constituants en initialie de la phrase est le suivant : textuel - interpersonnel - idéationnel. Et si cet ordre n'est pas systématiquement respecté, la composante idéationnelle intervient en tout état de cause en dernière position. Dans ces conditions, le thème est décrit comme incluant *tous les constituants en position initiale jusqu'au premier élément idéationnel*, tout le reste de la proposition appartenant au rhème. Différents exemples de thèmes multiples sont donnés dans la figure 3.4, le second faisant état de la nature *récursive* de la structure thème-rhème que l'auteur exploite pour rendre compte des phrases complexes.

On pourra s'étonner du fait que selon la définition de Halliday, quand plusieurs éléments à portée idéationnelle apparaissent en initiale, seulement la première sera considérée comme appartenant au thème. C'est par exemple le cas dans la phrase suivante :

> At eight o'clock this morning the President left from Barajas to attend the international conference to be held in Rome.

Dans cette configuration, seule la partie soulignée sera considérée comme thématique puisqu'il s'agit de la première composante à portée idéationnelle. Pourtant, comme le remarque Downing (1991), cette analyse semble incompatible avec une définition du thème en termes d'à propos : il est difficile de dire que la phrase est à propos de « at eight o'clock this morning » sans considérer « the President » qui, s'il fallait faire un choix, représenterait de façon plus pertinente l'à propos de la phrase.

Différentes solutions peuvent être apportées à ce problème. (Taboada, 1995) propose pour sa part

une nouvelle définition du thème qui ne pose pas la contrainte de l'à propos, afin de pouvoir y intégrer d'éventuels circonstanciels initiaux comme celui de l'exemple précédent. (Downing, 1991) propose quant à lui une classification alternative des fonctions des éléments initiaux, selon le type de « cadre » qu'ils définissent :

1. thèmes liés aux « participants », qui définissent des cadres « individuels » ;
2. thèmes spatiaux, temporels ou situationnels, qui définissent des cadres circonstanciels ;
3. thèmes de discours, qui définissent des cadres subjectifs et logiques.

Pour notre part, nous serons amenés par la suite à considérer les différentes éléments initiaux à portée idéationnelle comme différentes composantes d'un thème composite qui serait, pour l'exemple précédent, « the President at eight o'clock this morning », ce qui constitue selon nous une représentation « complète » de son à propos. Nous serons amenés à évoquer à nouveau ce même exemple dans la section suivante, le modèle des thèmes composites étant développé au chapitre 8.

3.1.2 La dualité sujet/topique

La distinction entre sujet et topique est posée dans (Li et Thompson, 1976) en tant que critère permettant de distinguer deux catégories de langues naturelles, qui privilégient des structures différentes au niveau de l'énoncé. La première catégorie, à laquelle appartiennent les langues indo-européennes, privilégie la structure en *sujet/prédicat* (SP), qui se manifeste dans sa forme canonique par une structure de type sujet-verbe-objet. La seconde catégorie, à laquelle appartient notamment le Chinois, privilégie une structure en *topique/commentaire* (TC), le topique correspondant ici à une expression extra-prédicative détachée. Pour illustrer cette distinction, les auteurs donnent l'exemple suivant (p. 483) :

John *sujet*	hit Mary. *prédicat*
As for education, *topique*	John prefers Russell's ideas. *commentaire*

On notera qu'il s'agit généralement de *préférence* pour l'une ou l'autre des formes, et que dans beaucoup de langues les deux formes existent (l'exemple ci-dessus en témoigne pour ce qui est de l'anglais). On notera également que certaines langues, comme le Japonais, utilisent indifféremment les deux formes, et qu'il en existe qui n'en utilisent aucune, comme le Tagalog[5].

Dans ce contexte, les propriétés d'un topique sont les suivantes : il doit être identifiable au sens défini dans la section précédente, mais ne fait pas nécessairement partie de la prédication principale. Le sujet, au contraire, n'est pas nécessairement identifiable, mais fait partie de la prédication principale. On remarquera que des distinctions analogues ont été établies par différents auteurs, souvent sous des dénominations différentes. Le critère d'appartenance à la proposition principale est notamment mentionné chez (Dik *et al.*, 1981), qui distingue ainsi *theme* et *topic*, le topique étant celui qui appartient à la prédication proprement dite, et le thème celui qui établit un cadre référentiel dans lequel on invite l'allocutaire à rechercher la pertinence de la prédication principale. Dik distingue également *theme* et *tail*, le premier correspondant à une expression non argumentale à gauche de la proposition, tandis que que le second correspond à une expression non-argumentale à droite. Cette même distinction se trouve également chez Lambrecht sous les étiquettes respectives de TOP, topique détaché à gauche, et A-TOP, topique détaché à droite ou anti-topique (Lambrecht, 1994, p. 128).

Nous nous intéresserons par la suite tout particulièrement à une catégorie particulière de topique non argumental que Chafe qualifie de *« Chinese-style » topics*, en référence à la préférence de la langue

[5]Langue des Philippines.

chinoise pour la forme topique-commentaire). Ces topiques « limitent l'applicabilité de la prédication principale », en « posant un cadre spatial, temporel ou individuel au sein duquel [elle] se tient » (Chafe, 1976, p. 50). En voici quelques exemples :

In volleyball, you stand still playing.

Dans l'Ouest, le taux de retard scolaire est maintenant en régression.

La nuit, tous les chats sont gris.

Pour reprendre la terminologie de Lambrecht (Lambrecht, 1994, p. 118), nous désignerons ces topiques par le terme de *scene-setting topics* (ou SST), qui nous semble exprimer leur fonction de façon assez précise. Nous serons amenés dans le chapitre 8 à nous intéresser à ces topiques de deux points de vue :

- Nous envisagerons d'une part la relation entre les SST et la notion d'*à propos*.
- Nous nous intéresserons d'autre part à l'intérêt de cette relation dans une approche *discursive* de la notion de thème.

Dans le premier cas, nous serons amenés à considérer un SST en tant que composante d'une représentation plus large de l'à propos. Il semble en effet nécessaire de prendre en compte de façon particulière la fonction topicale d'un SST, qui nous paraît sensiblement différente de celle qu'on peut attribuer à d'autres types de constructions détachées. Pour Lambrecht, les dislocations (constructions détachées à gauche et parfois à droite) sont considérées comme proche des constructions dites *présentationnelles*, en permettant de réintroduire un référent topical devenu inactif sous une forme lexicalisée. C'est par exemple le cas des énoncés (2) et (3) ci-dessous, la forme (1) représentant une forme présentationnelle « classique ».

(1) A wizard once was very wise, rich, and married to a beautiful witch.

(2) Now the wizard, he lives in Africa.

(3) As for the wizard, he lives in Africa.

Dans ces trois cas, le statut topical du référent « wizard » est sans ambiguïté du point de vue de l'à propos. Mais la situation est différente dans l'exemple suivant, que nous avons déjà mentionné relativement au modèle fonctionnel d'Halliday comme posant un problème si l'on considère le thème ou le topique relativement à cette notion d'à propos :

At eight o'clock this morning the President left from Barajas to attend the international conference to be held in Rome.

En effet, alors que le référent du syntagme temporel constitue bien un SST (et le thème pour Halliday), il ne peut pas être raisonnablement considéré comme représentation de l'à propos de l'énoncé qui est clairement « le président » (le topique chez Lambrecht ou Dik). De fait, Lambrecht considérera dans ce cas que les deux éléments sont bien topicaux, quand Dik y verra un thème suivi d'un topique.

Remarquons tout d'abord que ce type de construction n'est pas propre aux adverbiaux spatiaux ou temporels, puisque la fonction de *scene-setting* peut se matérialiser par d'autres types de syntagmes prépositionnels ainsi que par des constructions de type « concernant x » (dites *as-for* en anglais) :

(1) Dans l'enseignement primaire, le taux de retard scolaire est en forte diminution.

(2) En ce qui concerne le nucléaire, l'AIEA estime avoir épuisé sa mission de destruction et d'enlèvement des matières prohibées.

Pour décrire l'à propos d'énoncés de ce type, nous serons amenés à considérer les SST en corrélation avec une notion de topique « central » ou « principal », l'ensemble constituant différentes composantes de ce que nous qualifierons de thèmes « composites ». Nous rejoindrons ainsi Le Goffic (1994) :

> Même s'il n'apparaît pas, psychologiquement, comme ce dont l'énoncé parle, le circonstant initial de temps ou de lieu sert de repère, permettant de localiser ou de dater un événement ou un situation dans son ensemble. Il peut ainsi constituer un véritable thème ou une partie du thème (en liaison avec le sujet). (Le Goffic, 1994, p. 464)

Ainsi, cet auteur considérera que le thème de « la nuit, tous les chats sont gris » comprend aussi bien le cadre de la phrase que son sujet, formant un couple du type « les chats - la nuit », pour peu que la gent féline constitue bien l'à propos des propositions qui suivent (et non pas « la nuit », qui constituerait alors le thème à lui seul).

Concernant la dimension discursive des SST, nous nous rapporterons bien évidemment au modèle de l'encadrement de Charolles, sur lequel nous reviendrons longuement dans les sections 3.4.1 puis 7, mais nous examinerons également d'autres configurations discursives que nous considérerons comme équivalentes du point de vue qui nous intéresse ici.

Nous aurons l'occasion de revenir par la suite sur les différents problèmes soulevés dans cette section, notamment au niveau discursif avec Charolles (section 3.4.1). Notre propre approche du problème, orientée vers l'analyse thématique automatique, sera quant à elle développée dans le chapitre 8.

3.2 De la phrase au texte

3.2.1 Topiques de discours

La notion de topique de discours que nous visons dans l'optique de l'ingénierie documentaire est liée à un point de vue non seulement extérieur à la phrase, mais aussi à une approche *globale*. En effet, nous avons vu dans la section précédente (3.1) que l'analyse de l'analyse de la structure informationnelle de la phrase ne peut se concevoir si on considère cette dernière de façon totalement isolée, et que la dimension pragmatique est essentielle dans ce domaine. Toutefois, bien qu'ils puissent ainsi conduire à considérer des séquences de phrases, les modèles que nous avons évoqués dans cette section conservent un point de vue que l'on peut qualifier de *local* : chaque étape de l'analyse reste centrée sur un énoncé, même si l'on considère ce qui précède ou ce qui suit. Or, il est clair que cette approche locale ne permet pas de rendre compte du topique ou du thème d'un discours pris comme un tout, qui ne se réduit pas pas à la somme ou à la combinaison des topiques des phrases qui le composent (Fairthorne, 1969; Hutchins, 1977). Il est donc nécessaire de considérer un texte comme un tout cohérent pour accéder à son topique, en adoptant une approche globale.

Mais cette notion de topique de discours est beaucoup moins décrite par la linguistique que son « équivalent » phrastique, même si le terme apparaît régulièrement dans la littérature. Il est certes communément admis, si l'on considère l'à propos comme la relation fondamentale qui relie un thème (ou le topique) à l'objet auquel il se rapporte, que cette relation s'applique aussi bien à une phrase qu'à tout autre fragment textuel cohérent (paragraphe, document, collection, etc.). Mais bien que chacun d'entre nous bénéficie d'une perception intuitive de ce phénomène, il demeure extrêmement difficile d'en rendre compte d'un point de vue théorique, pour diverses raisons.

Nous avons pu constater précédemment qu'il n'existe pas de consensus fort concernant la notion de thème au niveau phrastique, fut-ce pour décrire des énoncés simples. On se sera rendu compte que ce problème est effectivement très loin d'être trivial : bien que la forme des énoncés soit fortement

contrainte par la syntaxe, cette dernière ne suffit pas à rendre compte de leur structure informationnelle, et l'introduction de critères moins formels, d'ordre cognitif ou plus généralement psychologique, est donc nécessaire. Compte tenu de ces considérations, on imagine aisément la complexité inhérente à la modélisation d'un phénomène analogue au niveau du discours : le passage à ce niveau ne relève pas d'un simple changement d'échelle, mais fait apparaître des structures protéiformes (la variété des phénomènes susceptibles de contribuer à l'organisation du texte est considérable) et beaucoup moins contraintes (il n'existe aucune théorisation reconnue de ce que pourrait être une « grammaire de texte », sauf cas très particuliers).

Un autre problème concerne le lien entre topique de discours et topique de phrase : au delà de la complexité inhérente au discours lui-même, il serait souhaitable pour une théorie du topique discursif qu'elle soit capable de rendre compte de ses relations avec les topiques phrastiques. Deux niveaux « d'intégration » sont alors possibles. On pourra considérer les deux types de topiques comme différents *par nature*, et chercher à décrire comment ils collaborent dans le texte. Ou bien on les considérera comme différentes manifestations d'un même phénomène, à des niveaux de grain différents, et on cherchera à produire une théorie capable d'en rendre compte de façon *uniforme*. En tout état de cause, il est manifestement impossible d'exporter de façon immédiate au niveau discursif les modèles de la structure informationnelle de la phrase, et il semble que ces derniers soient d'un secours limité dans ce domaine, voire une difficulté lorsqu'il s'agit d'unifier les deux ordres de grandeur.

En somme, la notion de topique discursif est liée à une problématique extrêmement vaste et complexe, dont aucune théorie actuelle n'est vraisemblablement capable de rendre compte en toute généralité. De fait, cette notion a souvent donné lieu à des définitions qui, bien que correspondant assez exactement à l'intuition que nous pouvons en avoir, n'en restent pas moins floues. Par exemple chez Jones :

> The ideas about theme developed in this study have their roots in the rather intuitive understanding of theme that most of us had in primary and secondary school - that is, that theme is « main idea » in a text. The theme-line of a text is its « central thread ». Theme also may be described as a « minimum generalization » of a text : a statement broad enough to represent the entire text, yet specific enough to represent its uniqueness. (Jones, 1977)

Plus récemment, la notion de topique de discours a été abordée par Chafe (2001), d'un point de vue plutôt conversationnel. L'auteur y fait appel à une notion de topique global, dit *supertopic*, valant pour le discours pris dans sa globalité (ou une large portion de celui-ci), qu'il distingue d'une notion de topique « intermédiaire », de moindre empleur mais plus structuré, dit *basic-level topic* ou simplement *topic*. Mais là encore les termes restent très vagues :

> A topic in this sense is a coherent aggregate of thoughts introduced by some participant in a conversation, developed either by that participant or by several participants jointly, and then either explicitly closed or allowed to peter out. (Chafe, 2001, p. 674)

Quand à la nature même d'un topique de discours, différentes hypothèses ont été envisagées, sans qu'aucun modèle prédominant n'apparaisse clairement :

- Le topique pourrait être représenté par un ou plusieurs syntagmes nominaux (Bradsford et Johnson, 1972). Bien que cette approche ait peu de crédit du point de vue linguistique, elle connaît le succès que l'on sait dans le domaine de l'indexation (manuelle ou automatique), comme nous l'avons vu plus tôt.
- De façon légèrement plus élaborée, le topique d'un fragment textuel peut être représenté sous la forme d'une proposition représentative du contenu de l'ensemble du fragment (Keenan et Schieffelin, 1976).

– La macro-structure d'un texte, c'est-à-dire un réseau de propositions représentatif du schéma global du texte, peut également être considérée comme représentative de son à propos, comme nous le verrons dans la section suivante.
– Brown et Yule préfèrent quant à eux la notion de « cadre topical » (cf. infra), qui vient circonscrire les différentes interprétations possibles de l'à propos d'un texte.

Ce flou qui entoure le concept de topique discursif doit probablement nous inciter à le considérer comme une notion intrinsèquement *relative*. On peut en effet douter de la possibilité d'en donner une définition précise qui reste valable pour tout type de discours, écrit ou oral, dialogique ou monologique, quel que soit son genre, son contexte, etc. Et au delà des propriétés du texte, il est également plausible que l'objectif de l'analyse elle-même ait une incidence sur la conception que l'on adoptera. Dans le cas particulier de l'ingénierie documentaire, on sera ainsi amené à considérer différentes formes de thème selon le degré de représentativité que l'on souhaite en obtenir : on pourra viser une représentation condensée du contenu sémantique global si cette représentation doit se substituer au document, ou, à l'extrême inverse, une représentation minimale (par exemple sous la forme de mots-clefs) si cette représentation ne constitue qu'un point d'entrée vers le document. Indépendamment du contexte, il paraît également probable que différents modes d'organisation du discours doivent être caractérisés par des thèmes de différentes natures, par exemple en fonction de la structure rhétorique du passage considéré. Plus généralement, on peut donc penser qu'il n'est pas possible, ni même souhaitable, d'obtenir une définition « universelle » de la notion de thème discursif, sauf à se limiter à des notions très générales.

Dans le même ordre d'idées, une des propositions formulées dans (Brown et Yule, 1983), qui constitue une contribution importante à ce sujet, est qu'un thème de discours n'a pas d'existence propre : un texte n'a pas de topique en soi, mais ce sont les locuteurs ou scripteurs qui « visent » un topique. Pour les allocutaires ou lecteurs, il existe pour un texte donné une variété de topiques possibles, au sein desquels il n'est pas possible de faire un choix objectif, dans la mesure où la notion d'à propos est propre à chaque interprétation. Cette approche conduit ces mêmes auteurs à préférer la notion de « cadre topical » (*topic framework*), qui a pour propriété de circonscrire les différentes conceptions « raisonnables » du topique d'un texte (ou d'un passage) donné :

> What is required is a characterisation of 'topic', which would allow each of the possible expressions, including titles, to be considered (partially) correct, thus incorporating all reasonable judgements of 'what is being talked about'. We suggest that such a characterisation can be developed in terms of a *topic framework*. [...] Those aspects of the context which are directly reflected in the text, and which need to be called upon to interpret the text, we shall refer to as *activated features of context* and suggest that they constitute the contextual framework within wich the topic is constituted, that is, the *topic framework*. (Brown et Yule, 1983, p. 75)

Les auteurs considèrent donc qu'il préférable de se limiter à l'énumération d'un ensemble de référents du contexte qui sont « importés » dans un discours (individus, objets, lieux, repères temporels, etc.) et nécessaires à son interprétation. En revanche, le choix d'une ou plusieurs propositions établissant un lien entre ces référents relèverait de la pure interprétation, et ne pourrait être fait objectivement par l'analyste. On notera que cette approche n'est pas pour autant sans rapport avec les notions de donné/nouveau, de topique/commentaire et d'activation que nous avons évoquées plus avant, puisque d'une certaine manière, l'établissement d'un cadre topical correspond bien à la sélection de référents « partagés » à un instant donné par les participants du discours, à partir desquels on considère une information nouvelle. De fait, les notions de connaissance partagée et d'activation sont explicitement mentionnées par les auteurs :

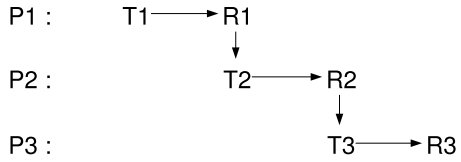
> The topic framework [...] represents the area of overlap in the knowledge which has

been activated and is shared by the participants at a particular point in a discourse. (ibid, p. 83).

Le problème du passage de la phrase au discours, enfin, est lui aussi abordé de façons très diverses dans la littérature, et nous servira de fil conducteur dans cette section, avant d'arriver aux approches que nous considérons comme proprement « globales », c'est-à-dire réellement centrées sur le texte. Une contribution importante à ce sujet du lien phrase-discours est due à Dik, qui tente d'inclure les deux notions de topique discursif et phrastique au sein d'une même approche fonctionnelle, qui sera brièvement discutée dans la section 3.2.3. Mais nous allons d'abord nous intéresser dans la section qui suit à la notion de progression thématique.

3.2.2 Progressions thématiques

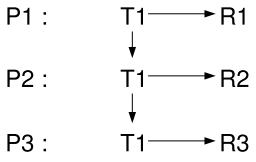
A partir du moment où l'on admet une dichotomie du type thème/rhème au niveau phrastique, une étape naturelle du passage vers le texte s'appuie sur l'étude de l'évolution des référents thématiques et rhématiques au fil de la séquence du texte, c'est-à-dire d'une phrase à l'autre. Cette évolution est souvent appelée *progression thématique*, et constituera l'objet de cette section en nous appuyant principalement sur (Combettes et Tomassone, 1988)[6]. Un premier type de progression thématique est dit *linéaire*, et correspond au cas « classique » où le rhème de chaque énoncé constitue le thème de celui qui suit :



Il est évident que dans cette configuration comme dans celles qui suivront, il n'est pas nécessaire que le constituant rhématique du premier énoncé soit repris *à l'identique* dans l'énoncé qui suit, il pourra par exemple s'agir seulement de l'objet de la prédication, d'un méronyme, d'un hyponyme, etc. Voici un exemple de progression linéaire (comme dans les exemples qui suivront, nous avons souligné les portions reprises du thème en pointillés, et celles du rhème par un trait continu) :

> Bien avant la naissance de l'enfant, le squelette se compose d'organes entièrement constitués de cartilages. Progressivement, ces "os" acquièrent leur dureté par formation de substance osseuse périphérique à partir du périoste et par remplacement du cartilage par du tissu osseux à partir de centres internes d'ossification. Ces transformations, déjà avancées à la naissance, se continuent jusqu'à l'âge adulte.
>
> Source : APH

Un autre type de progression thématique est dite à *thème constant*. Dans cette situation, le thème reste le même d'un énoncé à l'autre :

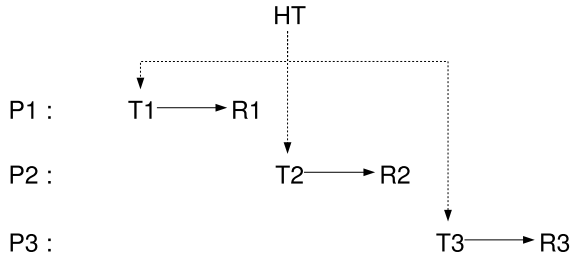


[6]Les exemples donnés dans cette section sont également empruntés à cet ouvrage, leur annotation étant en revanche de notre propre fait.

Combettes et Tomassonne remarquent que ce type de progression est sans doute le plus fréquemment utilisé, et peut être employé dans des fragments de textes longs dans la mesure où l'on bénéficie d'un « point d'ancrage » bien établi. Cette situation fait apparaître ce que les auteurs qualifie d'*hyperthème*, c'est-à-dire un thème qui se rapporte à une séquence d'énoncés. Voici un exemple d'une telle progression, où l'hyperthème serait « la gravitation » :

> Gravitation : attraction de la matière par la matière. C'est une force universelle, qui ne souffre aucune opposition. Rien ne lui fait écran. Elle nous retient sur la terre, maintient celle-ci arrimée au soleil... C'est encore elle qui rend le système solidaire de la galaxie.
>
> Source : VE

Au contraire des progressions précédentes qui sont purement linéaires, les progressions dites à *thèmes dérivés* font appel à des relations qui ne lient plus seulement les référents de deux énoncés consécutifs, mais les référents de plusieurs énoncés à une entité extérieure, textualisée ou non. Cette structure suppose en effet l'existence d'un hyperthème (HT dans le diagramme ci-dessous), dont les thèmes des différents énoncés sont des *sous-thèmes* :



Ce type de progression apparaîtra plus ou moins clairement selon que l'hyperthème est explicitement spécifié ou non. Voici un exemple de progression à thèmes dérivés où l'hyperthème (« les Touargegs ») est mentionné explicitement :

> Les Touaregs sont un peuple fier et belliqueux. Les hommes portent presque toujours un voile qui leur couvre tout le visage à l'exception des yeux. [...] Les hommes et les jeunes gens s'occupent des troupeaux ; les femmes et les jeunes filles ramassent le bois, font la cuisine, filent la laine, tannent et teignent le cuir, et prennent soin des enfants.
>
> Source : AF

Dans d'autres cas, l'hyperthème n'est pas textualisé, mais doit être « reconstitué » à partir des différents sous-thèmes. C'est par exemple le cas dans l'extrait suivant, où l'hyperthème est une zone géographique qui entoure Nairobi :

> Au nord de Nairobi, la vallée cultivée ne tarde pas à se trouver enserrée entre la chaîne de l'Aberdare et le massif du mont Kenya. [...] Au delà, le sol va encore trouver moyen de se plisser vigoureusement avec les collines de Samburu et de Metu. [...] À l'Ouest de Nairobi, même folie. En suivant cette plaine, vous déboucherez sur la vallée du Rift [...]. En remontant vers le nord, la plaine ne cesse de prendre de l'altitude.
>
> Source : KENYA

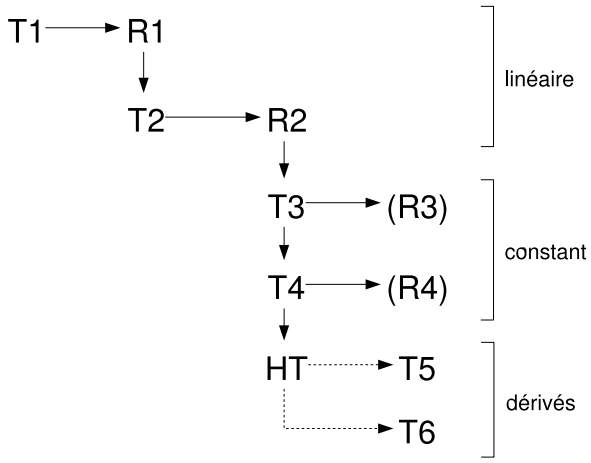


FIG. 3.5 – Progression thématique complexe

Remarquons que dans ce cas précis, on pourrait objecter que la progression thématique est très proche d'une progression à thème constant : on pourrait observer que l'hyperthème « Nairobi » est bien repris, explicitement ou non, en tant que thème de chaque énoncé, même s'il est systématiquement accompagné d'une restriction géographique. Ce type de construction, à la limite entre progression linéaire et à thèmes dérivés, nous semble appartenir à une classe de structures thématiques importante, où un même hyperthème est décliné selon un critère particulier, par exemple spatial. Ce phénomène, qui peut se manifester sous des formes diverses, a déjà été discuté brièvement dans la section 3.1, et sera plus longuement discuté dans le chapitre 8.

Il est évident que dans les textes « réels », les divers modes de progression thématique que nous avons évoqués jusqu'ici seront généralement utilisés en alternance, ne serait-ce que pour des raisons stylistiques. On trouvera également des structures *hybrides*, associant par exemple progressions linéaires et à thèmes dérivés comme nous l'avons déjà vu avec Hutchins (cf. figure 2.1 p. 33). Chez Combettes et Tomassonne, ces progressions sont dites *complexes*, dont voici un exemple :

> Le long du Niger inférieur, on trouve partout une population très dense$_{R1}$, sauf dans la région marécageuse du delta. Le groupe ethnique le plus important$_{T2}$ est celui des Ibos$_{R2}$. Les Ibos$_{T3}$ n'ont jamais eu de chef suprême. Ils$_{T4}$ sont divisés en petites tribus. La principale activité des Ibos reste l'agriculture$_{T5}$, mais beaucoup d'hommes ont quitté les villages pour chercher du travail en ville$_{T6}$.
>
> Source : AF

Dans cet extrait, on observe dans un premier temps une progression linéaire, qui permet de passer de R1 à T2 pour finalement introduire le groupe ethnique des Ibos (T3) dont il sera question par la suite, sous la forme d'une progression à thème constant (T3 = T4). Enfin, une progression à thèmes dérivés permet d'évoquer deux types de ressource des Ibos que sont l'agriculture ou le travail en ville[7] (l'hyperthème de cette troisième partie est donc « les ressources des Ibos »). Le schéma correspondant est représenté par la figure 3.5.

Il est important de remarquer qu'indépendamment des différents types de progression thématiques possibles, il arrive que la chaîne présente des *ruptures*, c'est-à-dire qu'aucun élément thématique ni rhématique ne fait l'objet d'une reprise explicite sous forme thématique. Dans ce cas, la cohérence

[7] Nous prenons ici une légère liberté par rapport à l'analyse de Combettes et Tomassonne, qui donnent les syntagmes « La principale activité... » et « Beaucoup d'hommes... » pour T5 et T6 respectivement.

globale sera bien sûr maintenue par d'autres moyens, mais cela montre bien que la notion de progression thématique ne suffit pas à en rendre compte totalement, du moins si on ne la considère qu'à un niveau local. C'est par exemple le cas dans cet extrait :

> θ **Bataille de palmes**
>
> § Pour remplacer les palmiers de la Croisette qui avaient gelé en janvier 1985, la ville de Cannes avait passé un marché avec une entreprise de Roubaix pour la livraison de trente palmiers adultes. Les arbres, de superbes spécimens de phoenix canariensis hauts de six mètres, facturés 23 000 francs chacun, ont été livrés comme convenu et replantés sur la Croisette.
>
> § Festivaliers et estivants pouvaient croire que ces altiers palmiers avaient grandi sous le bon soleil de la côte d'Azur...
>
> § Las ! Voici que la ville espagnole de Premia del Mar, près de Barcelone, réclame le retour de ses trente palmiers centenaires, arrachés dans un de ses parcs à l'insu du maire. L'adjoint à l'urbanisme, responsable de la transaction, a dû démissionner. Mais les édiles cannois, forts de leur bon droit, ne veulent rien entendre : les palmiers espagnols resteront sur la Croisette... jusqu'au prochain coup de froid.
>
> Source : LM1

Dans cet extrait, on voit que les thèmes des phrases initiales des paragraphes 2 et 3 (soulignés en pointillés) sont totalement nouveaux, puisqu'ils ne reprennent aucun référent qui aurait déjà été introduit dans le texte qui précède. De ce fait, ces thèmes ne sont pas liés aux progressions thématiques de ce qui précède, ce qui représente une certaine forme de rupture. Il est toutefois clair que la cohérence globale de l'ensemble n'est pas pour autant remise en cause, notamment grâce au référent des palmiers qui reste présent même s'il ne constitue pas systématiquement un thème au sens propre. La notion de progression thématique ne suffit donc pas à rendre compte de la cohérence du texte pris dans son ensemble, et il faudra donc faire intervenir des structures d'un autre ordre.

Pour conclure, on peut considérer la notion de progression thématique comme un première pas vers une approche discursive, dans la mesure où elle permet de rendre compte de relations entre des phrases qui ne sont pas nécessairement contiguës, par exemple dans le cas de progressions à thèmes dérivés. Ce type de progression fait également apparaître une notion d'hyperthème qui n'est pas toujours connectée à un référent explicitement introduit dans le discours, ce qui va dans le sens d'un thème discursif qui résulterait plutôt d'inférences sémantiques plutôt que du simple choix d'un référent proéminent. Enfin, la notion de progression thématique semble pouvoir se prêter à une analyse récursive, qui pourrait conduire à une analyse de fragments textuels relativement larges. Toutefois, comme nous venons de le voir, cette approche ne suffit pas toujours à rentre compte de la cohérence du discours, et il s'agirait donc plutôt d'une approche « locale étendue » qui nécessite de s'intégrer à une approche plus globale comme nous le verrons avec les macro-structures de Van Dijk.

3.2.3 Stratégies discursives de topicalisation

Comme beaucoup des travaux relatés ici, l'approche de Dik (1981, 1989) est basée sur la notion d'à propos, et étudie la notion de topique de discours en tant que « référent central au sein d'une unité de discours ». Une particularité importante du point de vue ainsi adopté par Dik est de viser explicitement à concilier les notions de topique de phrase et de topique de discours. Selon cet auteur, le discours pris dans son ensemble introduira le plus souvent plusieurs topiques, ordonnés de façon linéaire ou hiérarchique. Chacun de ces topiques sera initialement considéré comme « nouveau » (*new topic*), avant de devenir « donné » (*given topic*) lorsqu'il sera repris ultérieurement. Dik introduit la notion de

sous-topique (*sub-topic*), qui correspond à un topique qui n'est pas énoncé explicitement, mais qui est rendu disponible par inférence à partir d'un topique explicite, ainsi que la notion de topique réactivé (*resumed topic*), qui intervient lorsqu'un topique n'a pas été mentionné depuis longtemps, et doit être réintroduit pour accéder à nouveau au statut de topique donné.

On reconnaîtra bien sûr ici des concepts similaires à ceux présentés en section 3.1, notamment avec Lambrecht. Il est intéressant de remarquer qu'alors que ce dernier auteur adopte explicitement une posture centrée sur l'énoncé (Lambrecht, 1994, p. 117), il est conduit à développer une caractérisation très fine du statut cognitif des référents en faisant appel à des critères essentiellement pragmatiques. Inversement, la catégorisation des topiques de Dik est moins précise, mais emploie explicitement le terme de *discourse topic*. De fait, Dik décrit les procédés discursifs conduisant à l'instauration et la persistance des topiques de façon très explicite, alors que chez Lambrecht ils sont mentionnés en tant que critères pour l'analyse de la structure informationnelle, et non comme phénomènes discursifs en tant que tels. Dik emploie le terme de « stratégie de la topicalité » (*topicality strategy*) pour désigner ces procédés (Dik, 1989, pp. 268–277) :

1. Stratégies visant à introduire un nouveau topique de discours :
   (a) Énoncé méta-linguistique de ce qui deviendra le topique, par exemple : « Je vais vous raconter l'histoire de X ».
   (b) Construction existentielle (présentative chez Lambrecht), par exemple « Il était une fois X ».
   (c) Topique en position objet, par exemple « Dans le bureau nous vîmes entrer X ».
2. Stratégies visant à maintenir un topique en discours, se traduisant par des chaînes topicales (*topic chains*) : utilisation des moyens référentiels, ou de structures syntaxiques parallèles. Un topique donné peut également être maintenu grâce à un sous-topique, c'est-à-dire un topique qui peut être inféré selon diverses relations (hypéronymie, méronymie, etc.).
3. Stratégies visant à réactiver un topique : re-lexicalisation, déplacement topical explicite, etc.

Dans la perspective de Dik, c'est donc la mise en oeuvre des stratégies de topicalisation par le locuteur ou le scripteur qui se traduit en discours par l'établissement de topiques de discours.

3.2.4 De la nécessité d'adopter une approche « globale »

Il est intéressant de remarquer que si Dik revendique effectivement une approche discursive, celle-ci est manifestement conçue comme une « extension » d'une théorie par ailleurs centrée sur la phrase, cette dernière demeurant l'objet d'étude principal du courant fonctionnaliste auquel il appartient. Par conséquent, bien que certaines propositions de cet auteur soient explicitement formulées en termes de topiques de discours, on ne peut pas considérer son approche comme fondamentalement différente des modèles de la structure informationnelle qui incorporent la dimension pragmatique, comme chez Lambrecht.

Plus généralement, les différentes approches que nous avons eu l'occasion de discuter jusqu'à présent semblent converger vers un point qui serait « entre deux eaux », c'est-à-dire entre une approche centrée sur la phrase qui « loucherait » vers la pragmatique, et une approche orientée vers le discours qui tendrait à reproduire les schémas de la structure informationnelle de la phrase. Il est sans doute nécessaire, pour échapper à cette position de notre point de vue assez peu satisfaisante, de chercher à s'éloigner, ne serait-ce que temporairement, de la seule observation de l'organisation de référents en discours, qui semble trop intimement liée à des structures séquentielles et fondamentalement locales telles que les chaînes de références et les progressions thématiques (au sens inter-phrastique, cf. supra).

Il semble en effet que cette approche soit trop contraignante pour s'appliquer au discours dans sa globalité, pour plusieurs raisons.

On peut tout d'abord considérer que la notion même de référent est trop restrictive quand il s'agit de topique de discours : s'il semble raisonnable de considérer qu'un constituant donné d'un énoncé, pris dans un contexte donné par un interprétant donné, est bien en relation avec une entité conceptuelle clairement identifiée, il n'en va peut-être pas de même au niveau discursif. On peut en particulier douter du fait qu'un topique de discours puisse effectivement être représenté par un ou plusieurs référents choisis parmi ceux auxquels se réfèrent les phrases qui le composent.

Il semble au contraire plus raisonnable de penser que le topique d'un discours est le fruit de la combinaison de différents référents, aboutissant à une structure d'ordre supérieur, ce que nous discuterons par exemple au sujet des macro-structures de Van Dijk (1977). Il faut également prendre en compte la notion de cohésion lexicale (Halliday et Hasan, 1976), qui suppose que l'ensemble des référents explicites d'un discours est en fait lié à un réseau plus large de référents, dont certains ne sont pas explicitement mentionnés dans le texte mais qui peuvent néanmoins contribuer à la formation du topique du discours. En allant plus loin, on peut également penser qu'une telle structure ne se résume pas à la mise en relation des référents du discours, mais devrait idéalement rendre compte des interactions entre ces référents. On peut en effet considérer que la mise en relation de deux référents produit non seulement une structure d'ordre supérieur, mais conduit également à la modification mutuelle également les dits référents : la mise en présence de deux référents x et y ne produirait pas seulement une structure du type f(x,y), mais plutôt du type f(x',y') où x' est le produit de la modification de x par y et réciproquement. Nous aurons là aussi l'occasion de revenir sur cette question par la suite, notamment avec Rastier (1995a).

Un autre problème concerne le caractère évolutif des référents en discours : si un référent peut vraisemblablement être considéré comme « stable » au sein d'une chaîne référentielle de longueur réduite, on peut penser que ce n'est pas toujours le cas à des niveaux de grain plus importants, et qu'un référent « évolue » au fil des reprises et réactivations. Cela pourrait par exemple être le cas dans les textes où une même notion est définie par étapes successives, partant d'une définition intuitive pour arriver à une définition plus formelle[8]. Dans cette situation, un même terme pourra désigner des référents sensiblement différents au fil du texte, mais pas totalement distincts : il s'agirait plutôt de référents en évolution, en construction. De ce point de vue, on ne peut assimiler un référent de discours avec le référent d'un énoncé particulier dans la mesure où ce dernier ne constituerait qu'une vue « instantanée » sur un objet en constante évolution.

Un autre problème lié à l'étude « classique » du statut des référents et de leur propagation interphrastique repose sur un certain attachement à la notion de chaîne référentielle. De ce point de vue, et même en disposant d'une échelle d'identifiabilité/activation évoluée comme celle que nous avons rencontrée dans la section 3.1, il reste difficile d'envisager la valeur propositionnelle d'un énoncé en dehors des référents qui y sont introduits ou repris de façon *explicite*. Il existe pourtant des modes de propagation des référents qui ne reposent pas sur les mécanismes référentiels, comme nous le verrons notamment au sujet de l'encadrement du discours (Charolles, 1997) avec la notion de *portée* (cf. section 3.4.1). Or ces référents peuvent jouer un rôle particulièrement important dans la construction d'un topique de discours, puisqu'ils appartiennent justement à des structures intrinsèquement textuelles, puisque en marge de la structure prédicative des énoncés. Il paraît donc indispensable de prendre en considération une certaine diversité des modes de propagation des référents, qui n'est pas du seul ressort des mécanismes référentiels.

[8]C'est notamment fréquent pour les textes qui traitent de notions « fuyantes » comme la notion de thème elle-même.

3.3 Approches centrées sur le texte

3.3.1 Thème et macro-structures du discours

La notion de *macro-structure*, que nous avons déjà évoquée dans la section 2.1, a été principalement développée par Van Dijk (1977a, 1977b, 1985), et est fortement liée selon cet auteur au problème des topiques de discours, et adopte une approche réellement globale. Elle s'oppose à la notion de *micro-structure*, qui relève de la cohérence locale entre les propositions, phénomène que nous avons déjà évoqué à plusieurs reprise à travers les notions de structure informationnelle et de progression thématique.

La nécessité d'adopter une approche globale découle, comme nous venons de le discuter dans la section précédente, du fait que l'analyse des relations entre énoncés consécutifs ne suffit pas à rendre compte des phénomènes qui régissent l'organisation de fragments textuels plus importants. Par exemple, comme nous l'avons vu dans la section 3.2.2, deux paragraphes consécutifs peuvent être connectés sans que leurs dernière et première phrases respectives soient elles-mêmes connectées par des phénomènes locaux de cohésion et de cohérence (progression thématique, connecteurs, relations sémantiques, etc.). Il est intéressant de remarquer, alors que différents points de vue pourraient sans doute être susceptibles de rendre compte de la cohérence globale du texte, que Van Dijk considère la notion de macro-structure comme intimement liée à la notion de thème dans le sens de l'à propos d'un texte, de ce qui constitue l'information la plus importante ou la plus pertinente :

> We assume that, besides the local semantic structure, a discourse also has a global semantic structure or macrostructure. Thus a macrostructure is a theoretical reconstruction of intuitive notions such as 'topic' or 'theme' of a discourse. It explains what is most relevant, important, or prominent in the semantic information of the discourse as a whole. At the same time, the macrostructure of a discourse defines its global coherence. (Van Dijk, 1985, p. 115).

Van Dijk remarque également que s'il n'existait aucune contrainte d'ordre macroscopique, la cohérence globale du discours ne serait pas garantie dans la mesure où l'on pourrait construire un texte cohésif, et même localement cohérent (i.e. si l'on considère les énoncés deux à deux), qui ne constituerait pas un tout globalement cohérent. Par exemple :

This morning I had a toothache. I went to the dentist.
The dentist has a big car.
The car was bought in New York.
New York has had serious financial troubles.

La séquence d'énoncés qui précède ne constitue pas un discours globalement cohérent, dans la mesure où il ne se réfère pas à une « question » centrale qui unifierait sémantiquement son contenu propositionnel. Selon Van Dijk, c'est l'existence d'une telle « question centrale » (*central issue*) qui procure son unité à un fragment discursif, et qui peut également être qualifiée de *topique*. Comme l'illustre la figure 3.6, nous sommes donc en présence d'une modélisation du discours en deux « couches », la première, dite micro-structure, représentant les relations entre énoncés pris deux à deux, et la seconde, dite macro-structure, représentant les relations de cohérence globales . On remarquera dans cette figure que les liens entres les différents énoncés qui appartiennent à des fragments distincts de la macro-structure sont représentés en pointillés, afin de signifier que ces liens ne sont pas systématiquement présents puisqu'ils ne sont pas indispensables à la cohérence globale.

On aura compris que la notion de macro-structure est de nature purement sémantique : « une macro-structure est définie au niveau de la représentation sémantique du discours : il rend explicite

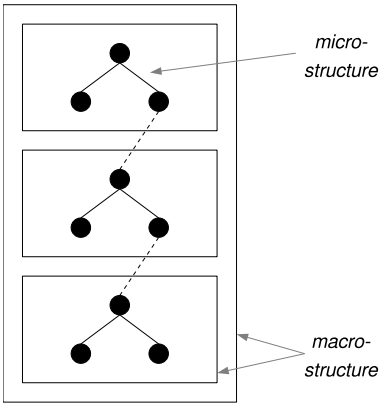


FIG. 3.6 – Micro- et macro-structures du discours.

le sens 'global' d'un discours » (Van Dijk, 1977a, p. 57). Plus précisément, Van Dijk revendique une approche *structuraliste* et *compositionnelle*. Elle est structuraliste dans le sens où il considère d'une part que le discours peut être analysé en tant que séquence de phrases, et d'autre part que les unités de sens qui peuvent être attribuées aux phrases sont des propositions au sens logique, sous une forme prédicative :

> We assume (1) that discourse expressions can be analyzed as sequences of sentences and (2) that the meaning units assigned to sentences are propositions, which consist of a predicate and a number of arguments that may have various (case) roles. (Van Dijk, 1985, p. 105)

L'approche de Van Dijk est également compositionnelle, puisqu'elle considère comme fondamental le principe de compositionnalité[9] qui postule que la valeur sémantique de toute expression complexe est fonction des valeurs sémantiques de ses constituants :

> According to the fundamental principle of semantics, that of functionality, a macro-structure of a discourse should be a function of the respective meanings of its sentences. This function, however, is not given by an added connectivity at the local level of the sequence, that is, the sum of all pairwise coherence links between sentences. Rather it is a kind of semantic transformation, mapping sequences of propositions of the text on sequences of macropropositions at more abstract, general, or global levels of meaning. (Van Dijk, 1985, p. 116)

Sans faire plus d'hypothèses sur la façon dont se construit le sens à l'intérieur des phrases, Van Dijk considère donc que la macro-structure d'un fragment textuel est le produit d'une fonction appliquée aux unités de sens des énoncés qu'il contient, représentées sous forme prédicative. Il faut souligner que cette fonction ne se résume pas à calculer le produit de la juxtaposition de ces unités (leur « somme »), mais implique diverses opérations sémantiques complexes, telles que la sélection, la réduction ou la généralisation. Ces opérations sont dites *macro-règles*, et permettent d'obtenir, pour un ensemble de propositions de la micro-structure, une *macro-proposition* appartenant à la macro-structure.

Les macro-règles sont donc des fonctions sémantiques du deuxième ordre, qui s'appliquent aux prédicats représentant les valeurs sémantiques des énoncés qui composent un fragment de discours. Le résultat obtenu, i.e. une macro-proposition, constitue lui aussi une proposition au sens argumental, qui pourra éventuellement être reformulée si on souhaite rendre compte du « sens global » du fragment de

[9] Van Dijk parle de principe de fonctionnalité (*principle of functionality*), cf. (Van Dijk, 1985, p. 105).

discours qu'elle représente. Par exemple, la reformulation en français de la macro-structure calculée à partir des énoncés du texte suivant pourrait être « Eva a pris le train pour Prague et a commencé un nouveau travail » (Van Dijk, 1977a, p. 56) :

> Eva awoke at five o' clock that morning. Today she had to start with her new job in Prague. She hurriedly took a shower and had some breakfast. The train would leave at 6h15 and she did not want to come late the first day. She was too nervous to read the newspaper in the train. Just before eight the train finally arrived in Prague. The office where she had found the job was only a five minutes walk from the station [...]

Dans cette perspective, la détermination d'un topique de discours ne se réduit pas choix de l'un des topiques des énoncés correspondants, fut-il proéminent (pour le texte précédent il pourrait s'agir de Eva, référent qui apparaît le plus fréquemment en tant que topique local). Au contraire, on cherchera à appliquer différentes règles permettant de retirer des informations jugées non essentielles, de procéder à la généralisation des référent afin d'obtenir des concepts plus généraux, et de grouper les conditions, constituants ou conséquences d'une action ou d'un évènement au sein d'une proposition plus générale (comme « prendre le train »). Il faut remarquer que l'application de ces macro-règles devra le plus souvent prendre en considération des connaissances extérieures au texte, ce qui est probablement l'une des causes les plus importantes du fait que différents interprétants établiront des macro-structures différentes en analysant un même texte.

On notera également que ce fonctionnement est considéré comme potentiellement *récursif* : des macro-règles pourront également opérer sur les macro-propositions pour obtenir des macro-propositions d'ordre supérieur, ce qui produit *in fine* une organisation hiérarchique représentative de la structure thématique du discours. Au sommet de cette structure, on trouvera une proposition représentative du contenu informationnel du discours pris dans son ensemble, qui constitue ainsi son « topique de plus haut niveau », et qui sera souvent exprimée dans son titre[10].

Inversement, si on la considère dans son ensemble, la structure hiérarchique formée par la macro-structure prendra souvent une forme caractéristique de son genre discursif, appelée *super-structure*. En d'autres termes, cette structure assigne un rôle discursif à différentes portions du discours en fonction d'un schéma global « typique ». Il est évident que ce principe ne peut s'appliquer à tous les textes, mais certains genres discursifs s'y prêtent particulièrement bien (avec souvent des conventions propres à un genre ou même à un domaine, et parfois des codifications explicites). Par exemple :

- Concernant les textes narratifs, la super-structure est souvent décrite comme une suite d'épisodes telle que : ouverture, cadre, déclencheur, développement, apogée, suspense final, clôture (Longacre, 1974).
- On est également confronté à des schémas figés dans beaucoup de textes scientifiques, par exemple (Hutchins, 1977) : problème, solutions, implications, où l'énoncé du problème inclura à son tour l'énoncé d'une hypothèse antérieurement formulée, la démonstration de son inexactitude, etc. (cf. figure 2.3 p. 35).
- Le style journalistique fait également apparaître des régularités macro-structurelles. En particulier, comme le remarque Van Dijk (1985, p. 122), ce genre impose que l'information principale apparaisse en première position, avant même que soient spécifiées les causes ou le contexte de l'événement décrit.

[10] Ce phénomène est typique dans le genre journalistique, avec des titres tels que « A tornado kills 500 people » ou « President will meet soviet leader » (Van Dijk, 1985, p. 117).

//siège//	/pour s'assoir/	/rigide/	/pour une pers./	/sur pieds/	/avec dossier/	/avec bras/
'chaise'	+	+	+	+	+	-
'fauteuil'	+	+	+	+	+	+
'tabouret'	+	+	+	+	-	-
'canapé'	+	+	-	+	+	0

FIG. 3.7 – Structure sémique du taxème //siège// (Pottier, 1964).

3.3.2 Le thème en sémantique interprétative

Dans le cadre de la sémantique interprétative (Rastier, 1987), la notion de thème est fondée sur l'étude de divers modes d'interaction en discours de différentes unités sémantiques en partant du sème, l'unité minimale de sens. Dans ce cadre, la notion de thème est envisagée dans une optique plutôt littéraire, et s'oppose frontalement à la tradition lexicographique qui considère qu'un thème est défini par un *mot-vedette*, choisi parmi un ensemble plus ou moins fermé de *termes* (cf. note D.1). Considérant que le sens d'un signe ne peut exister indépendamment de son interprétation en contexte, Rastier propose de relier la sémantique lexicale à la sémantique textuelle, en abordant l'analyse thématique du point de vue de la sémantique componentielle pour rapprocher la notion de thème de celle de *molécule sémique* (Rastier, 1995a). Cette approche caractérisera ainsi un thème par un réseau de cooccurrence sémantique, c'est-à-dire une structure stable de sèmes, matérialisés en discours par des lexicalisations diverses.

Il est bien sûr impossible d'exposer ici les fondements théoriques de la sémantique interprétative, et nous nous contenterons de rappeler les quelques éléments terminologiques dont nous ferons usage dans cette section. La démarche de la sémantique interprétative, qui s'inscrit dans le courant européen de la sémantique structurale et plus précisément de la sémantique componentielle (Pottier, 1964), postule que toute unité sémantique peut se décomposer en *sèmes*, unités minimales se sens. Dans l'optique différentielle qui caractérise la tradition structurale européenne[11], le sème constitue un élément conjoignant ou disjoignant[12] des *sémèmes*, qui représentent le signifié des *morphèmes*. Un morphème est un signe linguistique minimal dont la combinaison permet de former des *lexies* (c'est-à-dire des mots ou des syntagmes). Le signifié d'une lexie est appelé *sémie*.

L'ensemble des sèmes qui caractérisent un sémème déterminera son appartenance à telle ou telle *classe sémantique* (ou paradigme sémantique), ainsi que ses relations avec les autres sémèmes de sa classe. Les sèmes qui déterminent la classe d'un sémème sont dits *génériques*, et forment son *classème*. Les sèmes qui le distinguent des autres sémèmes de sa classe sont dits *spécifiques*, et forment son *sémantème*. Un exemple classique d'une analyse sémique fondée sur ces notions est donnée dans (Pottier, 1964), que nous avons reproduite dans la figure 3.7 en utilisant les notations habituelles (cf. infra). Dans cette figure, différentes instances du taxème //siège// sont distinguées par différents sèmes, le signe « + » ou « - » indiquant l'appartenance de chacun d'entre eux au sémantème de chaque sémème, et le signe « 0 » une indétermination.

Comme dans la figure précédente, les différents objets que nous avons introduits jusqu'ici sont traditionnellement désignés à l'aide de paraphrases, dont la nature précise est indiquée en utilisant diverses notations. En sémantique interprétative, les principales conventions sont les suivantes :

- Un signe est représenté entre guillemets, par exemple « fauteuil ».
- Un signifié (habituellement un sémème ou une sémie) est représenté entre guillemets simples, par exemple 'fauteuil'.
- Un sème (ou la récurrence d'un sème) est représenté entre barres obliques, par exemple /avec

[11] Cf. Greimas, Pottier ou Rastier, par opposition à la sémantique componentielle américaine, à la Katz et Fodor.
[12] Certaines approches attribuent au sème une fonction purement oppositionnelle, cf. (Perlerin, 2004, chap. 2).

dossier/.

– Une classe sémantique est représentée entre barres obliques doubles, par exemple //siège//.

Parmi les sèmes d'un sémème donné, il est important de distinguer les sèmes dits *inhérents* des sèmes dits *afférents*. Les premiers sont caractéristiques d'un sémème-type, c'est-à-dire d'un système considéré *en langue*. En d'autres termes, il s'agit de traits sémantiques valables hors contexte, qui possèdent un caractère définitoire. Par défaut, ces traits sont activés en contexte, à moins qu'ils soient « neutralisés » par un phénomène de *virtualisation*. Les sèmes afférents n'interviennent quant à eux que si lorsqu'on considère un sémème en contexte : ils ne sont actualisés que si le contexte le permet et le réclame. Il faut également distinguer différents types de sèmes génériques, selon le type de classe sémantique qu'ils caractérisent : les sèmes dits *microgénériques* caractérisent des classes minimales dites *taxèmes* (comme //siège//), les sèmes *mésogénériques* des *domaines* (comme //chimie//), et les sèmes *macrogénériques* des *dimensions* (classes très générales comme //animé// ou //inanimé//).

Pour clore ce rapide tour d'horizon des concepts liés à la sémantique interprétative, ajoutons que l'on distingue généralement trois paliers d'analyse sémantique, que sont la *microsémantique*, qui se limite au morphème et à la lexie, la *mésosémantique* qui s'intéresse aux paliers intermédiaires tels que le syntagme, la proposition ou la phrase, et la *macrosémantique* qui décrit tous les blocs supérieurs jusqu'au texte pris dans son ensemble.

Dans le contexte de la sémantique interprétative, le terme de *thème* est utilisé pour désigner « une structure stable de traits sémantiques (ou sèmes), récurrente dans un corpus, et susceptible de lexicalisations diverses » (Rastier, 1995a, p. 227). De la distinction entre sème générique et sème spécifique découlent les notions de *thème générique* et de *thème spécifique*. Un thème générique est caractérisé par la récurrence d'une *isotopie générique*, elle-même induite par la récurrence d'un même sème générique. Un thème spécifique se définit par une *molécule sémique*, c'est-à-dire une structure formée de sèmes spécifiques, mis en relation par des *primitives sémantiques*.

Le phénomène d'isotopie est lié à l'activation récurrente d'un même trait sémantique, soit par la récurrence d'un même morphème, soit, le plus souvent, par différentes lexicalisations permettant d'activer un même sème. Lorsqu'une isotopie concerne des sèmes génériques, elle est elle-même qualifiée de générique, et plus précisément de microgénérique, mésogénérique ou macrogénérique selon que le séme récurrent caractérise un taxème, un domaine ou une dimension (cf. supra). Et lorsqu'une isotopie générique est *dominante*, elle permet de caractériser un thème générique[13], que Rastier rapproche de la notion intuitive de « sujet », c'est-à-dire de la description très générale de ce dont il est question dans un texte.

La formation des isotopies n'est pas seulement liée à la récurrence d'un ou plusieurs traits sémantiques. Un aspect important de la sémantique interprétative concerne les interactions qui peuvent se produire entre les unités de sens. Lors de sa lexicalisation en discours, un trait sémantique dispose d'un certain « potentiel d'activation » se diffusant le long des structures morphosyntaxiques, qui pourront favoriser ou inhiber cette propagation. Cette propagation provoquera, selon les traits sémantiques qui « entourent » un sémème-occurrence, la virtualisation d'un sème inhérent ou l'activation d'un sème afférent. De même, la présence d'un sémème n'active pas seulement les sèmes qui lui sont propres, mais aussi, à un degré moindre, les sèmes qui caractérisent d'autres sémèmes de la même classe sémantique. L'actualisation d'un trait favorise également celle des traits voisins dans la même molécule sémique. Ce phénomène, produit typiquement par des anaphores associatives et dénommé *paratopie*, « justifie sémantiquement l'étude des cooccurrences lexicales[14] dans le domaine de l'analyse thématique » (Rastier, 1995a, p. 232).

La notion de thème spécifique intervient quant à elle à un niveau plus fin, et se traduit par la

[13] On remarquera que dans ce cas le thème n'est pas à proprement parler une structure, contrairement à ce que suppose la définition donnée plus haut.

[14] On trouvera dans (Ferret *et al.*, 1997) le pendant lexicométrique de ce principe, appliqué à la segmentation thématique.

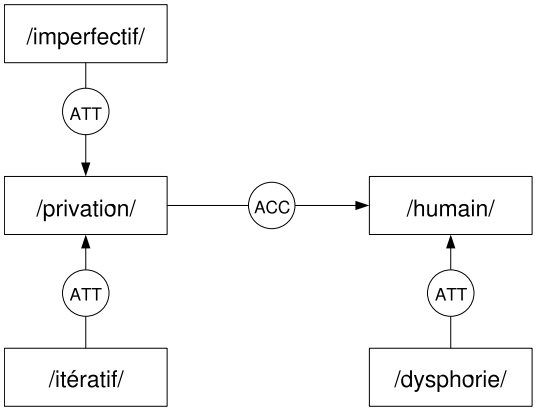


FIG. 3.8 – Molécule sémique du thème de « l'ennui » (Rastier, 1995a).

formation en discours de molécules sémiques, c'est-à-dire de sèmes spécifiques mis en relation par des primitives sémantiques (cas ou relations structurales), formant un réseau que l'on peut représenter par des graphes où les noeuds sont des sèmes et les arcs sont étiquetés par des primitives sémantiques. Par exemple, la figure 3.8 représente le thème de « l'ennui » (la notation « ACC » désigne le cas accusatif, et « ATT » l'attributif). Il est intéressant de constater que si le terme « ennui » est probablement le plus immédiat pour qualifier ce thème, la molécule sémique correspondante pourra émerger d'un texte même s'il n'apparaît jamais, comme dans cet extrait[15] :

> La conversation de Charles était (/imperfectif/) plate (/imperfectif/, /monotonie/) comme un trottoir de rue (/monotonie/), et les idées de tout le monde (/itératif/, /monotonie/) y défilaient (/imperfectif/, /itératif/), dans leur costume ordinaire (/itératif/, /monotonie/), sans (/privation/) exciter d'émotion (/euphorie/), de rire (/euphorie/) ou de rêverie (/euphorie/).
>
> Source : MB

Rastier envisage la relation entre thèmes générique et spécifique sous la forme d'un schéma figure/fond : « les molécules sémiques sont des formes simples, alors que les isotopies génériques sont des fonds sémantiques sur lesquels elles se présentent à la perception » (Rastier, 1995a, p. 231). En d'autres termes, les sèmes génériques qui participent à une isotopie générique viennent « compléter » les sèmes spécifiques qui composent une molécule sémique pour produire le schéma thématique global. Ce mécanisme intervient par exemple dans le cas de la métaphore : une même molécule sémique pourra révéler des sens différents selon la nature du « tissu » thématique sur lequel elle repose.

Pour la sémantique interprétative, l'instauration d'un thème est donc le produit d'une combinaison complexe d'unité sémantiques fines, ce qui nous éloigne significativement de la notion de « référent » que l'on manipule dans d'autres approches. La distance qui sépare les formes de la substance apparaît d'autant plus grande dans cette optique, où la détermination d'un thème ne peut plus se résumer au choix d'un ou plusieurs mots dont on considérera le référent comme thématique ou topical. Mais l'analyse lexicale n'en est pas moins essentielle pour la sémantique interprétative dans la mesure où les morphèmes demeurent le siège des composantes minimales du sens, et constituent en cela des « points d'accès » vers les thèmes qu'ils lexicalisent, en partie ou en totalité. De fait, il est possible de lexicaliser une molécule sémique en utilisant un seul mot, mais on ne considérera en aucun cas qu'il

[15] On notera que s'il n'appartient pas à la molécule proprement dite, le sème /monotonie/ est ici considéré comme la combinaison des sèmes /imperfectif/ et /itératif/. Le sème /euphorie/ appartient quant à lui à une autre molécule, ici évoquée négativement, qui représente le concept de possession d'un objet de valeur (cf. (Rastier, 1995a, p. 229)).

existe un lien déterminé entre ce mot et la molécule. Par exemple, la molécule sémique de la figure 3.8 peut être lexicalisée par « ennui » mais aussi par « dimanche », « monotone » ou « araignée » (Rastier, 1995a, p. 231), mais aucun de ces mots ne constitue une lexicalisation privilégiée de cette molécule, et cette dernière ne constitue pas non plus une représentation arrêtée de leur sens. Rastier caractérise ainsi la relation entre thème et unité lexicale :

- Les thèmes sont accessibles dans le texte par les lexèmes (et non pas définis par eux).
- Un lexème peut lexicaliser plusieurs thèmes comme il peut n'en lexicaliser aucun.
- Un thème peut se lexicaliser par différents lexèmes. Il peut en exister une lexicalisation privilégiée dans un contexte donné (« mot vedette » ou terme), souvent caractérisée par une forme plus synthétique et une plus grande fréquence.

Le mot, à partir duquel commence l'analyse thématique, n'en est donc pas l'objet, et cette analyse pose ainsi la question de la pertinence thématique de chaque lexème : comment savoir si un lexème donne ou non accès à un thème ? Pour passer de l'analyse lexicale à l'analyse thématique, Rastier propose d'exploiter le phénomène de cooccurrences des formes évoqué plus haut. Le terme de *cooccurrent* est utilisé pour désigner les formes pour lesquelles apparaît une certaine proximité physique dans le discours, alors que le terme de *corrélat* est utilisé pour désigner les cooccurrents pour lesquels on a identifié une relation sémantique, en tant que différentes lexicalisations d'une même molécule sémique. Pour accéder au statut de corrélation, la cooccurrence doit se trouver au sein d'un contexte isotopique (marquant un certain degré d'équivalence sémantique par le partage de sèmes), ou paratopique (marquant un certain degré de complémentarité sémantique favorisant l'émergence d'un même thème). Les corrélats sont à leur tour reliés par des relations sémantiques (casuelles ou d'équivalence partielle), pour former un réseau constitutif du thème : le réseau thématique.

Ces propositions sont bien sûr à rapprocher des méthodes d'analyse thématique dites quantitatives, que nous verrons dans le chapitre 4, même si l'étude de la distribution des formes des surfaces n'est que très approximativement révélatrice des phénomènes d'isotopie sous-jacents. On pourra par exemple remarquer une certaine proximité de ces concepts avec la notion de signature thématique telle que développée dans (Ferret et Grau, 2001) : si l'aspect sémantique n'y est pas pris en compte explicitement, les signatures lexico-numériques des thèmes produits par cette approche sont manifestement censées représenter des objets équivalents aux réseaux thématiques de Rastier.

On pourra d'ailleurs remarquer que certaines critiques habituellement formulées à l'égard des méthodes quantitatives semblent partiellement applicables au modèle de la sémantique interprétative. En particulier, quand les premières tendent à considérer un texte comme une séquence non structurée de formes de surface dont on étudie les propriétés distributionnelles, la seconde tend à considérer un texte comme une séquence d'unités sémantiques sans que les différents niveaux de structuration linguistique (syntaxiques ou discursifs) soient explicitement pris en compte. Car même si un observateur humain armé des outils de la sémantique interprétative prendra bien ces structures en compte dans son analyse, il apparaît que la sémantique interprétative ne fournit pas de modèle formel (et donc opérationalisable) des phénomènes qui interviennent à l'interface entre ces structures linguistiques et les structures sémantiques qui résultent de leur interprétation.

De ce fait, il nous semble que la sémantique interprétative seule ne permet pas de rendre compte de toutes les structures discursives qui participent de la structure thématique d'un texte. Il paraît en effet clair que cette dernière ne résulte pas seulement des phénomènes de récurrence et/ou de coocurrence de certaines unités sémantiques, mais aussi de structures d'ordre rhétorique (au sens large). Par exemple, il semble difficile d'ignorer le rôle des connecteurs de discours dans la structure thématique d'un texte, rôle dont le modèle componentiel ne rend pas compte de façon directe. En d'autres termes, si l'on adopte la terminologie proposée par Halliday (cf. section 3.1.1), il semble que les phénomènes décrits par la perspective componentielle seraient plutôt d'ordre idéationnel, alors que les dimensions interpersonnelle et textuelle jouent également un rôle primordial dans l'organisation thématique.

Ainsi, dans l'optique de l'analyse automatique, que l'on étudie la distribution des formes de surface ou celle des unités sémantiques si l'on s'en donne le moyens, il nous semble nécessaire de considérer les indices ainsi récoltés en corrélation avec des indices issus d'autres modèles sémantiques adaptés aux autres phénomènes discursifs. On ne saurait donc considérer le modèle de la sémantique interprétative, même à supposer qu'il soit intégralement opérationalisable, comme capable de rendre compte en toute généralité de la structure thématique d'un texte. Au contraire, il paraît intéressant de chercher à l'utiliser en corrélation avec d'autres modèles, ce que dont nous serons amenés à discuter plus avant.

3.4 Théories et modèles connexes

3.4.1 L'hypothèse de l'encadrement du discours

Présentation générale

Dans (Charolles, 1997), Charolles décrit un mode particulier d'organisation du discours fondé sur les notions d'univers et de cadre de discours. Il montre comment une expression détachée dite *introducteur* peut initier un segment textuel appelé *cadre*. Un tel cadre sera constitué d'un ensemble de propositions dont l'interprétation est contrainte par les critères donnés par l'expression qui l'introduit. Voici un exemple de cadre introduit par une expression temporelle, l'ensemble du segment devant être interprété relativement à la période qu'elle définit :

> De 1965 à 1985, le nombre de collégiens et de lycéens a augmenté de 70%, mais selon des rythmes et avec des intensités différents selon les académies et les départements. Faible dans le Sud-Ouest et le Massif central, modérée en Bretagne et à Paris, l'augmentation a été considérable dans le Centre-Ouest, en Alsace, dans la région Rhône-Alpes et dans les départements de la grande banlieue parisienne où les effectifs ont souvent plus que doublé.
>
> Source : HER

Une partie de l'analyse de Charolles se fonde sur la notion *d'univers de discours*, définie par Martin (1983) comme « l'ensemble des circonstances, souvent spécifiées sous forme d'adverbes de phrase, dans lesquelles la proposition peut être dite vraie ». Un univers de discours correspond donc un ensemble de circonstances attachées à un événement, fait ou procès, et déterminent par là-même les conditions dans lesquelles une proposition peut être dite vraie ou fausse (portée *véridictionnelle*) , et rejoint ainsi ainsi les notions de *framework* chez Chafe ou de *theme* chez Dik (cf. section 3.1). Comme le remarque Martin, ces circonstances sont généralement spécifiées sous la forme d'adverbes dits « de phrase »[16], c'est-à-dire pour Charolles de « groupes syntaxiques périphériques adjoints à la phrase ». Ces syntagmes, qui ont donc pour rôle d'installer des univers, et de régir la mise en relation des univers entre eux, seront ici qualifiés *d'introducteurs*.

Dans le modèle[17] de l'encadrement du discours, une propriété essentielle de ces introducteurs est qu'ils peuvent porter sur plusieurs propositions, et non pas seulement sur celle à laquelle ils sont rattachés. Ainsi, comme on peut le constater dans l'exemple ci-dessous, le ou les critères véridictionnels spécifiés par un introducteur peuvent se « propager » au fil du texte sans avoir à être repris explicitement : selon un « principe général d'attachement à gauche, les propositions arrivantes [prennent] place, sauf indication contraire, dans le dernier univers en cours au moment où elles apparaissent ».

[16] Cf. note D.3.

[17] Bien que le terme d'hypothèse soit semble-t-il préféré par la communauté, nous nous autoriserons parfois à employer le terme de modèle, partant du principe qu'un modèle n'est jamais qu'une hypothèse qui n'a pas (encore) été infirmée par l'expérience.

L'ensemble des propositions soumises à la portée d'un introducteur forment ainsi une unité appelée *cadre de discours* (nous verrons plus loin que cette appellation n'est pas seulement liée aux univers de discours, mais recouvre un phénomène plus général).

Les introducteurs de discours ont donc une fonction double, à la fois *instructionnelle* et *représentationnelle*[18]. La fonction représentationnelle est liée au fait qu'un introducteur d'univers véhicule une instruction véridictionnelle purement interprétative, par opposition à d'autres instructions détachées comme les expressions modales (par ex. « vraisemblablement... » ou « il se peut que ... »), qui véhiculent des instructions computationnelles en imposant un *mode de calcul* de la valeur de vérité d'une proposition. La fonction instructionnelle est quant à elle liée à la fonction *procédurale* et *cognitive* des introducteurs, qui permettent de répartir les contenus propositionnels dans des blocs homogènes, tout en favorisant la mobilisation des connaissances nécessaires à l'interprétation de ces blocs.

Ces différentes fonctions des introducteurs sont liées à une première famille d'opérations que l'on peut distinguer parmi celles décrites par Charolles comme intervenant dans la gestion des univers de discours, et qui président à :

– *l'installation* proprement dite des univers discours ;
– *l'intégration* sous un univers ouvert d'une ou plusieurs propositions.

Une seconde famille d'opérations concerne la mise en relation des univers entre eux, et concernent :

– *la projection* d'univers dits « parents » ;
– *l'unification* d'un nouvel univers avec un univers dit « virtuel » ;
– *la subordination* des univers les uns sous les autres.

Ces différentes opérations sont liées à différents phénomènes discursifs relatives aux constructions formées de plusieurs cadres, que nous allons pouvoir observer dans cet exemple analysé dans (Charolles, 1997) :

> En général, les gens se serrent la main droite quand ils se rencontrent ou se séparent, ou bien ils s'embrassent. Hello, bonjour, namaste ! Chez nous, un baiser est surtout une preuve d'amour et de tendresse à l'égard de quelqu'un de cher, mais chez certains peuples, c'est un salut courant. En Inde, les gens se saluent mains jointes sur la poitrine, comme s'ils priaient. Au Japon, les gens s'inclinent à plusieurs reprises, face à face, en joignant les mains. En France, les hommes faisaient le basemain aux femmes mariées en signe de respect, et les jeunes filles la révérence, mais cette coutume se perd de plus en plus.
>
> Source : SAL

Le premier introducteur rencontré dans cet extrait est « en général », qui ouvre un univers dit *générique*, contrairement aux introducteurs suivants qui ouvrent des univers spatiaux dits *spécifiques*, liés à des zones géographiques spécifiées soit par un déictique (« chez nous », qui se réfère à la situation d'énonciation), soit en se reposant sur les connaissances présumées connues (par ex. « en Inde »). Il faut préciser que l'introducteur « en général » correspond bien dans ce cas à un introducteur d'univers de discours, et non à une expression modale qui ne saurait jouer ce rôle (cf. supra). En effet, à la lecture du passage dans sa globalité, on s'aperçoit que cette expression pourrait ici être remplacée par une expression comme « dans la plupart des pays », et qu'elle introduit bien un univers spatial. En revanche, ce fait n'est pas connu dès le départ, et c'est seulement lors de la rencontre avec le premier introducteur spatial spécifique (« chez nous ») que l'on se rendra compte de sa nature effective. Pour cette raison, l'univers qu'il introduit sera qualifié de générique. Ces considérations sont directement

[18] On peut également qualifier ces fonctions de *textuelle* et *idéationnelle*, à la manière de (Péry-Woodley, 2000b) qui se rapporte ainsi aux méta-fonctions de Halliday (cf. section 3.1.1).

liées à l'approche *incrémentielle* adoptée par Charolles, qui prévoit des possibilités de réanalyse *a posteriori* pouvant conduire à la mise à jour des interprétations préalablement construites.

Ce premier introducteur donne dans un premier temps l'occasion de décrire l'opération de *projection*, qui nous intéressera tout particulièrement par la suite. Selon Charolles, l'installation d'un univers s'accompagne de la projection d'espaces associés, découlant de la relation de contraste qui naît naturellement des connexions entre univers de discours :

> Préciser qu'une proposition p est vérifiée sous une circonstance C, ou, plus généralement, relativement à un certain critère C, donne à entendre, vue que précisément on prend soin de préciser C, qu'elle n'est vraie que sous C, ce qui revient à évoquer toutes les autres circonstances dans lesquelles elles ne serait pas vraie. L'instanciation d'un univers de discours projette donc, par inférence locale, ce que nous appellerons un ensemble d'univers parents (correspondant à l'ensemble des circonstances différentes de C), univers qui demeurent virtuels tant qu'ils ne sont pas instanciés dans le texte. (Charolles, 1997)

La nature des univers ainsi projetés dépend bien sûr de celle de l'univers projetant, les circonstances associées aux univers projetés correspondant à la négation ou au complément de l'univers effectivement installé. S'en suit un phénomène d'attente de la part du lecteur :

> Le contenu de l'expression introductrice sélectionne plus ou moins par avance un ou plusieurs univers qui sont associés à l'univers qu'elle introduit. Il en va ainsi avec "en général" qui appelle un prolongement du type "en particulier". Bien entendu, il se peut parfaitement qu'une séquence introduite sous les auspices d'un "en général" demeure au plan des généralités, mais s'il est effectivement question, dans la suite du texte, de tel ou tel cas particulier, les lecteurs auront le sentiment d'aborder une étape potentiellement inscrite dans le développement du discours (*ibid.*).

L'opération d'*unification* correspond alors à l'instanciation d'un univers parent, c'est-à-dire à la reconnaissance d'un univers correspondant à un univers virtuel précédemment projeté. Dans le cas de l'exemple qui nous occupe ici, cette opération d'unification revient à reconnaître l'univers introduit par « chez nous » comme une instance de l'univers parent spécifique projeté par « en général », et permet du même coup de reconnaître rétrospectivement ce dernier comme un univers spatial.

L'opération de *subordination* apparaît quand un univers de discours inclut d'autres univers plus spécifiques. C'est par exemple le cas de l'univers introduit par « chez nous » relativement à celui introduit par « en général », qui au lieu de provoquer la fermeture de l'univers précédemment ouvert, constitue un univers de niveau inférieur. Le choix entre fermeture et subordination dépend de différents facteurs :

- Dans le cas de « chez nous », le fait qu'il s'agisse d'un univers spécifique immédiatement précédé par un univers générique induit nécessairement une relation de subordination.
- Dans le cas de « chez certains peuples », il apparaît clairement qu'il n'est aucune intersection référentielle entre les peuples en question et celui désigné par « chez nous » (du fait, en partie, de l'usage du connecteur « mais »), et que l'univers introduit par « chez nous » doit donc être fermé. L'univers nouvellement ouvert se trouve alors subordonné à l'univers générique « en général », en parallèle du précédent.
- Dans le cas de « en Inde », il est nécessaire de faire intervenir le contenu propositionnel des segments concernés pour se rendre compte que ce pays ne peut être inclus dans le cadre précédent, c'est à dire qu'il n'appartient pas aux « certains peuples » de l'univers précédent. On conclut alors à la fermeture du cadre précédent pour en ouvrir un nouveau, de même niveau.
- Enfin, dans le cas de « au Japon », ce sont les connaissances d'arrière plan qui indiquent que, s'agissant d'un pays différent, il y a là encore exclusion mutuelle et donc fermeture du cadre précédent.

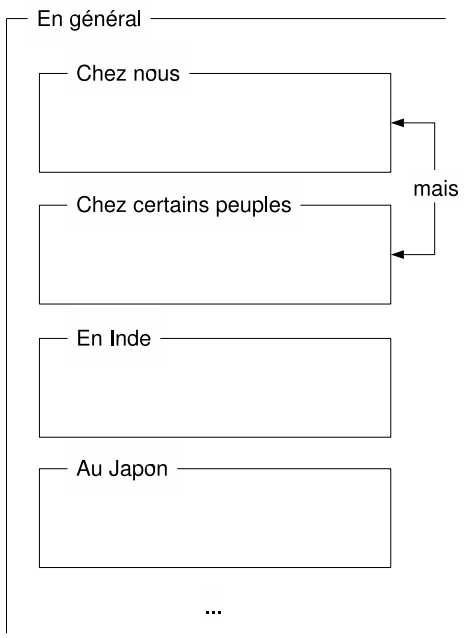


FIG. 3.9 – Analyse partielle des cadres de discours de l'extrait (SAL).

Finalement, l'ensemble de ces cadres spatiaux spécifiques sont subordonnés au premier univers (générique), pour produire une structure reproduite dans la figure 3.9 (on notera que ce schéma de reproduit pas l'analyse du texte dans son intégralité, que l'on pourra trouver dans (Charolles, 1997)).

Comme nous l'avons laissé entendre plus haut, le terme de « cadre de discours » recouvre en fait une certaine variété de constructions, dont les univers de discours ne sont qu'un cas particulier. Charolles définit en effet les quatre types de cadres suivants :

- les univers de discours ;
- les champs thématiques ;
- les domaines qualitatifs ;
- les espaces de discours.

Les *champs thématiques* (ou cadres thématiques sont introduits par des expressions telles que « concernant X », « pour ce qui est de X » ou « quant à X » (constructions qualifiées de *as-for constructs* en anglais, par exemple chez Lambrecht, cf. section 3.1). Par exemple :

> Les deux filles de Robert ne sont pas du tout dans la même situation. Concernant Louise, elle a une bonne situation et son père [...]

Selon Charolles, les expressions détachées de ce type se distinguent des introducteurs à portée veridictionnelle que nous avons rencontrés jusqu'ici par le fait que leur fonction principale n'est plus de spécifier des critères de vérité valant pour les contenus propositionnels d'un segment, mais plutôt d'introduire ce qui sera le thème (au sens de l'à propos) de ce segment :

> Ce que marquent au premier chef des formules comme "à propos de X", "au sujet de X", "concernant X", "pour ce qui est de X" c'est la volonté du locuteur de signaler que, au moins pour un temps, ce qu'il va dire porte sur X (et non sur Y ou Z), a pour objet X, bref, que le thème de son propos va être X. (Charolles, 1997)

Nous serons amenés à commenter ce point de vue dans le chapitre 8, où nous observerons des exemples où un introducteur du type « concernant X » n'a pas pour fonction de spécifier le thème *stricto-sensu* du segment qu'il introduit, mais se rapprocherait plutôt de la notion d'univers de discours dans un sens légèrement élargi. De notre point de vue, ce cas se présente lorsque, contrairement à l'exemple ci-dessus, le référent de l'introducteur ne fait pas l'objet d'une reprise immédiate (du type « concernant X, il... »), ce qui semble d'ailleurs mieux accepté à l'écrit.

Les *domaines qualitatifs* sont quant à eux introduits par des expressions qui renseignent sur les aspects qualitatifs des états de choses dénotés, tels que le but ou les motivations d'un participant à un procès. En voici un exemple :

> Pour faire plaisir à sa belle-mère, Paul mit une cravate. Il fit la vaisselle puis rangea son bureau. [...] Pour taquiner son beau-père, [...]

Enfin, les *espaces de discours* sont introduits par des expression portant sur les aspects méta-linguistiques de l'énonciation (comme « en un mot » ou « finalement »), ou par des marqueurs de l'organisation du discours (comme « d'une part » / « d'autre part »). Dans ce dernier cas, on parle également de cadres *organisationnels*, dont voici un exemple (tiré de (Jackiewicz et Minel, 2003)) :

> La dramatique situation du Mexique montre aujourd'hui que le problème de la dette demeure entier. Avec cette circonstance aggravante que le dispositif imaginé entre 1982 et 1984 arrive à expiration. Pour trois raisons :
>
> - tout d'abord, la récession envisagée comme méthode universelle pour ramener les compteurs d'un pays à zéro n'est plus acceptée ni acceptable par les pays du tiers-monde qui n'en voient pas la fin ;
>
> - de plus, l'approche "au cas par cas", telle que le FMI l'a conçue et pratiquée, a également tourné à l'échec ;
>
> - enfin, troisième et dernier point, les rééchelonnements pluriannuels sont devenus inopérants.
>
> Source : LMD

Liens avec les notions de topique et d'à propos

Le lien entre les adverbiaux cadratifs et la notion de topique ou de thème est un problème très complexe, dont nous ne pourrons explorer ici tous les méandres. Nous allons simplement faire état de quelques conclusions tirées de (Charolles, 2003), où cette question est amplement discutée. Charolles rappelle tout d'abord que « les expressions peu intégrées syntaxiquement figurant en zone préverbale sont plus ou moins destinées à fonctionner comme des thèmes ou des topiques » (p. 11). En effet, au delà du fait qu'en français (comme dans d'autres langues) la position initiale soit préférentiellement thématique, différents mécanismes de topicalisation (plus courants à l'oral) sont liés au détachement à gauche, accompagné ou non d'une reprise :

> Ma mère, je lui ai tout dit.
>
> L'opéra, je n'y connais rien.
>
> L'opéra, je suis incompétent.

De fait, ces constructions sont assez bien décrites dans le cadre de la structure informationnelle de la phrase (par exemple chez Lambrecht, cf. section 3.1). En revanche, même au niveau phrastique,

le statut topical des syntagmes adverbiaux détachés à gauche (qui sont plus courants à l'écrit et nous intéressent en cela plus particulièrement), est moins net. En voici un exemple, s'ajoutant aux différents introducteurs d'univers que nous avons déjà rencontrés :

> Dans le Calvados, la moitié des emplois du secteur des TIC correspond à une activité industrielle.

Nous avons déjà évoqué dans la section 3.1 différentes approches de ce cas précis : chez certains auteurs, ces topiques particuliers sont appelés *scene-setting topics* (Lambrecht) ou *Chinese-style topics* (Chafe), et n'excluent pas l'existence d'autres topiques dans la même phrase (cf. section 3.1.2). Chez Halliday, ils seront considérés comme thèmes en tant que premières expressions à portée idéationnelle, cette fois à l'exclusion des constituants qui le suivent (cf. section 3.1.1).

Face à ce problème, Charolles envisage quant à lui différents points de vue. Dans un premier temps, il considère ces expressions en tant que *topiques « par l'amont »*, c'est-à-dire relativement aux éléments antérieurs du discours. Il s'appuie pour cela sur (Haiman, 1978), qui discute du statut topical des expressions conditionnelles, et qui privilégie la conception du topique dans laquelle celui-ci se réfère à une information « donnée », « connue » ou « ancienne ». Dans cette perspective, l'accès d'un référent au statut de topique dépend directement de son rôle au sein du discours qui précède (par exemple en termes de degré d'activation chez Lambrecht, cf. section 3.1), où il doit nécessairement avoir été mentionné, ne serait-ce qu'indirectement, ou dont il doit être inférable.

Dans le cas des expressions conditionnelles détachées dont traite Haiman, le lien avec le discours qui précède est effectivement quasi-systématique, puisqu'il s'agit de subordonnées adverbiales ayant un sujet propre, et qui s'insèrent donc aisément dans une chaîne référentielle :

> [...] Sa capacité à survivre dépend de la durée de l'occlusion. Si celle-ci se prolonge, la pénombre évolue lentement vers une lésion totale.

Charolles remarque que des situations analogues peuvent être observées dans le cas des adverbiaux non phrastiques détachés en initiale. Il en va par exemple ainsi dans l'extrait suivant, où la seconde phrase constitue une sorte d'amorce pour l'adverbial détaché de la phrase suivante :

> Qu'est-ce que l'écriture ? L'usage a consacré différentes acceptions du mot. Dans une acception élargie, il peut désigner [...]

Mais cela n'est pas toujours le cas. Il suffit pour s'en convaincre d'observer les nombreux cas où un syntagme adverbial détaché apparaît en initiale d'un texte (même si cela peut également se produire dans d'autres circonstances), puisque dans ce cas cet adverbial ne peut évidemment pas faire référence à ce qui précède. Cette situation est relativement fréquente, comme le montre l'étude de Virtanen (1990) (sur l'anglais), qui relève que 34% des adverbiaux temporels figurent en tête de phrase, dont 60% en initiale de texte. En voici un exemple :

> Ø En 1991, à la Station INRA de Dijon, Patrick Étiévant et Bruno Martin commençaient l'analyse du vin jaune, produit seulement dans le Jura. Le goût spécifique de ces vins résulte de leur technique d'élevage : [...]
>
> Source : LR

Ce dernier fait conduit Charolles à adopter une conception assez large de la notion de familiarité qui fonde la dualité donné-nouveau :

> Le fait que les adverbiaux détachés en position initiale s'appuient sur des informations déjà évoquées dans le discours antérieur explique qu'on les considère comme des topiques. Cette idée d'appui n'est cependant pas facile à préciser. [...] Il semble préférable de partir de l'idée que les adverbiaux sélectionnent du fait de leur signification un attribut (temps, lieu, condition, cause, moyen, manière) des états de choses pour indexer la proposition en tête de laquelle ils apparaissent. Lorsque le critère sélectionné est générique, comme le temps et le lieu qui sont des traits inhérents à tous les états de choses, la décision consistant à l'exploiter comme un index n'a pas besoin d'être justifiée par le contexte précédent ce qui fait qu'ils peuvent apparaître à l'initiale de discours, surtout si le type de texte s'y prête. Lorsque le critère retenu pour indexer la proposition qui suit est moins prédictible, lorsqu'il met en avant un attribut particulier (comme le fait que telle action a été accomplie de telle manière, dans tel but...), on s'attend par contre à ce que son choix soit motivé par le contexte. (Charolles, 2003, p. 25)

Par la suite, c'est cette approche que nous adopterons à notre tour, car elle semble convenir assez bien à la problématique de la recherche d'information. Nous adopterons toutefois une approche plus large de ce que Charolles qualifie de critères génériques : si le temps et l'espace constituent de fait des traits inévitablement associés à tout phénomène, fait ou procès, nous montrerons dans la section 8.4 que d'autres traits peuvent intervenir sans justification dans le contexte précédent pour peu qu'ils jouent un rôle particulier dans les connaissances du domaine auquel se réfère un texte. De ce fait, on pourra très bien trouver en initiale d'un texte un adverbial comme « dans le secondaire », pour peu que le texte soit rattaché *a priori* au domaine scolaire (par exemple par son titre, ou le contexte extra-linguistique). En voici un exemple :

θ **Coopération secondaire**

§ Dans le secondaire, la politique du BCF s'organise autour de trois axes : la formation continue, la promotion des Technologies de l'Information et de la Communication (TIC ou TICE), et la promotion du français. [...]

Source : AFT

Le deuxième point de vue envisagé par Charolles nous invite à considérer les adverbiaux détachés comme *topiques « pour l'aval »*. Partant de la caractérisation des topiques proposée par (Jacobs, 2001) avant de s'en détacher partiellement, il formule les conclusions suivantes. Premièrement, un adverbial détaché en tête de phrase possède le caractère cadratif propre aux topiques, dans le sens où il est à même, comme nous l'avons vu plus haut, de porter sur une série de propositions formant un bloc homogène relativement au critère qu'il spécifie. Selon les termes de Charolles, « il fonctionne comme une sorte d'index que le lecteur ou l'auditeur doivent garder en mémoire pour le traitement de la phrase hôte de l'adverbial, jusqu'à l'occurrence d'indices signalant que sa portée est terminée ».

En revanche, Charolles remarque que les adverbiaux ne sont généralement pas « adressés » au sens de Jacobs, c'est-à-dire que la fonction qui consiste à « ouvrir un fichier destiné à collationner les propositions rassemblées dans un cadre » n'est pas nécessairement liée à la fonction qui consiste, selon les termes de Jacobs, à « marquer l'emplacement au sein des connaissances de l'allocutaire où l'information véhiculée doit être stockée au moment de l'énonciation ». En effet, cette fonction est liée selon Jacobs à la notion d'à propos, que Charolles ne reconnaît pas systématiquement aux adverbiaux détachés. Pour expliciter ce point, il donne ces deux exemples :

(1) Lola sortit faire un tour. En bas de l'immeuble, un homme faisait les cent pas. Il l'aborda pour lui demander si les taxis étaient en grève.

(2) Lola sortit faire un tour. En bas de l'immeuble, un homme faisait les cent pas. Une fine couche de givre recouvrait le sol. Des enfants rentraient de l'école en se chamaillant.

Dans le premier cas, il est clair que le circonstant locatif « en bas de l'immeuble » ne peut être considéré comme l'à propos du discours qui suit, du moins pas à lui seul, puisque c'est « un homme » qui est repris. En revanche, dans le cas (2), le discours porte bien sur ce qui se passe en bas de l'immeuble, et le circonstant locatif a donc bien une portée thématique au sens de l'à propos.

Sans aucunement remettre en cause cette distinction, nous préférerons par la suite considérer, dans le cas (1), que « un homme » constitue bien le thème « principal », que nous qualifierons de « noyau thématique », mais que le circonstant locatif associé possède lui aussi une certaine responsabilité quant à la définion de l'à propos. Ainsi, nous suivrons Le Goffic (1994), que nous avons déjà cité dans la section 3.1.2, et considérerons que le thème dans (1) est bien « un homme - en bas de l'immeuble ».

Terminons avec la troisième conclusion de Charolles, qui nous paraît particulièrement importante. Nous avons vu dans les exemples précédents que, indépendamment du statut thématique que l'on voudra bien accorder au circonstant spatial dans (1), la distinction entre les deux cas implique de prendre en compte la façon dont se déroule la suite du discours. En effet, on ne peut décider du statut thématique de « un homme » sans étudier le schéma référentiel en jeu dans les propositions qui suivent. On pourra au mieux lui attribuer un statut « par défaut », par exemple en considérant qu'en tant que sujet, il est le constituant qui possède le plus de chances d'accéder au statut de topique.

> Il n'est pas possible de statuer sur le topique (*aboutness*) d'une phrase isolée. Par défaut de contexte ultérieur, les adverbiaux détachés en tête de phrase, n'indiquent pas ce à propos de quoi est la phrase. Mais rien n'empêche que la suite oblige à leur restituer ce statut. Ce constat, s'il est bien fondé, milite en faveur d'approches dynamiques des phénomènes de topicalisation et conduit à réintégrer dans la discussion de ces phénomènes la notion de topique de discours. (Charolles, 2003, p. 47)

Pour conclure (temporairement) sur la notion de cadre de discours, remarquons que le modèle de l'encadrement du discours connaît depuis son introduction un succès certain auprès de la communauté du TAL, engouement que l'on peut expliquer facilement. D'une part, même s'il soulève des questions très complexes quand au phénomène de portée, le modèle est assez directement opérationalisable, s'appuie sur une notion d'introducteur qui se prête assez bien à l'analyse automatique, et produit des structures souvent hiérarchiques qui, en plus de décrire assez efficacement la structure de textes techniques, informatifs, expositifs, etc., se rapproche de structures de données omniprésentes en informatique. D'autre part, les phénomènes décrits par le modèle de l'encadrement en terme d'indexation et de répartition de l'information font immédiatement écho aux problématiques de la recherche d'information ou de résumé automatique.

De ce fait, différents travaux ont été menés avec pour objectif de procéder à l'analyse automatique de structures cadratives dans les textes. On peut notamment citer (Mourad et Schrepfer-André, 2002) au sujet des cadres énonciatifs, (Jackiewicz et Minel, 2003) au sujet des cadres organisationnels, ou encore (Ferret *et al.*, 2001b) ou l'analyse quantitative est utilisée pour clôturer des cadres thématiques. Pour notre part, nous avons proposé dans (Bilhaut *et al.*, 2003b) un système dédié à l'analyse des cadres spatio-temporels que nous décrirons dans le chapitre 7, généralisé sous la forme d'un système d'analyse thématique tirant partie de différents types de cadres véridictionnels (Bilhaut et Enjalbert, 2005a), que nous détaillerons dans le chapitre 8.

	$C_r(E_i) = C_r(E_{i-1})$ ou $C_r(E_{i-1}) = \varnothing$	$C_r(E_i) \neq C_r(E_{i-1})$
$C_r(E_i) = C_p(E_i)$	Continuation	Dépl. doux
$C_r(E_i) \neq C_p(E_i)$	Rétention	Dépl. brutal

FIG. 3.10 – Types de transition entre énoncés dans la théorie du centrage (Walker *et al.*, 1998).

3.4.2 Théorie du centrage d'attention (CT)

Evoquée informellement depuis les années 70, la théorie du centrage d'attention (*Centering Theory* ou CT) fut initialement formalisée dans (Grosz *et al.*, 1995), avant d'être plus largement développée dans (Walker *et al.*, 1998) et bien d'autres travaux. Elle s'inscrit au sein de la théorie de la structure du discours développée par Grosz et Sidner dans (Grosz et Sidner, 1986).

Alors que le modèle dit GST (pour *Grosz and Sidner Theory*) se concentre plutôt sur l'organisation globale du discours, représentée par la structure intentionnelle inter-propositionnelle (cf. section 3.4.3), la théorie du centrage s'attache à décrire la cohésion locale, produite par les relations entre les énoncés constitutifs d'un même segment. Selon ce modèle, la cohésion locale dépend du choix des expressions référentielles, qui détermine la charge inférentielle nécessaire à leur résolution dans un état attentionnel donné. On considérera par exemple que des expressions référentielles dont la résolution est trop complexe produira une diminution de l'effet de cohésion locale du discours.

Dans cette théorie, le *centre* désigne les entités d'une expression qui la relient aux autres expressions du même segment. Les auteurs soulignent le fait qu'un centre est associé à un énoncé et non à une phrase, et qu'il s'agit d'un objet sémantique, et non d'un objet textuel. Pour chaque énoncé d'un segment, on définit un ensemble de centres *anticipateurs* C_a et un centre *rétroactif* C_r. Le centre rétroactif d'un énoncé E_n est connecté à l'un des centres anticipateurs de l'énoncé E_{n-1}.

L'ensemble des centres anticipateurs d'un énoncé est ordonné pour refléter leur proéminence relative au sein de cet énoncé. Plus un élément est saillant au sein d'un énoncé E_n, plus il est probable qu'il soit relié au centre rétroactif de l'énoncé suivant E_{n+1}. L'élément le plus saillant est dit *centre préféré* et noté C_p.

A partir des différents types de relations entre les centres rétroactifs, on peut définir différents types de transition entre énoncés d'un segment, présentés dans le tableau 3.10.

Cette caractérisation des relations entre les énoncés du discours apporte des indices intéressants dans le cadre de la structuration du discours, en particulier sur le plan de la délimitation d'espaces textuels cohérents. Différentes réserves ont bien sûr été formulées au sujet de ce modèle, en particulier sur son approche purement linéaire des relations entre énoncés (cf. (Rastier, 1995b) ou (Ho Dac, 2000)), négligeant une structure plus globale du discours où des relations peuvent exister entre des énoncés non adjacents, souvent de nature sémantique ou pragmatique.

Ce modèle n'a toutefois pas cessé, depuis son apparition, de susciter un intérêt assez vif dans certaines communautés, et de nombreuses extensions du modèle initial ont été proposées. Parmi ces contributions, nous nous contenterons ici de citer Strube et Hahn (1996, 1999), qui proposent une approche fonctionnelle du centrage, et introduisent par ailleurs la notion de *coût* dont nous aurons plus tard l'occasion de nous servir.

Cette notion s'appuie sur l'idée que, étant donnés les différents types de transition définis par la théorie du centrage (cf. figure 3.10), certaines *séquences* de transitions sont plus « naturelles » que d'autres, ou plus exactement que leur traitement implique un moindre coût cognitif. Cette idée est elle-même liée au fait que, par définition, le centre préféré d'un énoncé « représente une prédiction au sujet du centre rétroactif de l'énoncé suivant » (Walker *et al.*, 1998). L'hypothèse formulée par Strube et Hahn est que lorsque cette prédiction n'est pas vérifiée, le traitement des transitions est cognitivement

	Continuation	Rétention	Dépl. doux	Dépl. brutal
∅	faible	élevé	n.a.	n.a.
Continuation	faible	faible	élevé	élevé
Rétention	élevé	élevé	faible	élevé
Depl. doux	faible	élevé	élevé	élevé
Depl. brutal	élevé	élevé	faible	élevé

FIG. 3.11 – Coût cognitif des différentes paires de transitions (Strube et Hahn, 1996).

plus coûteux.

Exprimée dans les termes de la théorie du centrage, cette hypothèse consiste à considérer qu'une paire de transitions est *peu coûteuse* lorsque le centre rétroactif d'un énoncé est correctement prédit par le centre préféré de l'énoncé précédent, c'est-à-dire si $C_r(E_{i+1}) = C_p(E_i)$. Inversement, une paire de transition sera considérée comme *coûteuse* quant $C_r(E_{i+1}) \neq C_p(E_i)$. Le résultat de l'application de ce principe aux différentes paires de transitions possibles est représenté sous forme de tableaux dans la figure 3.11. Par suite, les auteurs font état de statistiques effectuées sur l'allemand, qui tendent à montrer que les transitions peu coûteuses au sens ainsi défini sont effectivement beaucoup plus fréquentes que les transitions coûteuses, ce qui laisse à penser qu'il s'agit d'un critère pertinent, ou tout au moins discriminant.

Notons enfin qu'un certain nombre de contributions à la théorie du centrage visent à étudier les liens qu'elle entretient avec la notion de *topique*. Il est en effet intéressant de remarquer qu'aucun rapprochement n'est fait explicitement avec cette notion dans la formulation originale de la théorie, bien que ce rapprochement puisse paraître assez naturel, comme le remarque Péry-Woodley au sujet de la notion de centre rétroactif :

> Cette entité jouit d'un statut spécial qui lui est conféré par deux propriétés : c'est l'entité la plus centralement concernée par [l'énoncé dans lequel elle est évoquée], et elle fait le lien entre [cet énoncé] 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"*). (Péry-Woodley, 2000a, p. 65)

Cette voie a donc été explorée dans différents travaux, parmi lesquels comptent les approches fonctionnelles de (Rambow, 1994) et de (Strube et Hahn, 1996), qui visent à substituer, aux critères grammaticaux habituellement utilisés pour ordonner les centres anticipateurs, des critères liés à la structure informationnelle de l'énoncé, ainsi que les propositions de (Péry-Woodley, 2000a) où la notion de topique est envisagée comme pivot entre les modèles du centrage et de l'encadrement du discours. Pour notre part, nous utiliserons très ponctuellement la théorie du centrage augmentée de la notion de coût décrite plus haut, pour tenter de rendre compte d'un phénomène de quasi-détachement que nous rattachons à la notion de thème discursif (cf. section 8.2).

3.4.3 Théorie de la structure intentionnelle (GST)

Cette théorie, dite GST, pour *Grosz and Sidner Theory* (Grosz et Sidner, 1986), s'appuie principalement sur la description des processus qui sous-tendent la création ou l'interprétation du texte, eux-mêmes guidés par les intentions du locuteur, ou par celles que lui prête le lecteur. Le discours est ainsi scindé en segments, chacun d'entre eux étant doté d'un « objectif communicationnel » (DSP, *Discourse Segment Purpose*). Ces segments s'articulent selon trois structures inter-dépendantes. La *structure linguistique* divise le discours en segments. La *structure intentionnelle* est formée par les relations existant entre les rôles intentionnels attribués à chaque segment. Enfin, *l'état attentionnel*,

représentation abstraite du centre d'attention des acteurs du discours, est caractérisé par les objets et les relations saillantes à un point donné du discours. Au sein de la structure intentionnelle, on trouve la relation de dominance, liant un segment à un autre par le fait que la satisfaction de son objectif intentionnel participe à celle du segment dominant. D'autre part, la relation de « satisfaction-précédence » existe entre deux segments quand la satisfaction de l'un doit impérativement précéder celle de l'autre.

L'exemple suivant reproduit un texte tiré de (Moser et Moore, 1996), composé de trois phrases numérotées P_i :

> **(P1)** Come and see the L.A. Chamber Ballet's concert. **(P2)** The show should be very entertaining. **(P3)** It presents an all new choregraphy.

En définissant les relations intentionnelles $Intend_X\,(Y)$ et $Believe_X\,(Y)$ respectivement comme « X à l'intention que Y » et « X croit que Y », et en notant L le locuteur et A l'auditeur, la structure intentionnelle correspondante serait composée des relations de dominance suivantes :

- $I_1 : Intend_L\,(Intend_A\,(P_1))$
- $I_2 : Intend_L\,(Believe_A\,(P_2))$
- $I_3 : Intend_L\,(Believe_A\,(P_3))$

De ces relations dépend la structure linguistique associée, qui serait ici composée de trois segments DS_i imbriqués :

- $DS_1 = P_1 \cup DS_2$
- $DS_2 = P_2 \cup DS_3$
- $DS_3 = P_3$

On remarquera que ce type de structure ne peut manifestement pas rendre compte de l'organisation du discours en toute généralité, et qu'elle doit être considérée comme une dimension du discours parmi d'autres. Dans cette optique, on pourra par exemple prendre en considération les rapports entre structure intentionnelle et structure rhétorique (cf. note D.2 p. 270).

3.4.4 Théorie de la structure rhétorique (RST)

La théorie de la structure rhétorique (RST) a été initialement développée dans le cadre de la génération automatique de texte, avec pour objectif de se doter d'un modèle formel et opérationalisable de la structure du discours (Mann et Thompson, 1987). Depuis, cette théorie s'est développée au delà de ce contexte particulier, et est également utilisée en analyse de textes, tant pour la description linguistique que pour l'analyse automatique du discours.

Le terme de « rhétorique » dans ce contexte est relativement éloigné de la rhétorique dite classique ou ancienne, qui recouvre la science (ou l'art) de l'expression éloquente, et se consacre essentiellement à l'étude des diverses figures de style susceptibles d'y concourir. Ici, il s'agit plutôt de rendre compte de la *cohérence* des textes de façon purement objective : chaque constituant du discours (proposition ou segment d'ordre supérieur) se voit assigner un rôle relativement aux constituants qui le suivent ou le précèdent. Ce rôle correspond à sa *fonction* dans le processus de formation d'une suite cohérente de propositions, c'est-à-dire logique et ininterrompue, qui caractérise en général un texte « bien construit ».

C'est donc la notion de *relation* qui joue le rôle le plus central dans la RST, qui définit ainsi une typologie des rôles qu'un segment est susceptible de jouer relativement à un segment adjascent, de façon symétrique ou non. Une première catégorie de relations regroupe les cas où ce rôle est asymétrique, c'est-à-dire qu'il existe un rapport de subordination ou de rection entre un constituant et un autre. Dans ce cas, le constituant le plus « essentiel » sera qualifié de *noyau*, et celui qui dépend de lui sera qualifié de *satellite*. Par exemple, la relation de « démonstration » est définie comme liant

un noyau constitué d'une affirmation quelconque à un satellite apportant une information destinée à accroître la croyance du lecteur quant à cette affirmation, ce qui revient à considérer que l'élément « central » ou « essentiel » de cette relation est l'information démontrée et non la démonstration en tant que telle. Chaque relation sera plus précisément définie par les conditions à remplir par son noyau, son satellite, les relations entre eux, ainsi que l'effet attendu sur le lecteur. Il faut remarquer que l'ordre des segments ne fait pas partie de cette définition, même si, pour chaque relation, un ordre particulier sera plus vraisemblable ou plus fréquent. En voici un exemple de définition, concernant la relation de « concession » (on trouvera dans la figure 3.12 une définition simplifiée des différentes relations noyau-satellites du jeu défini par la RST « classique ») :

Nom de relation : Concession

Contraintes sur le noyau : Le scripteur (W) a un avis positif sur la situation présentée dans le noyau (N).

Contraintes sur le satellite : W ne prétend pas que la situation présentée dans le satellite (S) n'existe pas.

Contraintes sur la relation : W reconnaît une incompatibilité apparente entre les situations présentées dans N et S ; W considère les situations présentées dans N et S comme compatibles ; le fait de reconnaître la compatibilité entre les situations présentées dans N et S augmente le crédit attribué par le lecteur (R) à la situation présentée par N.

Effet attendu : Le crédit attribué par le lecteur (R) à la situation présentée par N augmente.

Quand aucun lien de dépendance ne peut être établi entre les constituants d'une relation, parce qu'aucun d'entre eux ne peut être considéré comme plus « central » que les autres, celle-ci est dite *multi-nucléaire*. Un exemple d'une telle relation est la relation de « contraste », qui est définie comme liant les différentes « possibilités dans une alternative ». Ces différentes possibilités pouvant être considérées comme équivalentes, elles constituent chacune un noyau. Les différentes relations multi-nucléaires définies par la RST « classique » sont représentées dans la figure 3.13.

En analysant les instances de ces différentes relations dans les textes, on peut en produire une représentation arborescente, qui reflétera leur structure rhétorique. Ces structures seront généralement représentées sous forme graphique, en utilisant la symbolique introduite par Mann et Thompson devenue conventionnelle. Par exemple, une analyse possible du texte suivant (résumé d'un article scientifique) est donnée dans la figure 3.14 :

> θ **(1) Lactose and Lactase**
>
> **(2)** Lactose is milk sugar ; **(3)** the enzyme lactase breaks it down. **(4)** For want of lactase most adults cannot digest milk. **(5)** In populations that drink milk the adults have more lactase, perhaps through natural selection.
>
> **(6)** Norman Kretchmer, Scientific American, page 70, October 1972.
>
> Source : SA

On notera que si la RST est maintenant utilisée, outre sa perspective originelle de génération automatique, comme cadre d'analyse du discours, et le problème de l'automatisation de cette analyse a également été envisagé. Par exemple, (Marcu, 1997) propose un algorithme d'analyse rhétorique se basant à la fois sur le modèle RST formalisé en logique du premier ordre et sur une analyse de surface utilisant des marques et indices linguistiques structurants.

Il est important de mentionner que la RST ne considère pas comme fermé l'ensemble des relations utilisées pour mener à bien une analyse, et prévoit au contraire la possibilité de définir de nouvelles relations selon les besoins. De ce fait, le cadre défini par cette théorie est extrêmement large, ce qui

Nom de la relation	**Noyau**	**Satellite**
Anti-condition	Action ou situation dont l'occurrence résulte de la non-occurrence de la situation conditionnante.	La situation conditionnante.
Antithèse	Idées approuvées par l'auteur.	Idées rejetées par l'auteur.
Arrière-plan	Texte dont la compréhension est facilitée.	Texte servant à faciliter la compréhension.
But	Une situation cible.	L'intention sous-jacente à la situation.
Cause délibérée	Une situation.	Une autre situation, cause de la première, du fait de l'action délibérée de quelqu'un.
Cause non délibérée	Une situation.	Une autre situation, ayant provoqué la première, mais pas du fait d'une action délibérée.
Circonstance	Texte exprimant les événements ou idées situés dans le cadre interprétatif .	Un cadre interprétatif temporel ou situationnel.
Concession	Situation défendue par l'auteur.	Situation apparemment incompatible, mais également défendue par l'auteur.
Condition	Action ou situation dont l'occurrence résulte de l'occurrence de la situation conditionnante.	La situation conditionnante.
Démonstration	Une affirmation.	Information destinée à accroître la croyance du Lecteur relative à l'affirmation.
Élaboration	Information de base.	Information supplémentaire.
Évaluation	Une situation.	Un commentaire évaluatif de la situation.
Facilitation	Une action.	Information destinée à aider le Lecteur à accomplir cette action.
Interprétation	Une situation.	Une interprétation de la situation.
Justification	Un texte.	Information légitimant l'énonciation du texte par l'auteur.
Motivation	Une action.	Information destinée à accroître chez le Lecteur le désir d'accomplir l'action.
Reformulation	Une situation.	Une reformulation de la situation.
Résultat délibéré	Une situation.	Une autre situation, causée par celle-ci, du fait de l'action délibérée de quelqu'un.
Résultat non-délibéré	Une situation.	Une autre situation provoquée par la première, mais pas du fait d'une action délibérée.
Résumé	Un texte.	Un court résumé de ce texte.
Solution	Une situation ou un procédé apportant une satisfaction complète ou partielle au besoin.	Un problème, une question, ou tout autre besoin exprimé.

FIG. 3.12 – Relations noyau-satellite de la RST « classique » (Mann et Thompson, 1987). Les traductions ici mentionnées sont celles proposées sur le site « officiel » de la RST, et qui semblent communément utilisées en dépit de leur caractère parfois approximatif.

Nom de la relation	Un segment	Un autre segment
Contraste	Une possibilité dans une alternative.	L'autre possibilité.
Jonction	*non contraint*	*non-contraint*
Liste	Un item.	L'item suivant.
Séquence	Un item.	L'item suivant.

FIG. 3.13 – Relations multi-nucléaires de la RST « classique » (Mann et Thompson, 1987).

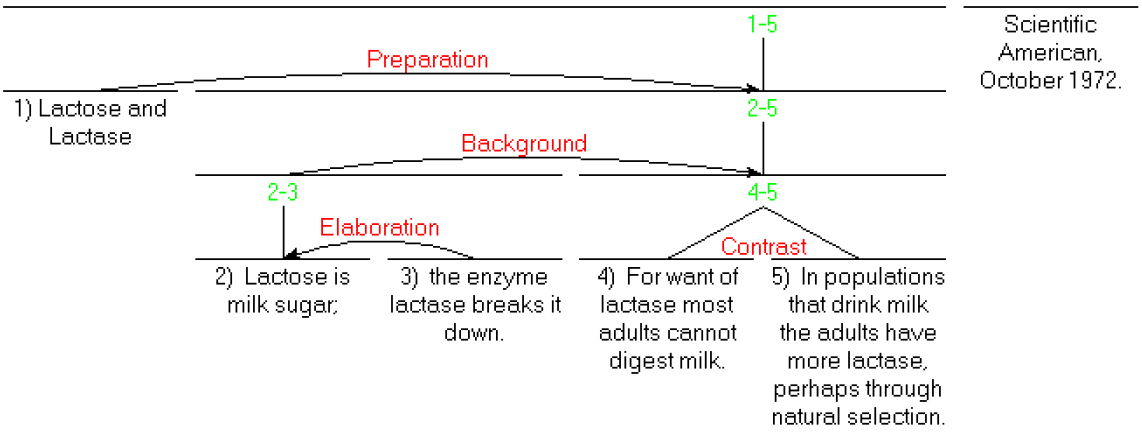


FIG. 3.14 – Exemple d'analyse RST (Lactose and Lactase).

est en soi un atout. Il faut toutefois remarquer que, une fois détachée de tout catalogue de relations particulier, elle ne s'appuie plus que sur des notions extrêmement générales et inhérentes à bien des structures linguistiques : la nature arborescente de l'organisation textuelle qu'elle décrit est partagée par beaucoup de modèles allant de la phrase au texte, et la distinction entre relations noyau-satellite et multi-nucléaires se fonde sur une distinction très classique entre subordination et coordination. De ce fait, il est difficile de se détacher de la RST tant ses principes fondateurs se retrouvent nécessairement dans toute modélisation de la structure du discours. Il nous paraît donc nécessaire de garder à l'esprit que, si la RST possède un jeu extensible de relations, c'est bien le jeu particulier que l'on utilise pour décrire un phénomène discursif qui importe, et non la structure arborescente sous-jacente, qui est en soi d'une totale généralité. Faute de quoi on tendrait à considérer que toute structure arborescente incluant les notions de coordination et de subordination est à rattacher à la RST, ce qui nous paraît inexact.

En revanche, il nous paraît très intéressant de chercher à comparer les analyses produites à partir d'un jeu orienté vers la dimension rhétorique du discours (comme celui proposé par Mann et Thompson) avec des analyses produites par d'autres modèles arborescents visant à décrire d'autres dimensions du discours. Ce point fera l'objet de la section 8.5, où nous utiliserons la RST pour comparer la structuration discursive issue d'une analyse rhétorique avec celle issue d'une analyse thématique.

Chapitre 4

Analyse thématique en traitement automatique des langues

Dans le domaine du traitement automatique des langues, la notion de thème apparaît essentiellement dans la tâche d'*analyse thématique*, qui vise la structuration des textes selon des critères relatifs à la répartition de leur contenu informationnel. Cette tâche est toutefois rarement traitée dans sa généralité, et bon nombre de travaux visent plus particulièrement la *segmentation* thématique, c'est-à-dire le découpage en segments dits « thématiquement homogènes ». Les applications de ce type de segmentation concernent bien sûr la recherche et la navigation intra-documentaire, mais aussi des tâches telles que le résumé automatique. On notera d'ores et déjà que dans ce contexte, la représentation du contenu des segments obtenus est souvent considérée comme secondaire, et semble de fait laissée pour une grande part à la communauté de la recherche d'information où l'on parlera plutôt d'indexation.

Devant la multiplicité des propositions qui ont été faites dans ce domaine, nous adopterons ici comme dans les précédents chapitres le parti-pris consistant à décrire assez précisément un nombre limité de travaux. Comme on le verra, ces derniers représentent deux écoles *a priori* bien distinctes, dont le rapprochement montre bien la largeur du spectre des méthodes applicables au problème. Toutefois, ce chapitre nous servira avant tout à poser un certain nombre de jalons nous permettant de positionner notre approche, sans que nous cherchions à mener une réelle étude comparative. On se rapportera plutôt à (Hernandez, 2004) pour obtenir une couverture « en largeur » de l'état de l'art du domaine.

Nous allons tout d'abord envisager les approches de la famille que nous appellerons *« text-tiling »* (du nom de la méthode de Hearst, cf. infra), qui appliquent des méthodes essentiellement numériques pour détecter les phénomènes de cohésion lexicale et en déduire des ruptures thématiques. Puis, dans un second temps, nous considérerons les approches visant à exploiter des indices plus « linguistiques » pour analyser la structure thématique des textes, et nous détaillerons plus spécifiquement une méthode visant à obtenir une structuration hiérarchique en se fondant sur la dualité thème/rhème.

4.1 Segmentation thématique par cohésion lexicale

Les approches habituellement qualifiées de « quantitatives » ou « numériques » se fondent plus ou moins explicitement sur la notion de cohésion lexicale (Halliday et Hasan, 1976), en exploitant la répétition des mots comme indicateur d'homogénéité thématique. Il s'agit notamment des travaux se plaçant dans la lignée de (Youmans, 1991), (Hearst, 1994), (Reynar, 1994) ou encore (Salton *et al.*, 1996), qui procèdent à une segmentation linéaire du texte, c'est-à-dire en segments adjacents, dite *text-tiling*. En se basant sur la distribution des formes de surface, la méthode « de base » ne requiert aucune ressource externe ou presque, mais néglige en contrepartie la dimension sémantique du phénomène de

cohésion lexicale. Différentes méthodes ont été envisagées pour répondre à ce dernier problème, sur lesquelles nous reviendrons à la fin de cette section.

Le principe commun à de nombreuses méthodes de cette famille est le suivant : chaque segment minimal (par exemple le paragraphe) est caractérisé par un vecteur associant à chaque descripteur (mots « bruts » ou lemmatisés, par exemple) une valeur numérique représentative de sa fréquence dans ce segment, généralement obtenue par $tf \cdot idf$. On notera que des techniques de type « fenêtre glissante » peuvent se substituer à l'utilisation de segments minimaux déterminés *a priori*, avec l'avantage d'être utilisables quand aucune structuration adéquate n'est présente au sein du document, et l'inconvénient de ne pas bénéficier de la pré-segmentation fournie par certains éléments textuels à dimension thématique forte comme les paragraphes. Une fois ces calculs effectués, une mesure de distance vectorielle permet d'évaluer la cohésion thématique de chaque couple de segments. Ces derniers pourront alors être regroupés, par seuillage sur cette distance, en unités homogènes. Du point de vue de la caractérisation des segments, on pourra utiliser les descripteurs les plus fortement pondérés de chaque vecteur, point que nous discuterons dans la section 4.3.

Plus précisément, la méthode de base (telle que décrite, par exemple, dans (Masson, 1995)) est la suivante. On procède tout d'abord à différents pré-traitements visant à déterminer les unités sur lesquelles portera l'analyse. Dans le cas le plus simple, on se contentera d'un découpage en mots et on travaillera sur leur forme de surface. Une première amélioration consistera à appliquer une analyse morphologique permettant de sélectionner certaines catégories (noms et verbes par exemple)[1], et/ou de prendre en compte les lemmes au lieu des formes « brutes ». On pourra également utiliser d'autres types d'unités (syntagmes nominaux complexes, par exemple), bien que cela ne soit pas sans conséquence sur l'applicabilité de la méthode, point sur lequel nous serons amené à revenir plus tard.

À la suite de ces pré-traitements, on considère un niveau de segmentation *a priori* définissant l'unité textuelle minimale, à moins qu'une fenêtre glissante soit utilisée (sans que ce choix n'ait réellement d'incidence sur la suite de la méthode). On procède, au sein de chacune de ces unités textuelles, au décompte des occurrences des différents descripteurs retenus lors du pré-traitement. Chaque unité S_i est alors représentée par un vecteur du type $(c_{i1}, c_{i2}, ..., c_{in})$ dont chaque composante c_{ij} correspond au nombre d'occurrences du j^{eme} descripteur.

Une fois ce premier vecteur obtenu, il convient de pondérer les décomptes d'occurrences pour obtenir le poids de chaque descripteur. Le but de cette pondération est de diminuer l'importance des descripteurs présents uniformément dans le texte, et inversement de renforcer celle des descripteurs localisés sur quelques paragraphes. Le facteur généralement utilisé est le $tf \cdot idf$ classiquement utilisé en RI (cf. chapitre 1) et défini comme suit, où tf_{ij} est le poids du j^{eme} descripteur dans le segment i, df_j est le nombre de segments dans lesquels apparaît ce descripteur, et N le nombre total de segments.

$$P_{ij} = tf_{ij} \cdot log\left(\frac{N}{df_j}\right)$$

Ce vecteur pondéré est enfin utilisé pour mesurer la cohésion thématique des segments pris un à un dans leur ordre d'apparition, en se basant sur des mesures vectorielles de distance entre les vecteurs de segments consécutifs. Parmi les diverses mesures existantes, on trouve par exemple le *coefficient de Dice* , qui a pour intérêt de désavantager les couples de vecteurs dont les composantes nulles ne correspondent pas. Pour deux vecteurs $X = (x_1, x_2, ..., x_n)$ et $Y = (y_1, y_2, ..., y_n)$, ce coefficient est défini par :

$$C(X, Y) = \frac{2\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} {x_i}^2 + \sum_{i=1}^{n} {y_i}^2}$$

[1] On notera que cette étape ne vise pas spécifiquement l'élimination des mots grammaticaux, que la méthode tend à éliminer par elle-même.

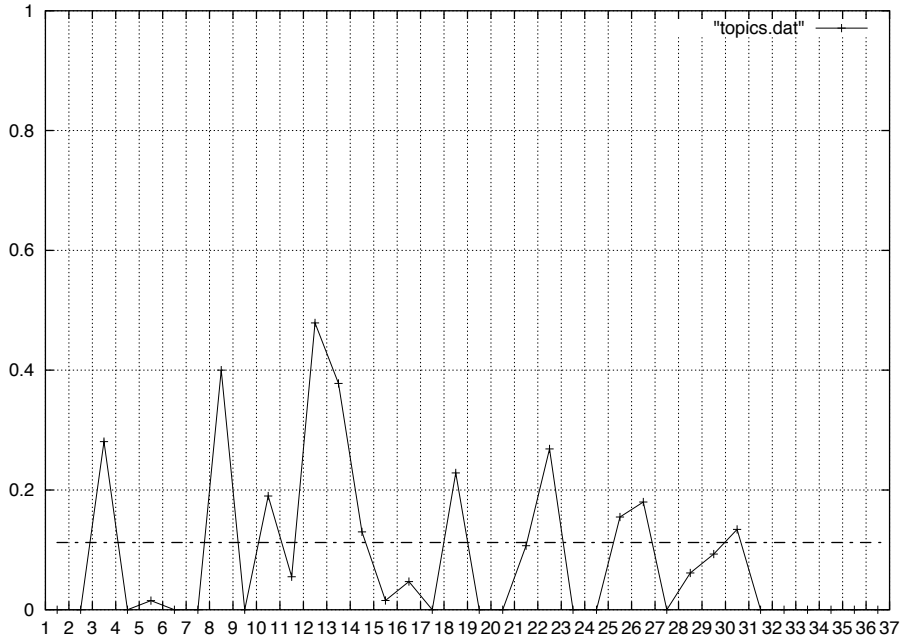


FIG. 4.1 – Exemple de graphe représentant les résultats d'une segmentation de type *text-tiling*. Les segments sont représentés en abscisse, et les coefficients aux intersections en ordonnée. Un seuil possible est représenté par la ligne horizontale discontinue, les valeurs situées sous ce seuil marquant une discontinuité thématique.

Ce coefficient prend ses valeurs dans [0; 1], les valeurs plus grandes témoignant d'une plus forte cohésion thématique entre les deux segments. Cette mesure de distance peut donc être utilisée pour regrouper des segments, en fixant un seuil en deçà duquel on considère que le discours présente une rupture thématique. Un graphe tel que reproduit par la figure 4.1 est un moyen courant d'observer les coefficients obtenus. On notera d'ores et déjà un premier défaut de la méthode, qui relève justement de la nécessité de fixer un seuil arbitraire et de toutes les difficultés qui en découlent : il semble que la seule solution « universelle » en la matière consiste à déterminer ce seuil en fonction du nombre de segments que l'on souhaite obtenir.

Bien entendu, la méthode que nous venons de présenter n'est qu'une possibilité parmi beaucoup d'autres, même si elle paraît assez représentative de la famille des méthodes fondées sur l'étude quantitative de la distribution des mots. Parmi les autres approches, il convient notamment de mentionner « Segmenter » (Kan *et al.*, 1998), qui s'appuie sur la notion de chaîne lexicale, ou encore « Dot Plotting » (Reynar, 2000) qui s'intéresse à la distribution « spatiale » des formes. Toutes ces méthodes souffrent toutefois des mêmes défauts, inhérents pour une part à leur caractère « surfacique », et d'autre part au modèle de segmentation linéaire lui-même.

Elles peuvent tout d'abord être mises en défaut par l'évocation d'un même concept par des mots différents mais sémantiquement proches. Par exemple, la méthode proposée dans (Kozima, 1993) se base sur la proximité sémantique entre les mots, calculée par une mesure de distance au sein d'un réseau sémantique. La pré-existence d'un réseau sémantique adapté demeurant une contrainte très lourde, d'autres méthodes se proposent de construire automatiquement un réseau lexical utilisable pour mesurer la proximité entre les descripteurs. L'approche proposée dans (Ferret *et al.*, 1997) consiste à construire un réseau de cooccurrences à partir d'un corpus suffisamment large, en partant de l'hypothèse classique que la cooccurrence fréquente de deux termes est un indice de proximité sémantique. Ce réseau de cooccurrences est ensuite utilisé de la même manière. Une méthode très similaire est

également décrite dans (Dias et Alves, 2005). Nous avons pour notre part expérimenté l'utilisation de ressources sémantiques componentielles dans ce cadre, expérience relatée en annexe B.

Une autre limitation forte des méthodes de ce type réside dans le modèle même de segmentation. L'approche linéaire considère en effet le texte comme une séquence de segments contigus entre lesquels se produit une rupture thématique franche. Or il est évident que la structure du discours est bien plus complexe, et qu'il s'agit d'un ensemble construit et cohérent dont la segmentation est artificielle en soi. On observe donc ici un fossé important entre ce qui relève de la réalité linguistique et les objectifs poursuivis par l'analyse automatique. Ce point nous confronte à la relation entre modélisation de la structure du discours d'une part et méthodes d'analyse d'autre part. Il convient de mentionner ici de rares propositions de méthodes à la fois purement numériques et s'appuyant sur un modèle discursif réellement structuré, comme par exemple dans (Ferret et Grau, 2001). Mais on peut néanmoins douter de la possibilité de mener une analyse réellement discursive sur des critères uniquement quantitatifs. Nous reviendrons que cette question dans la section 4.3, où nous évoquerons des méthodes hybrides où collaborent des critères quantitatifs avec des critères linguistiques tels que ceux que nous verrons dans la section suivante.

4.2 Structuration thématique fondée sur des critères linguistiques

Une autre approche consiste à exploiter des marqueurs, des indices, et plus généralement des formes linguistiques et dispositionnelles, porteurs d'indications de la structure thématique et plus généralement discursive. Un exemple est donné par la notion de cadre de discours, et d'abord de cadre thématique, c'est-à-dire constitué d'un introducteur tel que « en ce qui concerne X », « à propos de X », « considérant X », etc., dont on peut voir qu'il introduit un segment de plusieurs phrases ainsi présentées comme « relatives au thème X » (Porhiel, 2001). Des structures analogues concernant le temps ou l'espace peuvent également avoir une fonction thématique : des expressions telles que « dans les départements urbains » ou « dans les années 1950 » introduisent des segments d'une certaine portée, indexés par un critère sémantique spatial ou temporel. Nous verrons dans la partie II comment exploiter et généraliser ces remarques pour définir et calculer une organisation thématique arborescente du document, dans le cas de l'espace et du temps tout d'abord, puis autour de la notion plus générale de thème composite.

Bien d'autres marques de structuration peuvent être relevées. Dans (Jackiewicz et Minel, 2003) est définie la notion de marqueur d'intégration linéaire (« d'abord... ensuite », « premièrement... deuxièmement », etc.) et étudie leur rôle dans la structuration du discours à travers les cadres dits organisationnels. Des marqueurs de relations anaphoriques (par exemple, « ce X », « celui-ci », « ce dernier », etc. en début de paragraphe) sont aussi des signes possibles de continuation d'un thème. (Smolczewska et Lallich-Boidin, 2004) exploite ce type de marqueurs ainsi que des caractéristiques dispositionnelles fortes propres aux documents techniques (par exemple, des structures de listes) pour segmenter les documents techniques en unités homogènes. Citons encore un modèle des relations séquentielles dans les textes expositifs proposé par (Goutsos, 1996), qui propose un ensemble complexe de critères pour déterminer différentes relations entre segments consécutifs (encadrement, ouverture, fermeture et continuation).

Un autre système appartenant à cette catégorie, baptisé « Thema » puis « UniTHEM », est développé au GREYC autour de Nadine Lucas (Pinatel, 2003; Lucas et Giguet, 2005). La méthode employée se veut minimaliste en termes de ressources, et se fonde sur un modèle textuel multi-échelle où la notion d'unité thématique (dorénavant UT) relève de la dualité thème/rhème appliquée au niveau textuel, déjà mentionnée à plusieurs reprises dans les précédents chapitres (section 2.1 en particulier). Plus précisément, une UT représente ici un segment textuel décomposable en deux sous-unités, l'une étant thématique et l'autre rhématique. Cette structure est définie hiérarchiquement : une unité rhéma-

```
-Si l'UT est de premier niveau et compte 4 paragraphes ou plus :
Le thème est constitué du premier paragraphe.
Le rhème est constitué de tous les autres paragraphes.

- Si l'UT compte de 13 à 20 paragraphes (niveau différent de 1) :
On note n le nombre de paragraphes de l'UT, et m le quart de ce nombre,
éventuellement arrondi et plafonné à 10.
Le premier paragraphe du rhème est alors le premier paragraphe unaire (i.e.
constitué d'une seule phrase) rencontré entre les paragraphes numéros 3 et m.
Si celui-ci n'existe pas, on prend le paragraphe de plus petit cardinal (i.e.
celui comportant le plus petit nombre de phrases).

...
```

FIG. 4.2 – Exemples de règles appliquées pour identifier la frontière thème / rhème (extraites de (Pinatel, 2003)). Elles sont complétées par un processus de diagnostic permettant de valider ou d'infirmer la décision.

tique peut à son tour être scindée en sous-unités, et le document dans son ensemble constitue une UT de premier niveau. Le résultat de l'analyse est une segmentation conforme à ce modèle, où à chaque UT sont associés un ensemble de mots-clefs résultant de l'analyse conjointe de ses composantes thématique et rhématique. La figure 4.4 montre un exemple d'analyse produite par ce système.

L'ensemble du procédé est conçu pour s'appliquer à des documents de longueurs différentes, mais adapte son fonctionnement à chaque ordre de grandeur. Selon les cas, différentes mesures typographiques (MT) enchâssées sont considérées : unité appelée « virgulot » (suite de caractères délimitées par des ponctuations faibles, virgule, point-virgule, etc.), phrase, paragraphe, titre, section, chapitre, etc. Une fois identifiées ces différentes unités, le processus d'analyse ne tient pas compte de chaque type de MT particulier, et applique le plus souvent les mêmes règles quel que soit le grain. Différentes phases peuvent être distinguées au sein de ce processus :

- étant donnée une UT (au départ le document, puis par la suite des segments analysés comme tels), il s'agira de distinguer la partie thématique initiale de la partie rhématique qui lui fait suite ;
- étant donnée une unité rhématique, il s'agira le cas échéant de la subdiviser en UT plus fines ;
- pour chaque couple thème / rhème formant une UT, on cherchera enfin à déterminer un ensemble de mots-clefs représentatifs de son contenu informationnel.

À cette fin, le système exploite une variété d'indices relevant essentiellement de phénomènes de contraste ou d'écho constatés à des niveaux très différents. Pour la première phase, qui consiste à localiser la frontière entre les parties thématique et rhématique d'une UT, le système est tout d'abord guidé par des attendus *a priori* sur les proportions (en nombre de MT) des éléments thématiques et rhématique. Des exemples de règles s'appliquant à ce stade (extraite de (Pinatel, 2003)) sont reproduites dans la figure 4.2.

Par la suite, le système procède à un diagnostic permettant de valider ou infirmer le découpage obtenu. Cela consiste tout d'abord à chercher une « marque différentielle » entre les deux segments obtenus, comme le fait qu'une certaine classe d'indices soit représentée uniquement dans l'un ou l'autre. Une classe d'indices correspond ici à un ensemble de formes de surface fournies dans les ressources du système par des expressions régulières ou des critères sur la mise en forme. Par exemple, l'ensemble des déictiques, des marques de négation ou encore des expressions mises en relief typographiquement. Cela consiste d'autre part à chercher une marque clôture du rhème, en se limitant à ses deux dernières MT. Des exemples de règle s'appliquant à ce stade sont reproduits en figure 4.3.

```
-Si l'UT comporte 3 paragraphes ou moins, on recherche une phrase
contenant une conjonction de coordination.

- Si l'UT comporte 4 paragraphes ou plus, on recherche une paire de phrases
coordonnées (contiguës avec une conjonction de coordination en tête de la
seconde.

- Si aucune coordination n'a été trouvée, recherche de deux marques jumelles
contiguës, c'est-à-dire  présentant un identité de forme, de classe, etc.
```

FIG. 4.3 – Exemples de règles appliquées pour identifier la clôture d'un segment rhématique (extraites de (Pinatel, 2003)).

La seconde phase consiste à rechercher un éventuel découpage de niveau inférieur au sein des segments rhématiques. Là encore, un processus de type découpage / diagnostic est appliqué, la première étape recherchant des dispositions canoniques. On considérera par exemple que lorsqu'un segment rhématique de niveau supérieur à 1 est constitué de 2 MT, alors le rhème se subdivise en deux UT, l'une constituée des deux premières MT et l'autre de la troisième, ou encore qu'une UT de moins de deux phrases n'a pas lieu d'être segmentée. Dans le cas où aucune disposition canonique n'est trouvée, le système recherche une MT séparatrice. Sera considérée comme telle une MT présentant « un trait global discriminant par rapport aux MT qui l'environnent », celui-ci étant établi à partir de mesures statistiques sur l'ensemble des traits représentés dans chaque MT (ces traits appartiennent aux indices et peuvent marquer leur caractère déictique, temporel, défini / indéfini, spécifier une étiquette morpho-syntaxique, etc.).

Si aucune MT séparatrice n'est trouvée, on procède enfin à la recherche de configurations dites « appel-écho », c'est-à-dire faisant intervenir un phénomène d'attente entre un signal thématique et une clôture rhématique. Différentes relations de cet ordre sont proposées, comme « phrase définitoire / phrase conclusive » ou « phrase interrogative / phrase explicative » (ces propriétés d'ordre phrastiques sont déterminées à partir de patrons simples tels que la présence d'un pronom interrogatif ou d'une forme négative). Chaque relation d'appel-écho ainsi détectée forme finalement une nouvelle UT. Finalement, le processus de diagnostic du découpage obtenu fait intervenir des règles du même ordre que dans le cas précédent, recherchant à la fois un appareillage et des marques différentielles entre les UT.

La troisième et dernière phase consiste à identifier, pour chaque UT, un ensemble de mots-clefs caractéristiques, c'est-à-dire représentatifs de l'à propos du segment. L'intérêt du procédé proposé est qu'il ne s'appuie pas sur des critères quantitatifs mais sur la segmentation thème / rhème préalablement obtenue. Le système cherche en effet de préférence des substantifs définis (sont considérés comme tels les mots qui sont précédés d'un déterminant défini, sans faire appel à une réelle analyse morpho-syntaxique) dans le segment thématique et apparaissent également dans la partie rhématique de la même UT (ils sont dans ce cas dits « confirmés »). On notera que le procédé pourrait facilement intégrer des critères quantitatifs, par exemple pour sélectionner, parmi les mots confirmés, ceux qui sont numériquement les plus saillants.

De notre point de vue, l'originalité majeure de la méthode de d'analyse thématique proposée est de se fonder en partie sur des indices dont le rôle dans la structuration thématique est, à première vue, loin d'être immédiat. En effet, contrairement aux procédés qui s'appuient sur les indices immédiatement issus du modèle sous-jacent, comme un introducteur dans le modèle de l'encadrement du discours, celui qui est ici à l'oeuvre exploite des critères qu'il semble difficile de caractériser *a priori* comme indices de la structure thématique, mais qui peuvent ressortir d'une démarche privilégiant l'observation de corpus. Par exemple, si la fonction thématique d'une locution comme « il semble que » n'apparaît pas immédiatement, les auteurs ont observé (dans un certain corpus) que son apparition est caractéris-

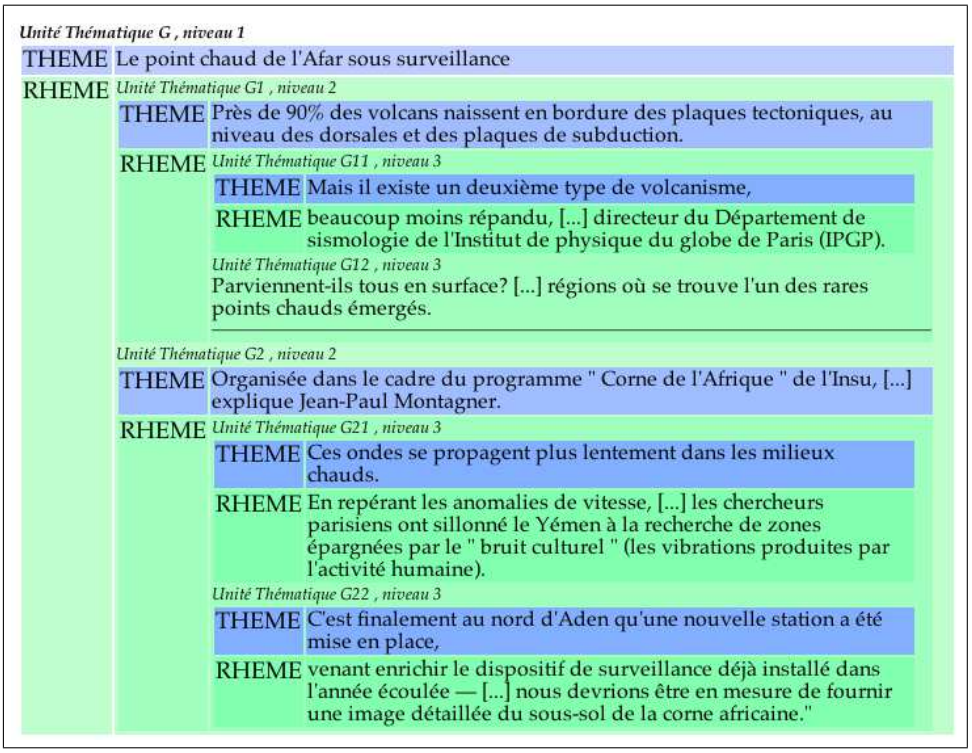


FIG. 4.4 – Exemple d'analyse thématique hiérarchique produite par UniTHEM.

tique d'une phrase conclusive qui, en écho avec une proposition négative, fournit un indice utilisable pour segmenter le discours. De même, on peut supposer que les intervalles fixés quant à la taille des UT pour choisir telle ou telle règle, ou le fait de reconnaître la clôture d'un segment rhématique par la présence d'une conjonction de coordination résulte d'une démarche empirique, même s'ils répondent vraisemblablement à une réalité psycho-cognitive.

L'inconvénient de cette approche est qu'il est difficile de discuter *in abstracto* du bien-fondé de ces critères, que seuls les résultats obtenus expérimentalement permettent de valider. Cette voie est sans aucun doute intéressante à explorer, même s'il nous est permis de douter du caractère minimaliste et surtout du degré de généralité de ces ressources. On peut aussi s'interroger sur la charge de travail nécessaire à leur constitution, et considérer l'opportunité, quitte à exploiter des indices dont la justification linguistique est parfois difficile à établir, de mettre en oeuvre des techniques d'apprentissage pour les obtenir.

4.3 Conclusion

Pour conclure, il nous faut tout d'abord remarquer que l'état de l'art en matière d'analyse thématique automatique apparaît comme largement dominé par les approches « quantitatives » évoquées dans la section 4.1, au détriment des approches « linguistiques » telles que décrites dans la section 4.2. L'attrait pour les premières s'explique facilement, si l'on excepte le penchant naturel de l'informaticien pour ce qui relève du numérique, par une faible complexité calculatoire et un degré de généralité très élevé. Elles sont en effet applicables indifféremment (à l'exception, éventuellement, de quelques pré-traitements) à des documents de toutes tailles, de toutes langues, de tous genres ou presque, et ce avec un coût qui les rend applicables à des volumes documentaires considérables.

Il nous paraît toutefois indispensable d'envisager sérieusement l'apport des approches linguis-

tiques si l'on souhaite progresser dans ce domaine, en commençant par (re)définir clairement la ou les tâches elles-mêmes : qu'entend-on par analyse thématique, et quel résultat attend-on des procédés qui s'attellent à ce problème ? On peut en effet s'étonner de l'absence, dans nombre des travaux concernés, de discussion visant à mieux définir ce que recouvrent les termes de « segmentation thématique » ou de « segment thématiquement homogène », le plus souvent définis de façon assez allusive en faisant appel à des notions certes intuitives mais très peu précises. De ce fait, il semble que la notion de segment thématique propre aux méthodes de la famille *text-tiling* soit avant tout définie par ces méthodes elles-mêmes, et non par un modèle posé *a priori*. En d'autres termes, elle serait conditionnée plutôt par les moyens utilisés pour l'obtenir automatiquement que par une quelconque modélisation du matériau traité, qui ne peut être que linguistique. Et de fait, ces méthodes sont à peine dépendantes de la nature de ce matériau, puisqu'elles pourraient indifféremment segmenter toute séquence de signes, mots ou autres.

Tout semble donc se passer comme si l'élaboration de procédés de segmentation thématique avait précédé la définition de leur propre tâche. Il nous semble que ce problème est à rapprocher, par exemple, de certains travaux visant à établir automatiquement des classes sémantiques sans pour autant caractériser précisément le type de relation que sont censés entretenir les constituants de ces classes. Les méthodes fondées sur des contextes communs permettent bien d'obtenir des listes de mots « sémantiquement proches », mais ce fait suffit-il à circonscrire une tâche reproductible, évaluable, améliorable ? Rien n'est moins sûr, mais dans ce domaine comme dans celui de la segmentation thématique, sont mises en place des campagnes d'évaluation qui se confrontent inévitablement à la nécessité de fixer des objectifs à atteindre. Et le manque de définition précise des objets à identifier apparaît alors on ne peut plus clairement.

Le cas de la campagne DEFT 2006 (2e DÉfi Fouille de Textes) portant sur la segmentation thématique nous paraît éloquent sur ce point. Le corpus d'apprentissage proposé est en effet constitué de trois ensembles de segments dits « thématiques » d'une frappante hétérogénéité, décrits comme suit dans la spécification de la tâche à réaliser par les équipes candidates[2] :

> Pour DEFT'06, la segmentation thématique des textes devra s'appuyer sur des corpus de différents domaines écrits en français : discours politiques, textes juridiques et ouvrage scientifique. Le but du défi consiste à déterminer les segments thématiques de ces différents textes (c'est-à-dire, les premières phrases de chaque segment thématique). Nous allons décrire avec plus de précision les corpus utilisés et la définition des segments thématiques pour chaque type de corpus :
>
> – Discours politiques. Le corpus est composé de discours politiques prononcés par des présidents de la république française. La segmentation thématique est basée sur la structure thématique des discours mis en ligne sur le site de référence.
> – Textes juridiques. Le corpus est composé d'articles de lois de l'Union Européenne. Les segments thématiques sont les articles des lois.
> – Ouvrage scientifique. L'ouvrage scientifique utilisé est le livre « Apprentissage Artificiel » d'Antoine Cornuéjols et Laurent Miclet. Avec ce corpus, les segments thématiques à retrouver sont les différentes sections (chapitres, sections, sous-sections, sous-sous-sections). Les titres des différentes sections ainsi que les figures, tableaux et les équations ont été supprimés. Le but est de déterminer la première phrase de chaque section.

Il est clair en effet que ces « segments thématiques » sont en fait des objets linguistiques de natures totalement différentes : articles de lois pour ce qui est des textes juridiques, chapitres et sections pour l'ouvrage scientifique, et segments qui s'avèrent être des paragraphes dans le cas des discours politiques. Il paraît donc extrêmement difficile d'en proposer une modélisation uniforme, et la tâche

[2] http://www.lri.fr/ia/fdt/DEFT06/

semble clairement orientée vers les méthodes numériques, qui peuvent vraisemblablement s'appliquer indifféremment dans les trois cas.

Dans ce contexte, s'il on veut bien admettre que les méthodes fondées sur la distribution des mots ne peuvent accéder à toute la subtilité de la structuration thématique des textes, il peut sembler nécessaire de « renverser la vapeur » en commençant par circonscrire les objets linguistiques qui semblent pertinents dans ce domaine, pour ensuite définir des tâches, et finalement développer des procédés d'analyse. On aura compris que notre démarche s'inscrit clairement dans cette dynamique, et nous tenterons à cette fin, dans le chapitre suivant, de tirer du panorama bibliographique que nous terminons ici des enseignements pouvant nourrir une approche de l'analyse thématique qui soit réellement fondée sur une définition *a priori* de ce que l'on qualifie de *thème*. La contrepartie de ce positionnement est bien sûr l'impossibilité de mettre au point une méthode d'analyse dont la portée serait aussi générale que celles que nous évoquions plus haut. Nous nous concentrerons au contraire sur certains phénomènes discursifs bien spécifiques, notre apport constituant plutôt une brique parmi d'autres briques telles que celles décrites dans la section précédente.

Nous souhaitons toutefois insister sur le fait que cet objectif ne réfute en aucun cas les apports des méthodes numériques. Au contraire, nous considérons avec attention la fructueuse collaboration qui se dessine entre les deux types de méthodes, à la manière du projet Régal (Ferret *et al.*, 2001b; Couto *et al.*, 2004; Hernandez, 2004) où les introducteurs de cadres thématiques sont exploités comme indices de rupture en complément de techniques quantitatives. Globalement, il paraît très intéressant d'utiliser ces dernières pour obtenir des indices à considérer *en complément* à d'autres indices. Ainsi, si le phénomène de cohésion lexicale apparaît comme jouant effectivement un rôle significatif au regard des structures que l'on cherche à analyser, il sera bien sûr utile de chercher à le reconnaître par tous les moyens possibles.

Une dernière remarque concerne la *représentation* des thèmes eux-mêmes. Au delà de la segmentation du texte, beaucoup d'applications visées par l'analyse thématique nécessitent de représenter le contenu informationnel des segments obtenus, par exemple à des fins d'indexation. Il nous semble qu'il s'agit là d'un autre problème important à considérer, dans la mesure où il est loin d'être toujours pris en compte explicitement. Dans le cas des méthodes numériques, il s'agit plutôt d'un produit indirect de la méthode employée, comme les vecteurs de coefficients fréquentiels pour le cas décrit dans la première section de ce chapitre : l'élaboration du procédé semble là aussi devancer la définition de l'objet. L'approche décrite dans la section précédente nous semble plus intéressante sur ce point, puisqu'elle considère *par hypothèse* la structure thématique qu'elle calcule (en l'occurrence segments thématiques et rhématiques) comme critère de sélection des mots représentatifs de l'à propos d'un segment. Nous nous inscrirons là encore dans cette catégorie de méthodes, en commençant par définir clairement les objet thématiques que nous souhaitons obtenir, avant de procéder à leur analyse effective.

Chapitre 5

Bilan

Nous nous sommes efforcé de montrer au long des précédents chapitres la grande diversité des concepts que recouvre l'idée de thème, en rapprochant différentes disciplines qui partagent ce même objet d'étude tout en adoptant des approches différentes. Du foisonnement qui en résulte émergent toutefois quelques notions transversales, dont l'étude nous semble riche en enseignements quant aux buts que nous poursuivons ici.

Mais bien que le choix des travaux relatés dans les précédents chapitres soit en lui-même orienté, les réalités qui y sont dénommées « thème », « topique » ou encore « sujet » ne sont pas nécessairement compatibles, et il faut prendre garde à ne pas rapprocher artificiellement des notions qui n'ont d'autre rapport entre elles que leurs dénominations. Pour distinguer clairement les axes horizontaux sur lesquels nous pourrons nous appuyer, il est donc nécessaire de préciser la posture qui nous autorise à poser un regard à la fois sélectif et unificateur sur la diversité des concepts abordés. Nos perspectives applicatives sont bien évidemment déterminantes dans la définition de cette posture, puisqu'elles circonscrivent *a priori* nos attendus.

Étant essentiellement liées à la recherche d'information, ces perspectives applicatives nous incitent à considérer la notion de thème du point de vue de l'*à propos*. Nous cherchons en effet avant toute chose à décrire et représenter ce dont il est question dans un texte ou un fragment de texte, de façon à faciliter l'accès à l'information qu'il véhicule. La notion d'à propos reste évidemment très floue, ce qui justifie ce chapitre. Mais ce choix nous permet de prendre immédiatement nos distances avec des approches fondamentalement différentes, comme celles qui associent *par définition* le thème à des propriétés purement formelles (position initiale par exemple), celles qui viseraient *uniquement* la segmentation thématique du texte sans chercher à décrire les thèmes des segments obtenus, ou encore celles qui considèrent la notion de thème *en dehors* des textes.

Une autre propriété que nous attribuons à la notion de thème est d'être proprement *discursive*. Pour cette raison, nous ne pourrons adopter immédiatement des modèles qui seraient trop fortement centrés sur des grains fins (tels que la phrase), ni des modèles axés sur des grains trop importants (comme le document). Nous chercherons plutôt à nous abstraire de tout grain particulier, le propre de l'organisation discursive d'un texte étant de mettre en jeu des unités d'ordre divers au sein d'une structure globale. Le rapport que cet objectif nous fait entretenir avec les différentes conceptions du thème que nous avons rencontrées n'est plus, cette fois, de l'ordre du filtrage, car trop peu de travaux portent explicitement sur la notion de thème discursif. Il sera donc nécessaire dans ce domaine de nous risquer au rapprochement entre théories ou même entre disciplines *a priori* distantes.

Enfin, une autre contrainte forte imposée par notre visée applicative concerne la mise en oeuvre informatique proprement dite : il s'agit bien évidemment de se doter de modèles opératoires (au sens du TAL) susceptibles *in fine* de fonder des outils fonctionnels. Précisons que dans la démarche que nous adoptons, cela ne constitue aucunement une restriction quant à la complexité des modèles linguistiques

ni quant à la nature des ressources que l'on s'autorisera à fournir à la machine : nous ne nous interdirons pas d'employer des analyses dites « profondes » et/ou des ressources dites « lourdes » lorsque cela sera nécessaire. En revanche, la contrainte opératoire impose la formalisation complète des modèles à mettre en oeuvre, ce qui est loin d'être immédiat lorsque l'on traite de notions aussi fuyantes que celles qui touchent à la définition de l'à propos d'un texte.

Après avoir fixé ce point de vue propre à notre visée applicative, nous pouvons tenter de discerner les lignes de force qui se dégagent de l'ensemble des travaux évoqués jusqu'ici, et de voir en quoi elles peuvent fonder une approche de la notion de thème qui serve au mieux nos objectifs.

5.1 Le thème comme « point de contact » avec un état de connaissances

Notre premier point concerne la dualité « donné » (ou « connu ») *vs.* « nouveau », que nous avons rencontrée très fréquemment dans les précédents chapitres. En toute généralité, il s'agit de distinguer des référents de discours que l'on considère comme appartenant déjà à l'espace mental d'un allocutaire, de ceux qui relèveraient plutôt de l'information nouvelle que l'on apporte sur les précédents. Cette problématique est bien sûr d'un intérêt primordial dans le contexte de la recherche d'information, puisqu'il s'agit bien, quand un utilisateur cherche de l'information relativement à une certaine notion, de lui fournir des éléments apparaissant dans un texte comme apportant de l'information nouvelle relativement à sa requête. À l'inverse, si un texte évoque le contenu de sa requête en tant qu'information nouvelle, on pourra peut-être considérer qu'il est moins pertinent, ou alors d'une autre manière. En tout état de cause, on voit se dessiner clairement le rôle de cette dualité « donné » / « nouveau » dans l'adéquation entre la formulation d'une requête et l'information effectivement fournie à l'utilisateur.

Cette dualité apparaît d'une part au niveau documentaire, pour fonder la notion de thème par opposition à celle de rhème. On se rappellera notamment les termes de Hutchins, qui qualifie le thème de « point de contact fourni au lecteur par le scripteur, permettant de relier son propos soit à un certain contexte ou environnement, soit à un discours ou texte antérieur » (cf. section 2.1). Dans cette optique, elle s'applique à différents niveaux (entre deux parties d'un même texte ou entre plusieurs textes), et ce même auteur va d'ailleurs jusqu'à souligner le parallélisme entre progressions thématique et relations bibliographiques dans la littérature scientifique (cf. figure 2.2 p. 34), suggérant que le principe d'accumulation d'information s'applique de façon équivalente à des niveaux de grain très divers.

Cette même dualité apparaît également au sein des travaux plus proprement linguistiques, notamment autour de la notion de structure informationnelle (cf. section 3.1). Comme nous l'avons vu, cette approche s'applique plus spécifiquement au niveau phrastique, tout en établissant une caractérisation très fine des référents du discours en termes d'identifiabilité et d'accessibilité. La notion inter-phrastique de progression thématique (cf. section 3.2.2) repose également sur cette dualité pour étudier les différentes modalités d'accumulation de l'information au niveau inter-phrastique. On remarquera que dans l'ensemble de ces travaux, les aspects extérieurs au texte semblent considérés comme secondaires, et la familiarité de l'allocutaire avec un référent repose avant tout sur les circonstances de son évocation au sein même du discours. De fait, on parlera plutôt dans ce contexte de « donné » que de « connu ».

On voit ainsi apparaître une certaine fragmentation entre les approches plutôt « documentaires » et plutôt « linguistiques » au sujet de la dualité « donné » / « nouveau ». Il nous semble toutefois que la réalité décrite par ces différentes écoles est bien la même, et qu'une vue globale sur le phénomène semble tout à fait constructive. Nous souhaitons à ce sujet reproduire à nouveau la citation suivante de Charolles, qui discute ce problème à propos de la topicalité des introducteurs de cadres (cf. section 3.4.1) :

> Le fait que les adverbiaux détachés en position initiale s'appuient sur des informa-

> tions déjà évoquées dans le discours antérieur explique qu'on les considère comme des topiques. Cette idée d'appui n'est cependant pas facile à préciser. [...] Il semble préférable de partir de l'idée que les adverbiaux sélectionnent du fait de leur signification un attribut (temps, lieu, condition, cause, moyen, manière) des états de choses pour indexer la proposition en tête de laquelle ils apparaissent. Lorsque le critère sélectionné est générique, comme le temps et le lieu qui sont des traits inhérents à tous les états de choses, la décision consistant à l'exploiter comme un index n'a pas besoin d'être justifiée par le contexte précédent, ce qui fait qu'ils peuvent apparaître à l'initiale de discours, surtout si le type de texte s'y prête. (Charolles, 2003, p. 25)

Charolles évoque ainsi le fait qu'un introducteur n'est pas nécessairement topical au sens strict (i.e. en termes d'identifiabilité ou d'accessibilité), puisque certains critères dits « génériques » peuvent indexer une proposition sans préalablement être explicitement introduits dans le discours. Il justifie ce point par le fait que ces critères génériques peuvent apparaître en initiale de texte.

Nous souhaitons prolonger cette idée en formulant l'hypothèse que le statut de topique en termes de « donné » ou « connu » peut être attribué à bien d'autres référents qui, sans spécifier « des traits inhérents à tous les états de choses » comme le temps et le lieu, sont supposés par le scripteur comme présents à l'esprit de l'allocutaire sans avoir à être préalablement introduits dans le discours. Le critère proposé par Charolles semble confirmer cette hypothèse puisque, comme le montre l'exemple reproduit dans la section 3.4.1, des concepts particulièrement « saillants » dans le domaine de connaissance auquel se rapporte un texte peuvent effectivement apparaître en initiale de texte, textualisés sous forme d'introducteurs.

En d'autres termes, il nous semble utile de considérer que la qualité de « donné », ou plutôt, dans ce cas particulier, de « connu », peut aussi dépendre de l'état mental propre à l'allocutaire et non pas seulement des référents qui sont explicitement introduits en discours. Nous rejoignons ici l'hypothèse déjà mentionnée dans la section 3.1.1, autour des notions de connaissance partagée ou de familiarité supposée, selon laquelle le texte constitue un tout confronté aux connaissances du lecteur. Insistons sur le fait que cette optique n'oblige en rien à prendre *uniquement* en considération ce qui relève des connaissances, mais plutôt à considérer le « donné » (qui dépendrait de l'accessibilité des référents au sein même du discours) et le « connu » (qui dépendrait de l'environnement de la perception du discours) comme *complémentaires*.

Il nous semble que cette hypothèse prend toute sa force, dans le cadre applicatif qui nous occupe ici, si l'on ne prend pas seulement en compte la familiarité du lecteur avec un référent telle qu'elle est *supposée* par le scripteur, mais aussi le fait que *chaque lecteur* sera amené à considérer un texte avec ses connaissances propres, qui s'éloigneront plus ou moins des connaissances supposées par le scripteur. Dans cette optique, la détermination du statut topical d'un constituant sera effectivement spécifique à chaque lecteur, et donc à chaque demandeur d'information. Cela implique, pour déterminer le thème d'un objet textuel donné, de considérer non seulement les suppositions du scripteur qui transparaissent ou non dans la forme du discours lui-même, mais aussi les connaissances effectives du lecteur potentiel, c'est-à-dire de l'utilisateur du système de recherche d'information. Finalement, il s'agit de prendre en compte parallèlement trois aspects pour décider de la représentation de l'à propos d'un texte :

- le texte lui-même, où un certain nombre d'éléments thématiques seront explicitement signalés comme tels ;
- les connaissances du lecteur telles qu'elles sont supposées par le scripteur, qui seront souvent en lien direct avec les connaissances propres au domaine dont traite le texte (ou plus généralement avec l'environnement du discours), et qui peuvent jouer un rôle dans la détection des éléments thématiques quand le texte lui-même ne les signale pas explicitement ;
- les connaissances effectives du lecteur, qui peuvent ne pas se conformer aux suppositions du scripteur, et ainsi rendre rhématiques des éléments supposés connus par le scripteur, ou au

contraire rendre thématiques des éléments supposés nouveaux.

Il s'agit donc de considérer que les connaissances supposées ou effectives du lecteur (ou, dans le contexte de la RI, du demandeur d'information), peuvent intervenir de façon significative dans la discrimination de ce qui relève du thématique ou du rhématique dans un texte. Nous rejoignons ainsi certaines des propositions que nous avons mentionnées au cours des précédents chapitres, notamment avec Maron, Hutchins et Hjørland.

On remarquera tout d'abord un certain rapprochement avec les propositions de Maron lorsqu'il cherche à rendre compte du degré de subjectivité qui caractérise la relation d'à propos (cf. section 2.3.1). On pourrait d'ailleurs chercher à faire correspondre les trois points évoqués ci-dessus avec la typologie que cet auteur a établie : la relation d'à propos qui lie un document et son thème tel que perçu par un lecteur donné appartiendrait à la catégorie *S-about* ; celle qui lie ce même document à l'expérience de son rédacteur pourrait probablement appartenir à la catégorie *R-about* (pour peu que ce rédacteur prenne en compte les spécificités du public visé) ; et celle qui le lie à une représentation objective (*O-about*, indépendante de toute expérience personnelle) pourrait être celle que l'on obtient en se limitant au texte lui-même. Cette classification reste en l'occurrence assez anecdotique, mais elle met en lumière le fait que « l'à propos peut être interprété de différents points de vue » (Maron, 1977, p. 40), fait que nous avions par ailleurs mis en évidence dans la section 2.2 à propos des approches logiques.

Nous avons également évoqué les positions de Hjørland (cf. section 2.3.2), qui rejette l'idée d'une notion de thème qui serait inhérente aux documents eux-même, et défend au contraire une posture *pragmatique* consistant à prendre en considération les habitudes liées à une pratique donnée pour obtenir une indexation plus efficace. Nous avons également vu avec Hutchins l'importance des connaissances des utilisateurs dans le contexte de la recherche d'information (cf. section 2.1). Ce dernier remarque que l'utilisateur cherche « un document qui présuppose un état de connaissances présentant des affinités avec le sien » (Hutchins, 1977, p. 29), en d'autres termes qu'il s'agit idéalement de trouver une information qui se situe dans le « prolongement » de ses propres connaissances, dans une direction choisie par lui. Ce même auteur conclut que les éléments thématiques d'un document doivent être recherchés dans les portions où « l'auteur établit des points de contact avec ce qu'il suppose être l'état de connaissance des lecteurs potentiels » (*ibid.*).

On peut donc considérer qu'un système de recherche d'information idéal serait capable d'établir une adéquation parfaite entre l'état de connaissances de l'utilisateur (ou d'un groupe d'utilisateurs) et les éléments thématiques des documents pertinents, ces derniers constituant le socle de l'information effectivement recherchée. Or, l'état de connaissances du ou des utilisateurs ne peut clairement pas être exprimée dans une simple requête au sens habituel du terme. Il devient au contraire utile de permettre à l'utilisateur de formuler d'une façon plus élaborée ce qui constitue des points de contacts avec ses connaissances propres, ou avec celles d'un certain domaine de connaissances. Ces points de contact pourront alors être utilisés pour rechercher plus efficacement ce qui relève du thématique dans les documents de la base documentaire considérée.

5.2 Le thème comme objet structuré

Nous avons été amenés à considérer, au cours de cette première partie, différentes réalités recouvertes par les termes de sujet, de topique ou de thème. Selon les théories, il pourra s'agir :

- d'une proposition chez Hutchins ou Hjørland (cf. sections 2.1 et 2.3.2) ;
- d'un référent du discours (au niveau de la structure informationnelle, cf. section 3.1) ;
- d'une macro-structure « épurée », à nouveau chez Hutchins mais avant tout chez Van Dijk (cf. section 3.3.1) ;

– d'une molécule sémique dans le cadre de l'analyse componentielle (cf. section 3.3.2) ;
– d'un ensemble de facettes dans la théorie de Ranganathan (cf. section 2.3.3) ;
– d'un vecteur d'information dans l'approche logique (cf. section 2.2) ;
– etc.

On le voit, il s'agit dans la majorité des cas d'objets complexes et structurés, ce qui paraît d'autant plus nécessaire lorsque l'on considère les thèmes *discursifs*, comme nous avons pu le voir dans les sections 3.3 et 3.2. Or, en recherche d'information et en traitement automatique des langues, le thème d'un objet textuel est le plus souvent représenté par de simples listes de descripteurs, que ce soit dans le cadre des méthodes de RI « classiques » (cf. section 1) ou de certaines méthodes de segmentation thématique automatique (cf. section 4). Le fait de considérer le thème d'un objet textuel comme un objet *structuré* constitue donc de notre point de vue une opportunité significative d'amélioration des systèmes de RI.

Reste à déterminer *quelle* structure, puisque la littérature évoque des alternatives très différentes, au sein desquelles on distinguera notamment les structures constituant une représentation synthétique du texte, de celles qui représentent uniquement la partie « thématique » de celui-ici.

Les macro-structures à la Van Dijk sont un candidat possible. Mais quelle que soit leur validité linguistique, il apparaît très difficile de procéder automatiquement à l'analyse complète de structures de ce type, qui nous ramènerait à la problématique de la compréhension automatique (Charnois et Enjalbert, 2005). Et même en supposant que cela soit possible, on peut douter du fait que ce type de structure corresponde effectivement à ce que nous attendons de la notion de thème dans le contexte de la recherche d'information. Dans la section précédente, nous avons longuement évoqué la dualité thème *vs.* rhème pour souligner l'intérêt du thème en tant que point de contact entre l'état de connaissances de l'utilisateur et l'information disponible dans un document. Dans cette optique, il ne s'agit clairement pas de représenter le contenu global d'un objet textuel, mais seulement le « socle » sur lequel il se fonde pour apporter de l'information nouvelle. La première option relèverait plutôt du résumé automatique (cf. section 1.6) que de l'indexation, la structure obtenue visant plutôt à se *substituer* au document lui-même plutôt que d'en *représenter* le contenu. Face à l'alternative relevée par Hutchins entre approche thématique et approche par résumé (cf. section 2.1), nous nous situons donc très clairement du côté de l'approche purement *thématique*.

Mais cela n'est bien sûr pas incompatible avec l'idée de représenter les thèmes de façon structurée. Au cours des précédents chapitres, nous avons mentionné cette possibilité tout d'abord au niveau documentaire avec le principe de classification par facettes (cf. section 2.3.3), et ensuite au niveau discursif avec la notion d'univers de discours (cf. section 3.4.1).

Rappelons que l'indexation par facettes consiste, dans sa version « assouplie », à déterminer, pour un domaine de connaissances donné, un ensemble de traits caractéristiques qui s'adjoindront à un thème « principal » pour décrire le contenu informationnel d'un document. Ce principe nous semble en parfaite adéquation avec l'idée de points de contact que nous avons évoquée plus haut, puisqu'il permet, au niveau de l'indexation, de situer le thème d'un document au sein de la structure des connaissances propre à un domaine. Pour l'utilisateur, cela autorise la formulation d'une requête qui comprend à la fois des descripteurs définissant l'information recherchée elle-même et des descripteurs définissant le contexte spécifique au sein duquel cette information doit se situer. Dans le domaine géologique, on pourra par exemple chercher de l'information relative au phénomène de « l'érosion » tout en spécifiant les facettes « type de sol » (ex : « limon rouge » ou « calcaire ») et « zone géographique » (ex. « Normandie » ou « Bassin parisien »).

Si le principe de l'indexation par facettes a été initialement conçu pour décrire des documents entiers, il nous semble particulièrement intéressant de le connecter à la notion de thème discursif. En effet, il apparaît qu'une composante importante de l'organisation du discours est liée au besoin du scripteur (ou du locuteur) de *situer* son propos relativement à des facettes multiples, c'est-à-dire d'instaurer un

thème qui comporte souvent plusieurs composantes. Ainsi, pour reprendre l'exemple précédent, si l'on souhaite discourir de « l'érosion des sols argilo-limoneux en Normandie », on pourra bien sûr instaurer ce thème en une seule fois, par exemple sous la forme d'un titre ou d'une lexicalisation complète. Mais bien souvent, l'établissement d'un tel thème fera appel à des outils discursifs qui permettront d'introduire progressivement ses différentes facettes. Cela sera par exemple le cas si l'on introduit d'abord la problématique générale (l'érosion), avant de la situer relativement à la facette « type de sol », pour enfin introduire une localisation spatiale. En d'autres termes, il s'agit donc de préciser progressivement les limites de l'univers de discours, limites qui seront le plus souvent mouvantes quand sont énumérées au fil du discours plusieurs valeurs possibles pour une même facette.

Dans ce cas, les structures discursives en jeu seront multiples car le texte est fortement contraint par sa linéarité, alors que les différentes facettes du thèmes doivent se superposer et non simplement se succéder. C'est ici que nous rejoignons la notion d'encadrement du discours de Charolles, ce type de structure discursive apparaissant comme un moyen privilégié d'instaurer des « facettes » parallèlement à un « thème principal ». Cette question est elle-même liée à la notion de « scene setting topic » que nous avons abordée dans la section 3.1.2 au sujet de la dualité « sujet » *vs.* « topique » : quand une phrase débute par « En Normandie, l'érosion... » ou « Sur les sols calcaires, l'érosion... », nous avons vu que que les différentes théories de la structure informationnelle peinent à s'entendre sur topique. Si l'on raisonne en terme d'à propos, s'agit-il des « sols calcaires » ou de « l'érosion » ?

Notre hypothèse est que, du point de vue de l'à propos, les deux composantes doivent être prises en compte simultanément pour décrire de façon précise le segment considéré. Cela peut être vu comme une application du principe de l'indexation par facettes à l'analyse thématique discursive, partant du principe que plusieurs composantes peuvent constituer différents points de contacts entre un état de connaissances et une information nouvelle. C'est cette hypothèse qui fondera la notion de thème composite que nous proposerons plus loin.

Mais au delà de la description des thèmes eux-mêmes, cette approche nous paraît également intéressante quant au problème de la relativité de la relation d'à propos. Comme nous le mentionnions plus haut, notamment avec Maron, il est difficile d'associer de façon objective un thème à un objet textuel donné. Mais il est tout aussi difficile, une fois que ce thème est déterminé, de juger de son adéquation avec la recherche effective d'un utilisateur : doit-on considérer que « ravinement » est à propos de « érosion », que « Normandie » est à propos de « Caen », ou encore que « limon » est à propos de « inondation » ?

Sans qu'elle apporte de réponse immédiate à des problèmes particuliers, l'approche logique que nous avons évoquée dans la section 2.2 semble intéressante de ce point de vue, car elle permet d'étudier les propriétés de l'à propos en tant que *relation*. La description purement mathématique ou logique ne constitue bien sûr pas comme une obligation, mais les notions proposées mettent bien en lumière l'impossibilité de définir objectivement une relation d'à propos universelle, valable en toute circonstance. On pourra ainsi chercher à définir des relations d'à propos plus ou moins spécifiques à un domaine ou à un groupe d'utilisateurs, mais cela peut encore sembler trop général. La notion de facette peut à nouveau intervenir ici car elle permet de définir *différentes* relations spécifiques à chacune des facette considérées. On pourra par exemple considérer la relation d'à propos liée à la facette spatiale comme symétrique ($Normandie \models Caen$ et inversement) mais non transitive ($Caen \models Normandie$ et $Normandie \models Rouen$ mais $Rouen \not\models Caen$), tout en adoptant des conventions totalement différentes pour d'autres facettes.

En contrepartie des avantages ici mentionnés, l'analyse par facettes appliquée au niveau intra-documentaire nécessite clairement de procéder à une analyse fine de la structure du discours, qui ne peut plus se limiter à l'étude quantitative de la distribution des mots. Il faut au contraire prendre en compte des structures telles que les cadres de discours, qui sont susceptibles de participer au phénomène de combinaison d'un thème « principal » avec un ensemble de facettes qui le situent relativement

à des notions plus ou moins spécifiques à un domaine de connaissance.

5.3 Le thème comme objet sémantique

Nous avons jusqu'ici très peu abordé la nature *concrète* des objets que nous avons qualifiés de thèmatiques. Cela est en effet impossible tant que l'on s'en tient au niveau très général qu'impose la revue transdisciplinaire entreprise dans cette première partie, et nous nous en sommes jusqu'ici tenu aux notions propres à chaque domaine ou discipline. Dans le contexte « documentaire » un thème correspondra de façon très évasive à un concept ou à un terme, ou encore à une entrée dans un thesaurus. Dans le contexte linguistique, il s'agira plutôt de référents du discours, notion mieux définie mais qui peut néanmoins recouvrir des réalités assez diverses. Dans les domaines du traitement automatique des langues et de la recherche d'information, il s'agira le plus souvent d'un ensemble de formes de surface numériquement représentatives, directement extraites du texte analysé.

Dans le cas de l'analyse automatique, on peut à première vue distinguer deux catégories de procédés. Les premiers, qui appartiendraient plutôt au domaine de la recherche d'information au sens strict, sont spécifiquement dédiés à la tâche d'indexation de documents entiers, et visent donc effectivement à produire une certaine représentation de l'à propos de ces documents (cf. section 1). Ce dernier est alors représenté le plus souvent par un ensemble de descripteurs qui ne sont autres que des formes extraites des documents où elles ont été jugées thématiquement saillantes. Les seconds concernent plutôt le traitement automatique des langues et la segmentation thématique intra-documentaire (cf. section 4). Dans ce cas, un problème pour leur application à la recherche d'information est que la production d'une représentation des thèmes des segments délimités est rarement une fin en soi mais plutôt un artefact de la méthode employée. Dans le cas du *text-tiling* par exemple, il s'agira des vecteurs de valeurs numériques associées aux descripteurs représentés dans chaque segment. Dans d'autres cas, la méthode se limite à la segmentation proprement dite et ne fournit aucune représentation exploitable des thèmes de segments. Plus généralement, on remarquera une forte prédominance dans ce domaine des méthodes fondées sur l'étude de la distribution des formes de surface, ce qui revient souvent à appliquer au niveau intra-documentaire des méthodes équivalentes à celles développées au niveau documentaire dans le domaine de la RI.

On aura compris à la lecture des précédentes sections que nous visons plutôt à nous éloigner des formes de surface, pour nous attacher autant que possible à la réalité sémantique des thèmes. Précisons de prime abord qu'il ne s'agit pas pour autant de s'appuyer systématiquement sur des ressources généralistes et/ou exhaustives. Nous tenterons au contraire de tirer parti d'un ensemble réduit et facilement constituable de ressources spécifiquement adaptées à un domaine de connaissance, en visant la représentation des thèmes par des *structures symboliques* à la fois porteuses de sens et accessibles au calcul.

Les limites des approches uniquement fondées sur l'analyse des formes de surface souffrent en effet des limitations bien connues dont nous avons déjà fait mention dans le chapitre 1, et qui ne sont bien sûr pas propres au problème de l'analyse thématique. Elles ne requièrent certes aucune ressource externe ni analyse profonde, mais elles ne peuvent pas prendre en considération la dimension fondamentalement sémantique du phénomène de cohésion lexicale sur lequel elles reposent plus ou moins explicitement, et dont la répétition exacte des mêmes formes de surface ne constitue jamais qu'un épiphénomène. Mais elles peuvent aussi se révéler totalement inefficaces face aux phénomènes d'ordre thématique qui ne relèvent pas de la cohésion lexicale. C'est par exemple le cas des cadres de discours (cf. section 3.4.1), le propre de la portée des introducteurs étant de faire valoir un critère sémantique sur des segments potentiellement larges sans nécessiter aucun effet de récurrence ni aucune chaîne référentielle. Des phénomènes similaires pourront être observés autour de l'architecture textuelle et notamment des titres.

De façon plus générale, il paraît peu imaginable de procéder à une analyse thématique fine sans tenir compte de la variété des phénomènes linguistiques en jeu, qui incluent des phénomènes sémantiques qui ne sont pas perceptibles sans représentations symboliques ni ressources adéquates, ainsi que de multiples modes d'organisation discursive dont la cohésion lexicale n'est qu'une instance parmi d'autres.

Pour cette raison, et en étant bien conscient des contreparties sur lesquelles nous reviendrons plus loin, nous privilégierons ici les procédés qui s'appuient sur une approche sémantique de la notion de thème. Cela n'exclut en rien les méthodes distributionnelles et numériques dont l'expérience a montré qu'elles pouvaient se révéler très complémentaires des méthodes plus symboliques et/ou linguistiques. Mais nous ne considérerons leurs résultats comme une information parmi d'autres.

5.4 Conclusion

Les travaux que nous présenterons dans la partie suivante s'appuient, à des degrés divers, sur les principes que nous venons d'esquisser, que nous pouvons résumer ainsi :

1. Le thème comme de *point de contact* entre la recherche d'un utilisateur et l'information véhiculée dans un document.
2. Le thème comme objet *sémantique* et *structuré*, en relation directe avec une certaine organisation des connaissances propres à un utilisateur, un groupe d'utilisateurs ou un domaine.
3. Le thème comme objet *discursif*, susceptible de décrire le contenu informationnel d'objets textuels de différents niveaux de grain.

Cette perspective nous conduira à prendre en compte à la fois les marques textuelles et des ressources relatives à un domaine pour procéder à la segmentation thématique du discours. Nous envisagerons également la façon dont ces mêmes ressources peuvent intervenir dans la description des segments ainsi délimités sous une forme structurée. Enfin, nous envisagerons différentes structures discursives dont l'analyse semble pertinente dans ce contexte.

Il va de soi que la prise en compte de ressources d'ordre ontologique et la mise en oeuvre d'analyses linguistiques profondes nous positionne *a priori* dans un contexte applicatif particulier : il ne sera pas question d'appliquer directement ces méthodes à des bases documentaires trop « ouvertes », ni *a fortiori* sur Internet. Mais ce type de méthodes peut néanmoins trouver des applications très pertinentes dans des contextes plus fermés mais tout aussi importants. On trouvera beaucoup de bases documentaires de ce type dans des systèmes d'information d'entreprise, institutionnels ou autres, et il est clair que des systèmes de RI adaptables à leurs besoins particuliers peuvent rendre des services non négligeables.

En outre, la notion de thème semble si complexe, et est encore si mal comprise qu'il semble illusoire d'imposer des contraintes trop fortes quant à l'analyse thématique automatique, et notamment de se priver de toute ressource spécifique. On peut certes chercher à développer des outils immédiatement applicables à grande échelle, mais on peut craindre dans ce cas de ne pas progresser significativement par rapport aux résultats obtenus par les méthodes existantes. Bien que laissant une part très importante à l'expérimentation sur corpus, notre démarche se veut peut-être plus fondamentale, dans la mesure où elle vise autant à mieux comprendre ce que recouvre la notion de thème que des applications directes à la recherche d'information.

Toutefois, la problématique qui ressort des différents axes dont nous avons fait état ici est évidemment trop large pour que nous puissions la couvrir intégralement. Nous nous concentrerons donc sur différents problèmes qui ont plus particulièrement attiré notre attention, sans chercher à obtenir une méthode d'analyse thématique dont la couverture serait aussi large que celles des méthodes plus

« naïves ». Nous décrirons ainsi, dans la partie suivante, différents procédés qui devraient prendre part à un système plus global d'analyse sémantico-discursive pour espérer obtenir une couverture totale des documents traités.

Nous nous pencherons tout d'abord sur le problème spécifique de la recherche d'information géographique, qui a pour avantage de faire apparaître très clairement l'intérêt de l'approche que nous venons de décrire. Nous verrons notamment que les facettes spatiales et temporelles revêtent une importance particulières dans ce type d'information, et comment elles interviennent dans l'analyse des textes et le processus de recherche d'information en général.

Nous détaillerons par la suite le problème de l'analyse automatique des cadres de discours spatio-temporels, qui s'est initialement posé à nous dans le contexte de l'information géographique mais qui a finalement abouti au développement d'une méthode d'analyse totalement indépendante.

Nous développerons enfin une méthode d'analyse thématique fondé sur la notion de *thème composite*, où nous explorerons plus largement les possibilités d'analyse thématique discursive exploitant la notion de facette, ou plus exactement la notion *d'axe sémantique* que nous expliciterons le moment venu.

Deuxième partie

Modèles et systèmes d'analyse

Cette partie est consacrée à la description de systèmes et de modèles de traitement automatique des langues s'appuyant sur les principes que nous venons de dégager. Nous décrirons tout d'abord des travaux portant sur la **recherche d'information géographique**, réalisés au sein du projet GeoSem. Nous présentons nos apports à ce projet, qui concernent tout d'abord l'analyse sémantique des expressions temporelles, tâche qui vise à reconnaître et marquer automatiquement ces expressions dans les textes, et surtout à calculer une représentation symbolique de leur valeur sémantique. Nous évoquerons d'autre part le problème de la mise en relation en discours des références à des faits socio-géographiques et de leur localisation spatio-temporelle, tâche qui constitue une certaine forme d'analyse thématique visant à établir une indexation par passages sous la forme de triplets « phénomène-espace-temps ». Nous présentons enfin le prototype de moteur de recherche que nous avons conçu pour exploiter les résultats obtenus et permettre l'expérimentation *in situ* des différentes techniques développées dans la cadre du projet.

Nous traiterons par la suite des deux axes de recherche qui ont découlé de ce projet, et auquel nous nous sommes tout particulièrement intéressé. Le premier concerne l'**analyse automatique des cadres de discours temporels**, basée sur l'hypothèse de l'encadrement de Michel Charolles, qui décrit des segments dits « cadres de discours », homogènes par rapport à un critère sémantique (en l'occurrence une localisation temporelle) spécifié par une expression détachée en initiale de phrase dite « introducteur de cadre ». Les cadres sont présentés comme des marqueurs d'indexation permettant de « répartir les contenus propositionnels dans des blocs homogènes relativement à un critère spécifié par le contenu de l'introducteur ». Nous proposons une méthode développée en collaboration avec l'ERSS qui, en s'appuyant sur différents pré-traitements tels que l'étiquetage morpho-syntaxique ou l'analyse sémantique des expressions temporelles, permet de reconnaître automatiquement les bornes des cadres temporels, en proposant un début de solution au délicat problème de l'analyse automatique de la portée des introducteurs. Nous décrivons le procédé d'évaluation que nous avons commencé à mettre en oeuvre pour tester les résultats obtenus, ainsi que les premières conclusions, encourageantes, que l'on peut en tirer.

Le second concerne la notion de **thème composite**, que nous avons développée dans le prolongement des travaux précédemment décrits, et qui constitue un modèle pour l'analyse thématique d'une certaine variété de structures discursives liées à la notion d'univers de discours. Nous présentons le modèle en lui-même avant de décrire la méthode d'analyse thématique automatique qui en découle. Nous envisageons les liens entre ce modèle et des concepts existants tels que l'encadrement du discours, la structure informationnelle ou encore la théorie de la structure rhétorique, avant d'introduire la notion d'axe sémantique que nous posons comme pivot entre l'organisation des connaissances d'un domaine et la structure thématique des textes qui s'y rapportent.

On notera que l'ensemble des procédés décrits dans cette partie étant implémentés à l'aide de **LinguaStream**, le lecteur pourra juger utile de consulter par avance certains éléments de la partie dédiée à cette plate-forme, et plus précisément le chapitre 9 qui en donne un aperçu général.

Chapitre 6

Recherche d'information géographique

Ce chapitre a été pour partie rédigé « à deux plumes » avec Patrice Enjalbert, pour publication dans (Enjalbert, 2005c).

Les dernières années ont été le témoin d'un intérêt croissant pour l'information géographique, c'est-à-dire pour l'accès à des données de nature économique, sociale, politique aussi bien que physique, associées à des territoires. Cela est dû pour partie au développement des systèmes d'information géographiques (SIG), bases de données spécifiquement adaptées au traitement d'entités spatialisées telles que des circonscriptions administratives, des infrastructures de communication, ou des entités physiques (rivières, reliefs...) qui sont décrites grâce à un système de coordonnées terrestres (on parlera d'entités géo-référencées). À ces entités peuvent être associées divers types de données, par exemple de nature socio-économique (Laurini et Milleret-Raffort, 1993).

Il existe toutefois une autre source d'information, elle aussi de plus en plus aisément accessible grâce aux technologies du Web. Il s'agit du document géographique, qui propose une analyse des données « brutes » (éventuellement fournies par les SIG) restituée sous la forme symbolique de textes et/ou de cartes (et autres représentations graphiques). Ce type de document est aujourd'hui massivement produit et utilisé par les administrations, publiques ou privées, qui sont chargées d'administrer ou de gérer un territoire, et bien sûr par les géographes eux-mêmes. L'objectif général des travaux ici présentés est de développer des outils permettant d'accéder le plus aisément, et de la manière la plus pertinente possible, à cette forme particulière d'information.

Le document géographique pose des problèmes très intéressants à la recherche d'information. En premier lieu, il requiert un traitement spécifique des expressions qui expriment une localisation spatiale ou temporelle : une recherche par mots-clés est clairement inadéquate dans ce cas, et une analyse sémantique est nécessaire pour pouvoir identifier dans le texte la mention des « zones spatio-temporelles » qui intéressent l'utilisateur. D'autre part, nous sommes très fréquemment en présence de documents longs ou très longs, et donc dans une problématique de recherche de passages – plutôt que de documents – nécessitant une analyse discursive des textes. Or cette analyse, nous allons le voir, est d'une certaine manière facilitée par la structure même de l'information géographique.

Les travaux présentés dans ce chapitre ont été réalisés dans le cadre du projet GeoSem, du programme CNRS interdisciplinaire « Société de l'information », et ont fait l'objet d'une collaboration entre le GREYC, l'ERSS (Toulouse), le laboratoire ESO (Caen), et l'EPFL (Lausanne) (Enjalbert, 2005b). Notre contribution à ce projet concerne principalement les points suivants, qui seront détaillés dans ce chapitre à la suite d'une présentation générale du projet :

– l'analyse sémantique des expressions temporelles ;
– la localisation spatio-temporelle des phénomènes en discours à des fins d'indexation multi-

dimensionnelle (c'est à dire incluant des critères spatiaux et temporels) ;
– la conception d'un prototype de moteur de recherche exploitant cette indexation.

6.1 Présentation générale du projet

6.1.1 Corpus et tâche

L'application vise un type de document particulier, caractérisé par une forte composante en géographie humaine. Un ensemble de traits spécifiques en font un cas très intéressant pour l'expérimentation de procédures avancées de recherche d'information.

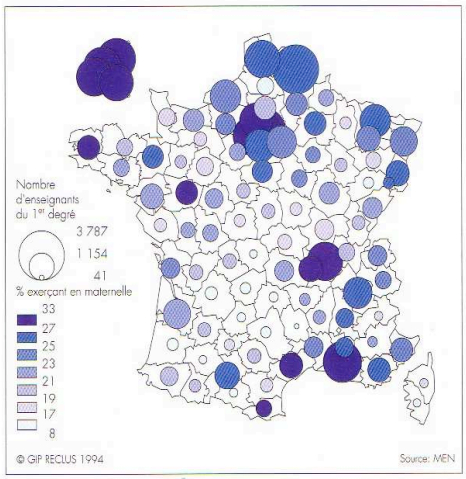
En premier lieu, la spécificité de l'information géographique est de relier un phénomène observé (quoi) à une localisation géographique (où), ainsi que, très souvent, à une certaine période temporelle (quand) à propos de laquelle le phénomène est décrit ou analysé. Cette propriété se projette sur le texte et contribue très fortement à le structurer, comme on pourra s'en rendre compte dans l'extrait de la figure 6.1. Corrélativement, une requête naturelle de recherche documentaire portera sur un triple critère phénomène-temps-espace (dorénavant P-T-E) : « Où puis-je trouver des informations sur tel phénomène dans tel espace à telle période ? », l'une des composantes étant évidemment susceptible de faire défaut. Or, d'après nos observations, l'expression de la localisation spatiale et temporelle dans les documents géographiques présente une très forte régularité, propice à une analyse linguistique « raisonnablement générique ». Développer des analyseurs sémantiques permettant une interrogation sur une conjonction de critères spatiaux et temporels – joints à des critères « thématiques » usuels – devient alors un objectif réaliste.

Une autre caractéristique du corpus visé est de comporter fréquemment des textes longs ou très longs tels que des rapports administratifs, des études socio-géographiques universitaires, ou encore des « atlas thématiques »[1]. En conséquence, le retour d'un document entier en réponse à la requête de l'utilisateur serait d'un intérêt limité ; au contraire, nous voudrions pouvoir indiquer un ensemble de passages du texte (paragraphes, sections, ou encore segments délimités dynamiquement). Ce phénomène est encore accentué par la nature multidimensionnelle (P-T-E) de la requête qui en accentue la sélectivité. Autrement dit, nous nous plaçons dans une perspective de délimitation et de caractérisation sémantique d'unités documentaires.

Notons une autre propriété du document géographique qui est d'être de nature composite, avec de nombreuses cartes (ainsi d'ailleurs que d'autres représentations graphiques, histogrammes et tableaux statistiques divers) comme illustré par l'extrait de la figure 6.1. Le projet GeoSem comporte donc un volet visant à l'interrogation de ces composantes graphiques et cartographiques du document, ouvrant sur d'intéressantes questions de sémiotique graphique et de collaboration entre composantes textuelles et graphiques. Nous avons par exemple envisagé dans (Widlöcher *et al.*, 2004) la possibilité de mener en parallèle des analyses du discours et des cartes associées pour améliorer le traitement des relations de contraste ou de similarité. Cet aspect de notre travail restant toutefois très secondaire, et ne sera pas développé ici.

Résumons cette première présentation par quelques exemples. Supposons que l'on trouve dans un texte géographique la phrase : « Jusqu'au milieu des années 1980, dans l'Aveyron, à Paris ou dans les Pyrénées-Atlantiques, seulement un enfant de 6e sur trois est en retard scolaire ». Une requête concernant « le retard scolaire dans le Midi vers 1970 » devra sélectionner le passage indiqué, mais pas une requête sur le même phénomène social portant sur l'Ouest ou les années 1990. Ou encore, la thématique générale du document dans son ensemble étant connue (ici, la scolarisation en France) une

[1] Précisons qu'un atlas n'est pas ici un simple recueil de cartes, mais un document, composite, comprenant à la fois du texte et des cartes ou autres graphiques, décrivant l'inscription d'un phénomène socio-économique dans un espace géographique.



§ [. . .] alors que, dans les départements les plus urbanisés, un instituteur sur quatre, voire sur trois, enseigne en école maternelle et moins de deux sur trois en école primaire. Trois enseignants sur quatre exerçant dans le premier degré de l'enseignement public sont des femmes, proportion qui n'a guère augmenté depuis les années 1960. Dans les écoles maternelles, la part des hommes ne cesse de croître mais reste minime (4%). En école primaire, la moyenne est de deux institutrices pour un instituteur, les instituteurs étant, relativement, un peu plus nombreux dans les postes de directeurs et dans les classes spécialisées.

§ La féminisation du corps enseignant du premier degré est très variable selon les départements : un instituteur pour deux institutrices dans le Pas-de-Calais, en Ariège et en Lozère, mais un pour quatre, voire pour cinq, dans les départements parisiens. En règle assez générale, la féminisation du premier degré est d'autant plus grande que l'urbanisation est importante, les vieilles régions industrielles et le Languedoc faisant exception.

§ Dans l'enseignement public, un instituteur sur trois a plus de 45 ans, un autre moins de 30 ans. L'âge moyen est de 38 ans, tant pour les femmes que pour les hommes, mais avec des écarts d'un département à l'autre. Les instituteurs en fin de carrière [. . .]

Source : HER

FIG. 6.1 – Une page composite (texte et carte) du corpus (HER). Croisement de localisations spatiales et temporelles avec différentes thématiques.

requête pourra porter sur une région et/ou une période d'intérêt : « la région parisienne », « le Midi et les années 1950 », « vers 1970 », etc. Le système devra retourner les segments textuels (et les cartes) associant ces diverses caractérisations.

6.1.2 Architecture générale

Le système de recherche d'information est organisé en deux phases tout à fait classiques. La première (réalisée *off-line*) opère une analyse du document afin d'en extraire des descripteurs adéquats, selon les trois axes P-T-E, reliés à des unités documentaires bien délimitées. La seconde phase (*on-line*) est constituée par l'exploitation de cette indexation pour répondre à la requête d'un utilisateur.

Concernant la première phase, on distingue deux types d'expressions linguistiques, donnant lieu à deux types de traitement. Le premier est constitué par des groupes prépositionnels (éventuellement nominaux) exprimant une localisation spatiale ou temporelle : « dans le Sud-Ouest », « dans le départements ruraux de l'Ouest », « dans les zones maritimes », « de 1965 à 1985 », « au milieu des années 1960 », etc. Ces expressions sont repérées et font l'objet d'une analyse syntaxique et sémantique ; le résultat est une représentation symbolique, manipulable calculatoirement, de l'aire géographique ou de la période temporelle dénotée, intégrée au document grâce à un marquage XML. Comme nous le verrons dans une prochaine section, notre travail sur ce point a concerné l'analyse des expressions temporelles, sur laquelle nous reviendrons dans la section 6.2

Le second type d'expressions est constitué par les groupes nominaux relevant du phénomène sociologique étudié lui-même : « le nombre de lycéens », « l'allongement des scolarités », « la féminisation du corps enseignant du premier degré », etc. Leur diversité ne permet bien sûr pas, dans le cas général, de procéder à une analyse sémantique approfondie comme pour l'espace et le temps, et nous adoptons ici une démarche de recherche documentaire plus classique, couplée avec des méthodes terminologiques et de segmentation thématique sur lesquelles nous reviendrons plus loin. Nous ferons par ailleurs état dans le chapitre 8 de la notion de « thème composite » que nous avons développée dans le but d'apporter des améliorations sur ce point, et qui peut être conçue comme une généralisation des triplets P-T-E allant bien au-delà du document géographique.

Étant donnés ces deux types d'expressions, il nous faut ensuite déterminer si telle ou telle information thématique est relative à telle ou telle zone spatiale et/ou période temporelle (mise en relation du quoi, du où et du quand) : par exemple dans l'extrait de la figure 6.1, l'augmentation des effectifs lycéens est reliée à la période 1965-1985 et un ensemble de zones géographiques comprenant le Sud-Ouest, le Massif Central, etc. On obtient ainsi une indexation de segments bien délimités du texte (unités documentaires), selon les trois « axes » phénomène, espace et temps. Ce problème sera discuté dans la section 6.3, et prolongé dans le chapitre 7 où nous traiterons de l'analyse des cadres de discours spatio-temporels qui est mise à contribution sur ce point.

Une fois réalisé ce travail « préparatoire », la seconde phase des traitements exploite l'indexation obtenue pour sélectionner et retourner des passages du texte en réponse à une requête de l'utilisateur du type de l'extrait figure 6.2. La requête est formulée en langue naturelle et analysée selon les mêmes principes. On obtient ainsi un codage symbolique des critères spatiaux et temporels exprimés par l'utilisateur, qui peut être comparé aux structures correspondantes extraites du texte pour sélectionner les passages pertinents. Cette procédure est mise en oeuvre par un moteur de recherche spécifique, que nous décrirons dans la section 6.4.

? Retard scolaire dans l'Ouest de la France dans les années 1950
? Évolution de la scolarisation entre 1960 et 1970
? Politiques de sécurité maritime dans la Manche
? Évolution du vote FN dans le Grand Ouest entre 1981 et 2002

FIG. 6.2 – Exemples de requêtes sur une base de documents géographiques.

6.2 Analyse sémantique des expressions temporelles

Les expression temporelles ici considérées sont des syntagmes nominaux ou prépositionnels, ou encore des adverbes, référant soit à des périodes (« à partir du début des années 1950 », « à la rentrée 1981 », « entre les années 1950 et 1980 »...), soit à des durées quantifiées (« en 30 ans », « pendant 10 ans »...). Dans tous les cas, il s'agit d'expressions portant sur un temps historique, donc d'une granularité variant entre le jour et le siècle. Notre but est de repérer et d'analyser ces expressions de manière à construire une représentation symbolique exploitable pour l'indexation du document. Il ne s'agit donc pas de traiter toutes les questions liées à la temporalité, mais de nous concentrer sur un type particulier d'expressions temporelles qui sont explicitement datées, c'est-à-dire directement reportables sous la forme d'un point ou d'un intervalle unique sur l'axe du temps. Nous écartons donc *a priori* celles qui demanderaient, pour diverses raisons, des inférences plus complexes pour être exploitées, comme « depuis lors », « tous les ans », « après la deuxième guerre », etc. Nous adoptons ici une démarche « classique » en deux étapes :

– une étape compositionnelle, produisant une expression formelle représentant une « sémantique abstraite », codée sous forme de structure de traits ;
– une étape d'interprétation contextuelle et référentielle, qui produit un intervalle entre deux dates.

La première est réalisée par une grammaire locale d'unification décrivant à la fois la structure syntaxique des expressions concernées et la forme sémantique associée à chaque type de construction. Elle se fonde bien sûr sur un découpage préalable en mots, mais ne fait pas appel à un étiquetage morphologique. Elle incorpore en effet, outre un lexique des mots spécifiques aux expressions temporelles (nom des mois ou des saisons, « années », « siècle », etc.), un lexique des mots grammaticaux nécessaires (essentiellement prépositions et déterminants).

La structure de cette grammaire est fortement récursive, puisqu'on observe une forte propension des expressions temporelles à se *composer* à partir d'expressions « atomiques » et de ce que l'on pourrait appeler « opérateurs ». Les expressions atomiques sont celles qui constituent, au regard du grain ici considéré, une localisation plus ou moins précise mais explicite, comme « le 20 mai 1976 », « le mois de juin 2006 », « au printemps 91 », « les années 80 », « le XXe sièce », etc. La construction de ces expression est donnée par des règles séquentielles relativement simples, et leur sémantique sera exprimée soit sous la forme d'une date, soit sous la forme d'un intervalle entre deux dates pour les expressions qui correspondent intrinsèquement à une période.

Étant données ces structures élémentaires, on peut distinguer deux principaux modes compositionnels permettant de construire des expressions complexes :

– Un premier mode consiste à *construire un intervalle* à partir d'une ou deux localisations directes. Dans ce cas, on doit prendre un compte une certaine variété de connecteurs du type « entre X et Y » ou « depuis X jusqu'à Y », ainsi que la possibilité de trouver des intervalles à demi ouverts avec des expressions telles que « depuis X » ou « avant Y ». La structure sémantique résultante est bien sûr elle-même obtenue par composition des structures décrivant les « opérandes » des connecteurs détectés.
– Un second mode consiste à *sélectionner un sous-intervalle* à partir de celui donné par une ex-

pression élémentaire. Il s'agit typiquement d'expressions telles que « dans la première moitié du siècle » ou « au début des années 90 ». Là encore, la structure sémantique produite l'est par composition, et non par réduction de la structure de l'opérande : par exemple dans le cas du « début des années 90 », on se contente d'ajouter un opérateur plutôt que de réduire directement l'intervalle décrivant les années quatre-vingt-dix proprement dites, ceci afin de laisser à la phase d'interprétation le soin de décider ce à quoi correspond « le début », ou même de pouvoir retarder ce choix jusqu'au moment de la phase de requêtage.

Finalement, les structures obtenues (avant interprétation) sont à l'image de celle reproduite ci-dessous, qui décrit l'expression « du milieu des années 80 jusqu'à la fin du siècle » :

$$\left[\begin{array}{ll} type: & \text{periode} \\ debut: & \left[\begin{array}{ll} zone: & \left[\begin{array}{ll} grain: & \text{decennie} \\ clef: & 80 \\ localisation: & \text{milieu} \end{array}\right] \end{array}\right] \\ fin: & \left[\begin{array}{ll} zone: & \left[\begin{array}{ll} grain: & \text{siecle} \\ clef: & ? \\ localisation: & \text{fin} \end{array}\right] \end{array}\right] \end{array}\right]$$

La phase d'interprétation vise quant à elle à calculer une représentation référentielle canonique sous forme d'intervalle « standard » entre des dates définies. Ceci suppose que pour chaque « opérateur temporel » soit définie une fonction de transformation d'intervalles. Par exemple, « début » va prendre le premier tiers ; un opérateur d'intervalle prend comme bornes les points extrêmes des périodes bornes ; etc. On notera que ce procédé nécessite la connaissance d'une date de référence permettant de résoudre les expressions « contemporaines » de l'écriture du texte, telles que « ce siècle » ou les notations d'années à deux chiffres (on notera toutefois que cette information constitue la seule donnée « externe » exploitée par l'analyseur). Dans le cas de l'expression ci-dessus, on obtiendra par exemple :

$$\left[\begin{array}{ll} type: & \text{periode} \\ debut: & \left[\begin{array}{ll} date: & \left[\begin{array}{ll} jour: & 01 \\ mois: & 01 \\ annee: & 1985 \end{array}\right] \end{array}\right] \\ fin: & \left[\begin{array}{ll} date: & \left[\begin{array}{ll} jour: & 31 \\ mois: & 12 \\ annee: & 1999 \end{array}\right] \end{array}\right] \end{array}\right]$$

L'analyseur a été développé sous LinguaStream à l'aide du formalisme EDCG, un extrait de la grammaire étant reproduit en annexe E.2 (seul le « coeur » de la grammaire est représenté, à l'exclusion de règles utilitaires ou relevant du lexique). La chaîne de traitement associée (permettant d'appliquer l'analyseur temporel isolément, notamment à des fins de tests), est reproduite en figure 6.3. On peut également voir dans la figure 6.4 le résultat, tel que représenté en sortie de la plate-forme, du passage de l'analyseur sur un document, ainsi qu'une autre vue du même marquage en mode « concordancier » dans la figure 6.5.

Une adaptation à l'allemand a été réalisée, dans les mêmes conditions, par Svenja Tantzen (2004), d'où il ressort, comme on pouvait s'y attendre, que les structures syntaxiques et sémantiques sont extrêmement proches. Une question importante, non encore résolue en français mais étudiée dans l'application en allemand, est le traitement de l'anaphore et des déictiques, phénomène massif sur le plan temporel dans le corpus qui nous occupe. Une autre question laissée en suspens concerne la datation relative à des événements, telles que « après la Révolution » ou « dix ans après la réforme »,

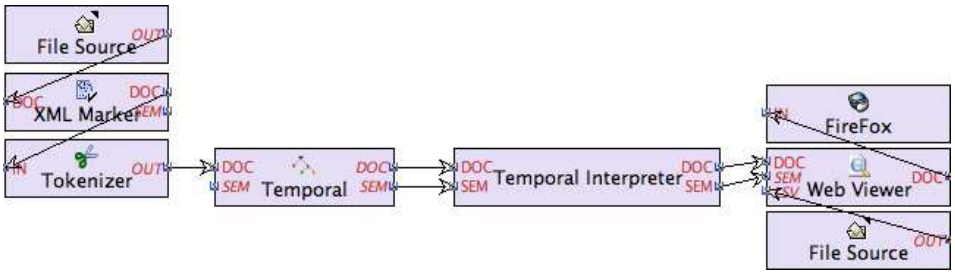


FIG. 6.3 – Chaîne de traitement LinguaStream permettant d'appliquer l'analyseur temporel. On voit que la phase de reconnaissance des expressions proprement dites est réalisée par un composant distinct (« Temporal ») de la phase d'interprétation qui lui fait suite (« Temporal Interpreter »).

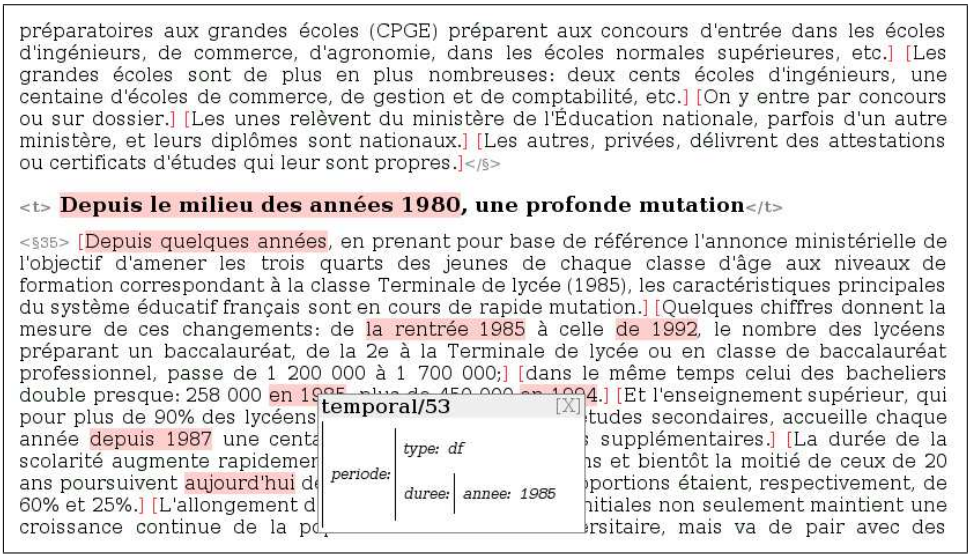


FIG. 6.4 – Résultat de l'analyse des expressions temporelles (observé dans le document lui-même).

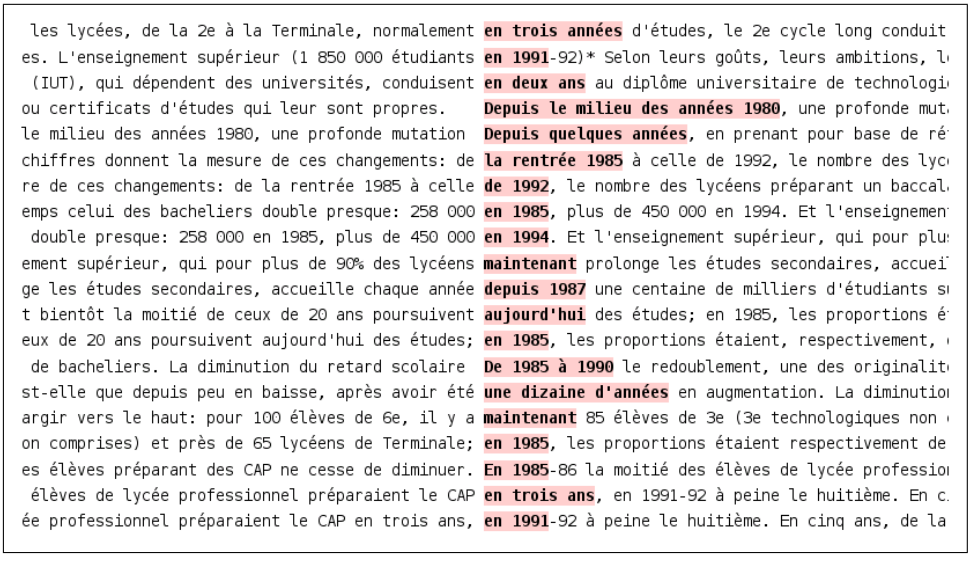


FIG. 6.5 – Résultat de l'analyse des expressions temporelles (observé en mode concordancier).

dont on devine aisément les implications en termes de connaissances « externes » et de prise en compte du contexte.

6.3 Localisation spatio-temporelle des phénomènes

Rappelons la tâche typique visée par le projet : on souhaite pouvoir répondre à des requêtes associant une composante de type « phénomène » (composante que l'on pourrait encore qualifier de « thématique » au sens propre) et des composantes spécifiant une localisation spatiale et temporelle de ce phénomène ; par exemple : « retard scolaire dans l'Ouest dans les années 1950 ». En réponse, il s'agit de présenter à l'utilisateur des extraits de la base documentaire associant ces trois critères. Insistons sur l'importance dans le cadre du projet de cette mise en relation, caractéristique du travail du géographe qui s'intéresse tout particulièrement à la répartition spatiale de formations socio-économiques, et à son évolution temporelle. Un aspect important du problème posé est donc d'être capable de repérer ces corrélations en discours.

Mais avant cela, il convient de détecter les expressions qui font référence aux « phénomènes » eux-mêmes, ce qui revient à mener une analyse thématique au sens classique. Notre approche en la matière est relativement « standard », à l'exception peut-être du fait que nous nous chercherons ici à décrire les « phénomènes » par des syntagmes complexes et non pas par de simples mots-clefs. Mis à part cela, nous appliquons des méthodes numériques habituelles que nous nous contenterons de décrire ici succinctement.

Nous supposons d'autre part que nous disposons *a priori* d'une segmentation du texte en blocs « thématiquement homogènes ». Ce qui suit est totalement indépendant de la méthode employée pour les déterminer, et il pourrait s'agir de méthodes de type *text-tiling*, d'un simple découpage en paragraphes, où encore de méthodes de segmentation telles que celle que nous proposerons dans le chapitre 8. La suite de l'analyse se place donc au sein de chacun de ces segments, et commence par déterminer quels sont les syntagmes (nominaux) les mieux représentatifs du thème de ce segment. La méthode employée à cette fin est en tout point analogue à celle décrite dans l'annexe A, dont nous nous contenterons ici de rappeler les étapes principales :

- détection des syntagmes nominaux à l'aide d'une grammaire EDCG (reproduite en annexe E.1) ;
- calcul du coefficient $tf \cdot idf$ au niveau des mots ;
- calcul d'un coefficient $tf \cdot idf$ cumulé au niveau des syntagmes eux-mêmes.

On obtient ainsi, pour chaque segment, un ensemble ordonné de syntagmes représentant les phénomènes, dont on peut ensuite chercher à examiner les relations avec les expressions spatiales et temporelles. Notons toutefois que ces deux processus – identification des phénomènes pertinents et leur localisation spatio-temporelle – ne sont pas tout à fait indépendants : nous verrons en effet plus loin que, afin de tenir compte de l'importance des composantes spatiales et temporelles dans l'information géographique, nous considérerons les localisations elles-mêmes en tant qu'indices de pertinence pour les phénomènes auxquels elles sont liées.

Nous allons maintenant décrire ce procédé de localisation spatio-temporelle des phénomènes, qui s'appuie sur les hypothèses suivantes :

- les phénomènes évoqués sont toujours localisés dans l'espace et/ou le temps, que ce soit explicitement ou implicitement ;
- les localisations spatiales et temporelles sont en interaction dans le discours, et ne doivent pas être envisagées indépendamment ;
- les localisations spatiales et temporelles sont en interaction avec la dimension thématique, et peuvent participer à l'identification des objets textuels thématiquement pertinents.

La premier volet de la méthode employée consiste à exploiter les relations syntaxiques, au sein

même de la phrase. A partir des relations de dépendance verbe-sujet ou verbe-objet (dans lesquelles le sujet ou l'objet évoque un phénomène) d'une part, et verbe-circonstanciel (spatial ou temporel) de l'autre, on pourrait établir la localisation du phénomène. Ceci pourrait être réalisé aisément si l'on disposait d'une analyse syntaxique exhaustive. Mais ce type d'analyse étant d'un coût non négligeable pour des résultats d'une fiabilité toute relative, c'est une autre méthode qui a été mise en oeuvre, exploitant un mécanisme de mise en attente : un phénomène « attend » une localisation spatiale et/ou temporelle dans la phrase, et réciproquement. Nous qualifions ces relations de « dépendances pseudo-syntaxiques », car il s'agit de liens intra-phrastiques entre les syntagmes spatio-temporels et les syntagmes lexicalisant les phénomènes, mais qui ne sont pas toujours de nature strictement syntaxique.

Ce procédé très simple permet de tenir compte du phénomène produit par la simple proximité de deux composantes dans le discours, qui nous autorise à connecter « subjectivement » différentes composantes de la phrase. On notera cependant qu'il ne revient pas seulement à connecter systématiquement les différentes composantes apparaissant dans la même phrase, puisque nous suivons ici l'ordre de la lecture et que différents critères sont appliqués pour atténuer progressivement la « portée intra-phrastique » des composantes considérées (on se reportera à (Bilhaut, 2002) pour plus de détails sur ce point).

Mais l'exploitation de liens syntaxiques intra-phrastiques est en fait tout à fait insuffisante, car nous devons faire face à des problèmes de portée de certains groupes prépositionnels spatiaux ou temporels. Observons par exemple l'extrait reproduit ci-dessous :

> De 1965 à 1985, le nombre de collégiens et de lycéens a augmenté de 70%, mais selon des rythmes et avec des intensités différents selon les académies et les départements. Faible dans le Sud-Ouest et le Massif central, modérée en Bretagne et à Paris, l'augmentation a été considérable dans le Centre-Ouest, en Alsace, dans la région Rhône-Alpes et dans les départements de la grande banlieue parisienne où les effectifs ont souvent plus que doublé.
>
> Source : HER

Ici, l'expression « De 1965 à 1985 » a une portée qui s'étend sur tout le paragraphe, et situe sur le plan temporel le jugement effectué sur l'augmentation du nombre de collégiens et lycéens dans les diverses régions mentionnées. Si le lien entre le phénomène et sa spatialisation relève bien ici de mécanismes intra-phrastiques, la localisation temporelle met clairement en jeu un mécanisme d'une autre nature. En fait, nous sommes en présence d'un phénomène absolument massif dans notre corpus (et dans ce type de textes en général), théorisé notamment par M. Charolles sous la forme de l'hypothèse de l'encadrement du discours que nous avons déjà discutée dans la section 3.4.1. Le traitement de ces configurations discursives fait donc partie intégrante du système de localisation des phénomènes ici décrit. Il fera toutefois l'objet d'un chapitre à part entière (7), et ne sera donc pas détaillé outre mesure ici.

Comme on l'a évoqué plus haut, le tissage des liens entre les phénomènes et leurs localisations permet d'affiner les pondérations attribuées aux phénomènes. Les localisations représentent en effet des éléments structurants du discours géographique, et peuvent donc être utilisées pour y distinguer les phénomènes importants. Pour cela, on attribue également des pondérations aux localisations, qui sont ensuite propagées aux phénomènes qui leur sont corrélés. Toute localisation se voit systématiquement attribuer un poids non nul, car on considère que le simple fait d'être connecté à une localisation représente pour un phénomène un indice de pertinence thématique. Il semble en effet que l'on peut raisonnablement induire cette règle de celle qui veut que, dans un document géographique, les phénomènes importants soient localisés. On applique en outre un coefficient supplémentaire aux introducteurs de cadre (leur poids initial étant proportionnel à l'envergure du cadre qu'ils introduisent). En effet, un phénomène en relation immédiate (i.e. pseudo-syntaxique) avec un introducteur de cadre se révèle

§ De la fin du siècle dernier jusqu'aux années 50, l'école primaire a été le pilier du système scolaire français. Elle inculquait les connaissances de base, lire, écrire et compter, qui serviraient toute la vie. Elle avait aussi pour mission de former les citoyens de la République. Elle délivrait le certificat d'études qui, pour le plus grand nombre, attestait de la réussite des études et marquait l' entrée dans le monde du travail. Les sessions du certificat d'études n'ont plus lieu. Nombre d'écoles communales de campagne ont été fermées, ou vont l'être, faute d'enfants à accueillir. Et l' école primaire n' est plus que le premier degré de scolarités ayant maintenant pour objectif le collège puis le lycée. La loi d' orientation de 1989 l'organise en cycles, depuis la maternelle jusqu'à la dernière année des études élémentaires ; et les instituteurs font dorénavant partie du corps des professeurs. La « communale » de Jules Ferry et de la Troisième République appartient au passé.

Source : HER

l'école primaire	de la fin du siècle dernier jusqu'aux années 1950	9.0
le pilier du système éducatif français	de la fin du siècle dernier jusqu'aux années 1950	4.2
le certificat d'études	de la fin du siècle dernier jusqu'aux années 1950	3.9
l'entrée dans le monde du travail	de la fin du siècle dernier jusqu'aux années 1950	3.0
la vie	*non localisé*	3.9
le premier degré de scolarités	maintenant	3.0
le collège	maintenant	3.0
le lycée	maintenant	3.0
les instituteurs	*non localisé*	2.0
la loi d'orientation	de 1989	2.0

FIG. 6.6 – Résultat de l'indexation d'un segment thématique. Les objets sont ici représentés textuellement, mais les résultats réels contiennent bien sûr les représentations sémantiques sous forme de structures de traits. Les valeurs numériques correspondent au poids attribué à chaque phénomène.

généralement particulièrement pertinent dans son segment. Cette configuration est par exemple très souvent employée au début d'un segment pour annoncer le phénomène localisé dont il y est question (ce que nous qualifierons plus tard de « noyau thématique »).

Le système ici décrit a été implémenté par une chaîne de traitement LinguaStream. Il produit *in fine* une indexation détaillée de chaque segment du document (ici les paragraphes, cf. *supra*), constituée de l'ensemble des phénomènes qui ont été jugés thématiquement pertinents, chacun étant associé à ses localisations spatio-temporelles s'il y a lieu. Le tableau de la figure 6.6 représente les résultats obtenus pour le segment reproduit au dessus. On notera que le processus n'a été effectivement implémenté que pour les expressions temporelles, car les calculs sur leur représentation sémantique sont facilement réalisables, contrairement aux expressions spatiales qui nécessitent l'emploi d'un système d'information pour les raisons déjà évoquées.

6.4 Moteur de recherche « sémantique et multi-dimensionnel »

Afin de concrétiser les résultats obtenus par les différents procédés développés dans le cadre du projet GeoSem, et surtout de rendre possibles des expérimentation *in situ*, nous avons développé un premier prototype de moteur de recherche capable de s'appuyer sur les indexations générées. Ce système était alors spécifiquement dédié aux besoins particuliers du projet, en reposant explicitement sur

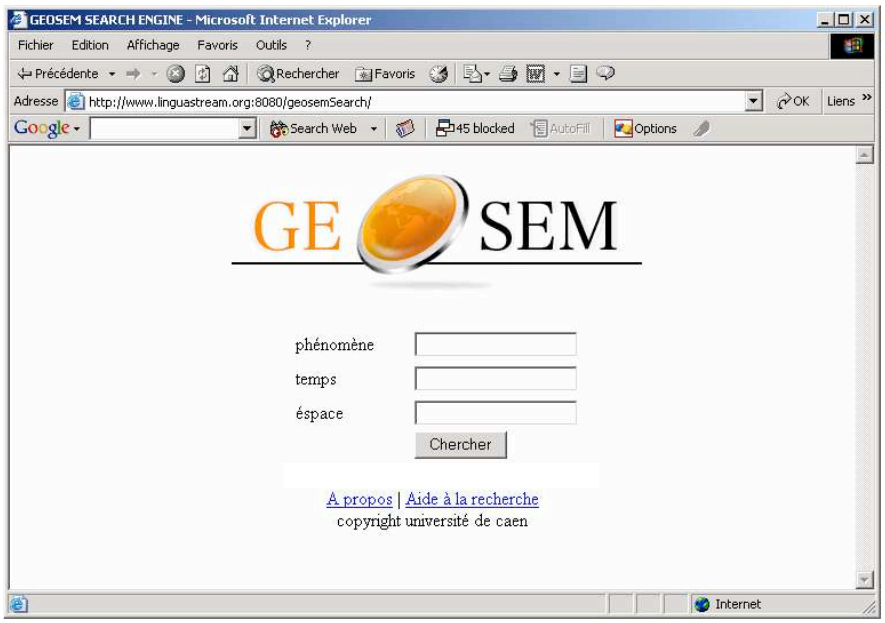


FIG. 6.7 – Interface du moteur de recherche dans le cas du projet GeoSem. Elle est générée en fonction des axes spécifique à l'application. Extrait de (Benallel, 2005).

le triplet P-T-E caractéristique de l'information géographique.

Dans un second temps, nous avons conçu un système plus général, susceptible de s'appliquer à d'autres types d'information, pourvu que le procédé d'indexation utilisé s'attache à la description du contenu informationnel de *passages*, et ce dans une optique *multi-dimensionnelle*. En tant que tel, ce moteur est immédiatement capable d'exploiter d'autres procédés d'analyse thématiques tels que celui que nous présenterons dans le chapitre 8.

La réalisation de ce seconde version du moteur a fait l'objet d'un projet de Master confié à Djamal Benallel (2005). Le résultat obtenu est actuellement entre les mains de Nicolas Hernandez qui procédera à sa mise en oeuvre effective dans le cadre d'expérimentations auprès d'utilisateurs.

6.4.1 Principe de fonctionnement

Comme nous l'avons dit plus haut, ce moteur n'est pas spécifiquement lié à un type particulier de document ou d'indexation. Son utilisation suppose donc un paramétrage visant essentiellement à spécifier quelles sont les « dimensions caractéristiques » de la base documentaire considérée. Dans le cas du projet GeoSem, il s'agit bien sûr des trois dimensions « phénomène », temps et espace (nous verrons plus loin en quoi consiste précisément la phase d'adaptation du moteur aux propriétés des documents indexés). Étant données ces dimensions, le moteur génère tout d'abord une interface de recherche en conséquence. Dans sa version actuelle, il s'agit d'un formulaire comportant autant de champs que de dimensions, l'utilisateur étant libre d'en renseigner tout ou partie (certains peuvent toutefois être déclarés comme obligatoires). L'interface générée dans le cas du projet GeoSem est reproduite dans la figure 6.7.

Les critères spécifiés par l'utilisateur le sont en langue naturelle. Pour cette raison, certains des procédés appliqués pour l'analyse des textes eux-mêmes seront à nouveau mis à contribution pour l'analyse de la requête. Il sera bien sûr possible et même souhaitable de permettre à l'avenir d'utiliser des modes d'expressions des critères qui soient spécifiques à chaque dimension. On pourra par exemple

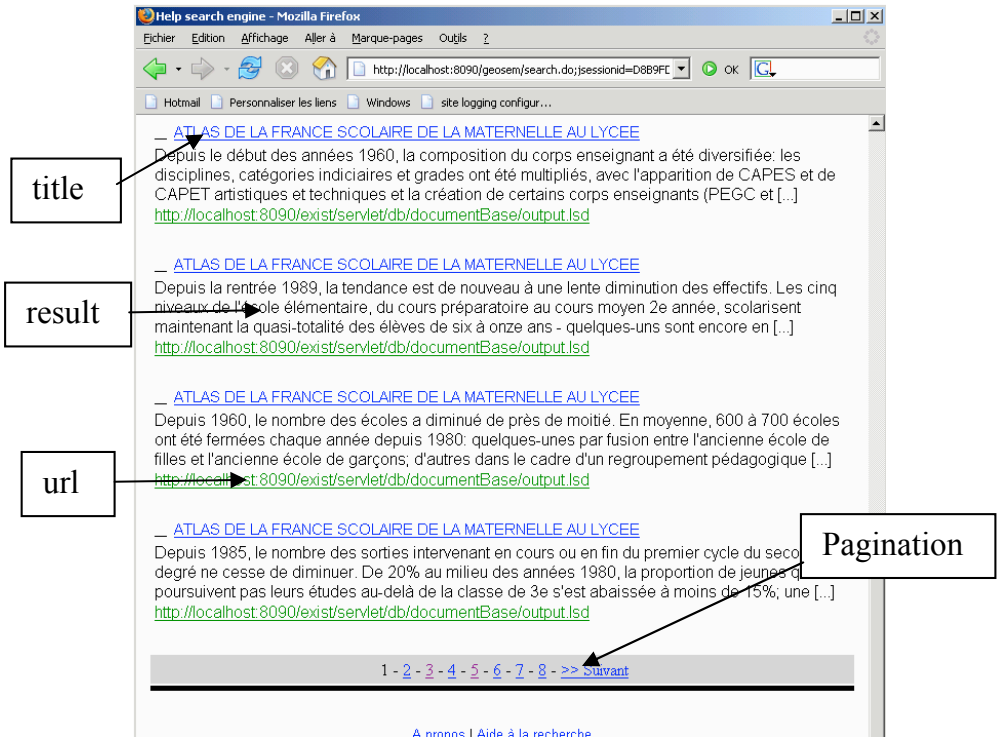


FIG. 6.8 – Résultats d'une requête tels que présentés par le moteur de recherche. Extrait de (Benallel, 2005).

mettre en oeuvre un système cartographique pour spécifier des contraintes portant sur la dimension spatiale. Cette question n'est toutefois pas prise en compte par le moteur dans sa version actuelle.

Après analyse des critères fournis par l'utilisateur, le moteur est capable d'exploiter des méthodes spécifiques pour confronter ces derniers à l'index qui lui a été fourni. Par chacun des segments candidats, il combine par la suite les résultats obtenus sur chacune des dimensions de la requête, et sélectionne finalement les plus pertinents. Les résultats sont présentés à l'utilisateur de façon classique, sous la forme d'une liste ordonnée telle que représentée par la figure 6.8. Finalement, la sélection par l'utilisateur de l'un des passages présentés permet de se rendre dans le document concerné, à l'emplacement précis où il figure.

6.4.2 Architecture

Comme nous l'avons dit plus haut, la conception du moteur suppose, sans être spécifique aux dimensions P-T-E, que les constituants de la base documentaire bénéficient d'une indexation *sémantique*, *multi-dimensionnelle*, et *par passages*. Aucune autre hypothèse n'est faite sur la nature précise des documents, pourvu qu'ils soient stockés au format XML. Un passage peut correspondre à n'importe quelle unité textuelle, même s'il sera généralement de l'ordre du paragraphe ou du groupe de phrases. Les passages analysés sont marqués dans les documents par un balisage LinguaStream, identifié par un « type d'analyse » donné (cf. partie III). Le ou les types d'analyse correspondants sont spécifié en tant que paramètre du moteur de recherche.

Les fichiers d'index fournis au moteur sont également fournis au format XML, et peuvent être produits à partir des documents analysés à l'aide d'un composant LinguaStream prévu à cet effet. Ils

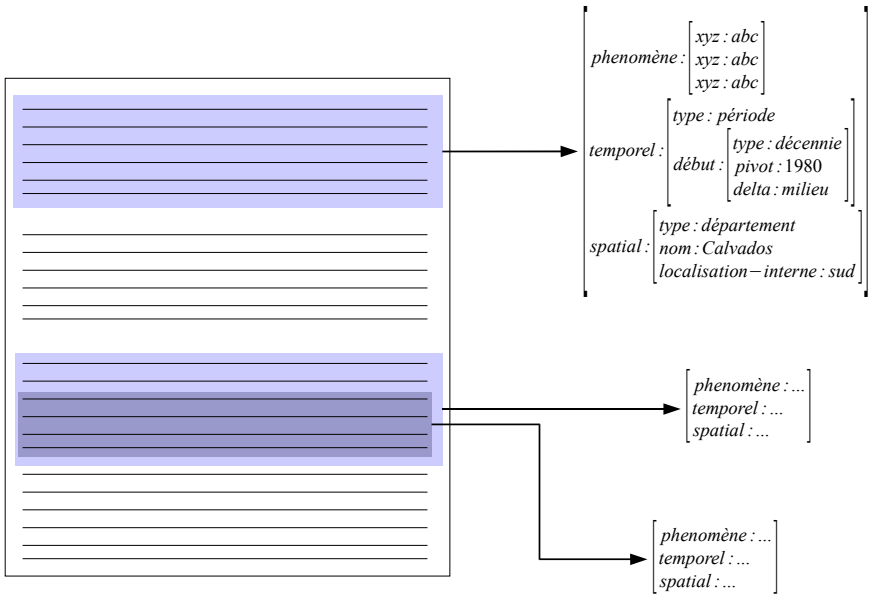


FIG. 6.9 – Indexation sémantique et multi-dimensionnelle d'un document géographique.

correspondent à une liste de segments identifiés par des liens XPointer qui pointent directement vers les passages concernés dans les différents documents de la base. Chaque passage est caractérisé par un ensemble de structures sémantiques correspondant à chacune des dimensions considérées, comme le montre la figure 6.9.

Le processus global de recherche dans la base documentaire à partir d'une requête est représenté en figure 6.10, et inclut les étapes suivantes :

- Chaque composante de la requête est analysée en faisant appel à une chaîne de traitement spécifique. Il s'agira généralement des mêmes analyseurs que ceux appliqués sur les documents eux-mêmes. Par exemple, dans le cas de l'information géographique, il s'agira des grammaires d'unification évoquées plus haut, qui permettent de reconnaître et d'analyser sémantiquement les expressions spatiales et temporelles ainsi que les syntagmes nominaux. Chacune de ces chaînes de traitement renvoie une structure de traits représentant la valeur sémantique de la composante de la requête analysée (au format XML). Durant cette phase, le moteur interagit avec la plateforme *via* son API. La spécification des différentes chaînes de traitement à appliquer fait partie de la configuration du moteur.
- L'ensemble des passages présents dans l'index est consulté, et chacune de leurs dimensions est comparée à la structure correspondante de la requête. Le résultat de cette comparaison est une valeur numérique représentant un degré de pertinence. Le procédé appliqué pour évaluer la valeur de pertinence d'une composante de la requête relativement à une composante d'un passage dépend bien sûr de chaque dimension particulière. Pour la dimension temporelle par exemple, un simple calcul d'intervalles suffira (taux de recouvrement par exemple), alors que la dimension spatiale fera vraisemblablement intervenir des connaissances géographiques (SIG ou autre). Là aussi, la spécification de la méthode de comparaison constitue un élément de la configuration du moteur.
- Finalement, les degrés de pertinence des différentes composantes de chaque passage sont combinés au sein d'une valeur unique qui permettra de les ordonner. L'importance de chaque composante relativement aux autres est spécifiée par une valeur de pondération, spécifiée là encore dans la configuration du moteur.

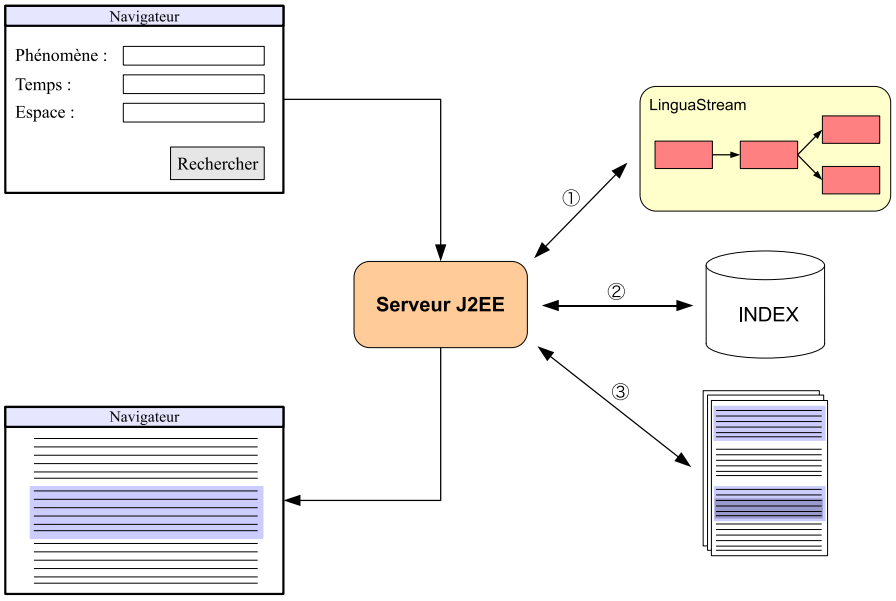


FIG. 6.10 – Processus de recherche dans la base documentaire.

On le voit, l'adaptation du moteur à chaque application particulière demande un paramétrage conséquent, incluant les éléments suivants (encore une fois sous forme de fichier XML) :

- Le ou les types d'analyse correspondant aux segments identifiés lors de l'indexation.
- La liste des axes à prendre en compte, avec pour chaque axe :
  - Un identifiant (utilisé en interne) et un nom (présenté à l'utilisateur).
  - La chaîne de traitement permettant d'analyser le critère correspondant tel qu'il est donné (en langue naturelle) dans la requête.
  - Le nom de la classe Java (implémentant une interface bien définie) permettant de procéder à l'unification des structures de traits spécifiques à cette composante.

6.4.3 Réalisation

L'ensemble du système a été réalisé sous la forme d'une application Web J2EE (Java 2 Enterprise Edition), et se conforme à l'architecture schématisée en annexe F.3. Sa mise en oeuvre s'appuie donc sur la plate-forme Java, et exploite une certaine variété de logiciels *open-source*.

La partie « interface Web » se conforme au *pattern* MVC[2], en utilisant JSP[3] et le *framework* Struts. Le prototype de test a été déployé sous le conteneur Web Apache Tomcat[4].

Le stockage des documents constituant l'index s'appuie sur la base de donnée XML native eXist[5]. Le choix de ce logiciel fait suite à une évaluation que nous avions préalablement confiée à Naima Bennai dans le cadre de son projet de DESS (Bennai, 2003).

[2]Modèle-Vue-Contrôleur
[3]Java Server Pages
[4]`http://tomcat.apache.org`
[5]`http://exist.sourceforge.net`

6.5 Conclusion

On l'a dit, l'objectif poursuivi par le projet était de développer des outils effectivement utiles aux utilisateurs visés, mais aussi et surtout de faire naître de nouvelles perspectives de recherche au sein de l'équipe. Le premier objectif est en voie d'être atteint, puisque des prototypes ont été conçus et sont sur le point d'êtres expérimentés *in situ*. Le second a également atteint à plusieurs titres. Il s'agit, pour ce qui nous concerne, de problèmes touchant à l'analyse automatique du discours, et tout d'abord celle des cadres spatiaux et temporels, qui fera l'objet du prochain chapitre. D'autre part, la généralisation des principes développés dans le cadre du projet a débouché sur la notion de thème composite dont nous traiterons dans le chapitre 8.

Chapitre 7

Analyse automatique des cadres de discours spatiaux et temporels

Nous avons présenté dans la section 3.4.1 l'hypothèse de l'encadrement du discours due à Michel Charolles (Charolles, 1997), qui décrit un mode particulier d'organisation du discours en identifiant des segments (dits « cadres de discours ») homogènes par rapport à un critère sémantique spécifié par une expression détachée en initiale de phrase (dite introducteur de cadre ou IC). Cette segmentation permet selon l'auteur de « subdiviser et répartir les informations apportées par le discours au fur et à mesure de son développement ». Nous allons ici nous intéresser à l'analyse automatique de cadres spatiaux et surtout temporels, c'est-à-dire introduits par une expression de ce type, et donc sémantiquement caractérisés dans leur ensemble par une localisation dans l'espace ou le temps.

Rappelons que cette préoccupation provient très directement de la tâche que nous nous étions fixée dans le contexte de l'information géographique concernant la localisation spatio-temporelle des phénomènes. La tâche typique visée par le projet GeoSem étant de répondre à des requêtes associant une composante de type « phénomène » et des composantes spécifiant une localisation spatiale et/ou temporelle de ce phénomène en présentant à l'utilisateur des passages associant ces trois critères, la prise en compte du modèle de l'encadrement du discours s'impose en effet d'elle-même dans le processus de mise en relation de ces différentes composantes. Observons par exemple l'extrait suivant, que nous avons déjà rencontré à plusieurs reprises :

> De 1965 à 1985, le nombre de collégiens et de lycéens a augmenté de 70%, mais selon des rythmes et avec des intensités différents selon les académies et les départements. Faible dans le Sud-Ouest et le Massif central, modérée en Bretagne et à Paris, l'augmentation a été considérable dans le Centre-Ouest, en Alsace, dans la région Rhône-Alpes et dans les départements de la grande banlieue parisienne où les effectifs ont souvent plus que doublé.
>
> Source : HER

L'expression « De 1965 à 1985 », adverbiale détachée en initiale de phrase, constitue ici un critère d'interprétation pour la proposition qui suit, à travers une localisation temporelle instaurant un univers situationnel réduit. En même temps, et c'est ce qui nous intéresse ici, elle constitue une instruction sur le plan textuel car elle peut étendre sa portée au-delà de la proposition ou de la phrase à laquelle elle appartient. Par cette extension de sa portée, elle a la potentialité de constituer un segment textuel, ou cadre de discours, homogène par rapport au critère d'interprétation qu'elle fournit. Dans le but d'exploiter ces structures discursives, une tâche spécifique a été consacrée à la question de l'analyse automatique des cadres spatio-temporels, aboutissant à l'élaboration d'une méthode opérationnelle que

nous présentons maintenant. Sur le plan linguistique, ce travail s'appuie notamment sur (Ho Dac *et al.*, 2003), (Le Draoulec et Péry-woodley, 2001) et (Laignelet, 2003).

7.1 Méthode d'analyse

La première phase de notre méthode, la détection des introducteurs de cadres, est la moins problématique. L'analyse des expressions spatiales et temporelles, présentée dans le chapitre 6, fournit l'ensemble des expressions candidates, auxquelles on attribue ou non le statut d'introducteur en fonction de critères positionnels relativement simples. Les introducteurs sont en effet caractérisés par leur position détachée en tête de phrase, séparée des propositions suivantes par une ponctuation faible (généralement la virgule). Ces critères ne sont pas absolus : on se heurte aux problèmes de la délimitation des phrases, des introducteurs multiples ou précédés d'un ou plusieurs mots, ou encore au défaut de ponctuation. On rencontre aussi des introducteurs très complexes dont la reconnaissance serait loin d'être triviale :

> Le jour même où M. Pierre Joxe, ministre de l'intérieur, annonçait le renforcement du dispositif de surveillance de l'espace aérien au-dessus de Paris (le Monde daté 14-15 août), un appareil a survolé la capitale vers 22 heures.
>
> Source : LM

Mais ces cas problématiques sont en pratique peu fréquents, et on obtient facilement de bons résultats (précision/rappel de l'ordre de 90%, performance bien sûr intimement liée à celle de l'analyseur d'expressions spatio-temporelles) (Laignelet, 2003).

Le problème de la portée, c'est-à-dire de la délimitation de la borne finale du cadre, est plus complexe. On l'a dit, il n'existe généralement pas de marque explicite, et même l'apparition d'un nouvel introducteur ne constitue pas systématiquement un indice de terminaison, en raison de l'éventualité d'une imbrication de cadres (le nouveau cadre peut être inclus dans le cadre précédemment ouvert, cas prévu par la théorie et rencontré en corpus). En pratique, le problème de la clôture des cadres est donc livré à l'interprétation subjective du lecteur, et l'expérience a montré que des divergences apparaissent entre les analyses de différents lecteurs. Cependant, l'étude linguistique fait apparaître un certain nombre d'indices exploitables de façon automatique, qui forment dans leur ensemble un faisceau raisonnablement fiable. Notons que la relative indétermination du résultat à obtenir, et donc du résultat obtenu, ne remet pas en cause ses applications en termes de recherche d'informations : une marge d'erreur de une ou deux phrases est tout à fait acceptable lorsqu'il s'agit d'extraire (ou de baliser) un passage complet.

Le faisceau d'indices ainsi élaboré est particulièrement hétérogène, puisqu'il regroupe à la fois des critères de surface, morpho-syntaxiques, sémantiques et distributionnels. On exploite tout d'abord des critères liés à la structure logique du document, et plus précisément le découpage en paragraphes. L'observation de corpus montre en effet qu'un cadre s'étend très rarement au-delà du paragraphe auquel il appartient, et cette limite peut donc être utilisée de façon statistiquement fiable quand aucun autre indice n'est venu interrompre le cadre auparavant.

Nous utilisons également un critère lié à la cohésion des temps des verbes, en nous appuyant sur une analyse linguistique de l'impact de ce paramètre sur la cohésion textuelle, (Le Draoulec et Péry-woodley, 2001), indiquant que le changement de temps verbal peut être considéré comme indice de terminaison d'un cadre temporel (l'applicabilité de ce critère aux cadres spatiaux, *a priori* beaucoup moins immédiate, resterait à démontrer). La mise en oeuvre automatique de ce test dépend évidemment de l'analyse morphologique effectuée en amont, mais la fiabilité des outils actuellement disponibles dans ce domaine est satisfaisante pour cette tâche. Notons que dans son implémentation actuelle, le

système exploite une classification particulière des temps verbaux qui consiste, pour les temps composés, à prendre en compte uniquement le temps de l'auxiliaire. On assimile ainsi le passé composé au présent, le plus-que-parfait à l'imparfait, etc., ceci afin de relâcher partiellement ce critère qui semble souvent conduire, s'il est appliqué trop strictement, à une fermeture prématurée des cadres. Le procédé d'évaluation que nous décrirons plus loin nous permettra éventuellement de valider ce procédé, ou au contraire de lui préférer une autre approche plus efficace. D'autre part, l'exploitation d'une caractérisation en termes aspecto-temporels, plus fine que le simple temps verbal, constitue une direction de recherche intéressante sur ce point. Un projet dans ce sens, exploitant les travaux de Cédric Person (2004), est actuellement à l'étude.

La méthode exploite aussi un critère de cohésion sémantique, ce que nous permettent les analyseurs sémantiques des expressions temporelles et spatiales. On peut en effet utiliser les valeurs symboliques produites par ces derniers pour évaluer la compatibilité référentielle de l'introducteur d'un cadre temporel (resp. spatial) avec les expressions temporelles (resp. spatiales) qui lui succèdent en discours. On considère l'apparition d'une expression incompatible comme un indice de rupture, puisqu'elle n'appartient pas à l'univers défini par l'introducteur. Par exemple, un cadre temporel introduit par une expression faisant référence à la période 1965-1985 (voir l'extrait reproduit plus haut) serait interrompu par l'apparition d'une expression temporelle faisant référence à l'année 2000 (qu'il s'agisse d'un autre introducteur ou non).

L'application de ce critère reste toutefois problématique. Il serait tout d'abord préférable de prendre en compte le cas où une expression sémantiquement incompatible apparaît sans pour autant provoquer la fermeture du cadre, même s'il se produit moins fréquemment que l'on pourrait le croire. Il s'agit généralement de constructions particulières (propositions relatives, incises, comparatifs, parenthèses, etc.) marquant explicitement un « échappement » temporaire de l'univers courant, par exemple pour évoquer une situation de contraste. Il faut d'autre part remarquer la difficulté de l'application de ce critère aux cadres spatiaux : l'évaluation de la compatibilité sémantique entre deux expressions spatiales est beaucoup plus problématique, notamment en raison d'une plus grande complexité du modèle sous-jacent et de la nécessité de faire intervenir des connaissances « géographiques » conséquentes (accès à un SIG par exemple).

Notre dernier critère, contrairement aux précédents, n'a pas pour fonction de détecter la fin du cadre mais plutôt d'éviter une clôture inappropriée. On se base ici sur la collaboration en discours des dimensions spatiales et temporelles, en analysant au sein d'un cadre la distribution des deux types d'expressions. Le discours géographique est en effet caractérisé par une répartition claire et univoque des informations relativement à des repères spatio-temporels, ces deux types de repères étant rarement ajustés simultanément. Par exemple, on évitera de changer d'univers temporel en même temps que l'on décrit la variation d'un phénomène d'une zone géographique à l'autre. C'est pourquoi le système détecte, à l'intérieur d'un cadre temporel (resp. spatial) les zones à fortes densités d'expressions spatiales (resp. temporelles), de façon à en éviter la clôture à l'intérieur de cette zone. Cela permet par exemple d'éviter la clôture d'un cadre temporel au beau milieu d'une énumération sur l'axe spatial.

On pourra globalement remarquer qu'un certain nombre des critères ici mentionnés semblent très rigides, et on imagine sans peine des situations où, par exemple, apparaîtrait une expression temporelle non « compatible » avec un introducteur de cadre sans pour autant interrompre sa portée. Il reste de fait à envisager comment les différents critères pourraient collaborer plutôt qu'être appliqués indépendamment, par exemple en exigeant la concordance d'au moins deux indices. Il nous faudra également envisager d'autres indices de continuation comme celui que nous mentionnions à l'instant, ou encore la possibilité de désactiver localement certains indices de clôture (cf. supra). Ces mécanismes n'étant pas encore en place, on pourra donc s'attendre à ce que l'analyseur tende à clôturer les cadres trop « tôt ». Nous verrons toutefois plus loin, sans que cela remettre en cause la validité théorique de ces remarques, que les premières évaluations que nous avons menées ne confirment pas cette intuition.

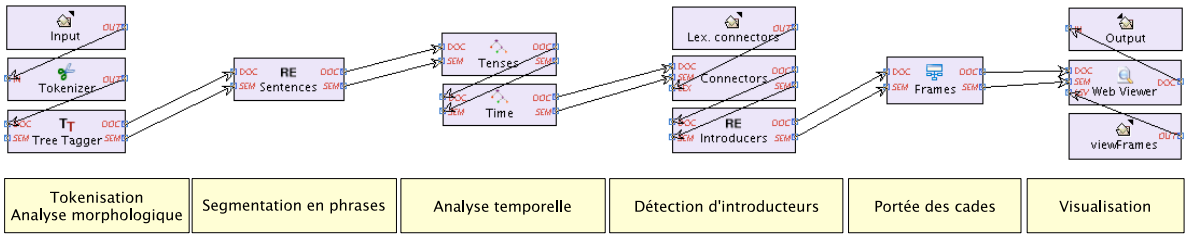


FIG. 7.1 – Chaîne de traitement de l'analyseur de cadres temporels.

MRE

```
1 {type : phrase, anchor : start}
2 <introducteur>
3 {type : connecteur}? {type : temporel} /as $t
4 </introducteur> /sem {axe : temps, valeur : $t}
5 ","
```

FIG. 7.2 – Règle de détection des introducteurs de cadres temporels.

7.2 Implémentation

L'implémentation effective du procédé d'analyse des cadres de discours concerne principalement les cadres temporels. Il s'agit en effet, dans le type de corpus considéré, d'un cas relativement favorable dans la mesure où les calculs sur les représentations sémantiques sont particulièrement aisés (calculs sur des intervalles), ne nécessitent aucune ressource externe, et que les critères linguistiques de clôture des cadres sont suffisamment efficaces. Il en va autrement en ce qui concerne l'espace : même si la mécanique informatique reste quasi identique, la complexité du modèle employé ainsi que la nécessité d'avoir recours à des données externes rendent les calculs sur les représentations sémantiques difficiles. D'autre part, nous avons pu constater que les critères linguistiques appliqués avec un certain succès dans le cas des cadres temporels sont moins efficaces dans le cas des cadres spatiaux. Différentes tâches restent donc à réaliser pour parvenir à une détection fiable des cadres spatiaux.

Comme nous l'avons dit, l'analyse de cadres temporels est, quant à elle, totalement implémentée, et ce à l'aide de la plate-forme LinguaStream. L'implémentation de ce procédé se décompose en plusieurs sous-problèmes, qui sont traités séparément en se reposant sur le principe de modularité sur lequel s'appuie notre plate-forme (cf. section 10). Nous serons également amenés à exploiter, pour chaque sous-tâche, des formalismes différents pour l'écriture des règles correspondantes. Ceux-ci seront décrits en détail plus loin (section 12), et nous nous contenterons ici de mentionner très succinctement les propriétés principales qui motivent le choix de chacun d'entre eux.

La première phase, qui concerne la détection des introducteurs, se décline elle-même en deux sous-tâchse : l'analyse des expressions temporelles d'une part, et celle des introducteurs (s'appuyant sur elle) d'autre part. Comme nous l'avons vu dans le chapitre 6, l'analyse sémantique des expressions temporelles fait l'objet d'une grammaire locale d'unification (EDCG), exprimant pour sa part des contraintes sur les résultats d'une analyse morpho-syntaxique préliminaire, et associant aux expressions reconnues une représentation de leur « sens » sous la forme de structures de traits.

Sur cette base, la détection des introducteurs peut être mise en place à l'aide de critères essentiellement positionnels. Les contraintes exprimées sont fondamentalement séquentielles, puisque nous recherchons des zones de texte vérifiant des motifs imposant la présence, dans un ordre fixé, d'éléments immédiatement successifs. Ces règles sont donc simplement exprimables à l'aide du formalisme MRE

(macro-expressions régulières). On notera qu'en plus des expressions temporelles, nous exploitons ici le marquage des phrases et des connecteurs de discours, le premier étant produit par un analyseur relativement basique (également décrit au format MRE), et le second à partir d'un lexique LSL des connecteurs de discours.

La règle complète est donnée dans la figure 7.2. Les contraintes sur les structures de traits produites en amont (entre accolades) ainsi que sur les formes de surface (ici, uniquement la virgule en fin de motif) permettent de délimiter l'introducteur. Nous détectons ainsi les éléments précédés d'un début de phrase, composés d'un éventuel connecteur de discours et d'une expression temporelle. Le reste de l'expression correspond au marquage à produire (balise type XML) et à l'annotation produits en sortie. L'élément reconnu aura le type « introducteur » et sera associé à l'annotation sémantique qui lui fait suite. Précisons que la variable *t* permet de faire « remonter » l'information contenue dans la structure de traits associée à l'expression temporelle, pour un usage ultérieur.

Pour la détermination de la portée de l'introducteur, la méthode présentée dans la section précédente s'appuie sur des critères énonciatifs tels que la cohésion des temps verbaux, sur la structuration en paragraphes, et sur des calculs sémantiques de cohérence entre l'introducteur et les autres expressions temporelles. La nature de ces contraintes diffère radicalement des précédentes. D'une part, nous pouvons désormais nous abstraire de la linéarité du texte : il est ici très simple ignorer un certain nombre d'éléments du flot textuel, sans avoir à les « consommer » explicitement. D'autre part, s'il existe bien des contraintes interprétatives entre l'introducteur et certains éléments de la zone introduite, ces contraintes n'imposent aucun ordre strict entre les éléments à prendre en compte (verbes, expressions temporelles et spatiales, éléments de la structure logique, etc.).

Pour ces raisons, les formalismes traditionnels permettant de formuler des patrons de différentes sortes (grammaires, expressions rationnelles, etc.) ne peuvent s'appliquer. Une première implémentation a donc été réalisée sous la forme d'un programme *ad-hoc*, prenant en compte l'ensemble des contraintes mentionnées plus haut. Il ne s'agit pas pour autant d'un logiciel réalisé *ab initio*, mais d'un script s'intégrant directement au système EBMS de LinguaStream[1] (cf. section 11.3). Le code correspondant est donc relativement court (de l'ordre de 100 lignes, cf. section E.4 en annexe). On notera que ce procédé a pour avantage de permettre l'incorporation de divers mécanismes de traces permettant de suivre le déroulement de l'analyse, notamment, en l'occurrence, d'associer à chaque cadre un trait spécifique permettant de mémoriser quel est le critère qui a abouti à sa clôture. On peut ainsi, à l'aide de l'outil « Grapher » de LinguaStream, générer des graphiques tels que celui reproduit en figure 7.3 où apparaissent les proportions respectives des cadres fermés par application de chacun des critères.

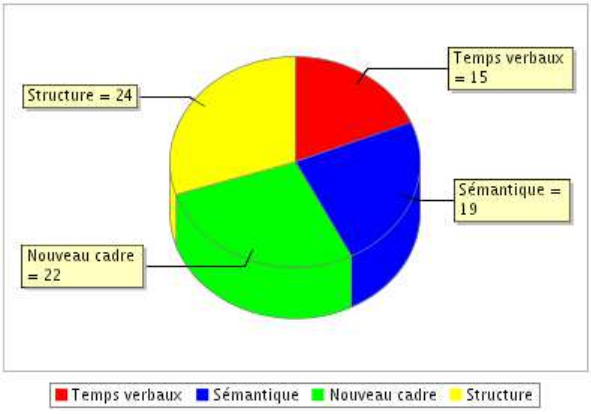


FIG. 7.3 – Critères de fermeture effectivement appliqués par l'analyseur de cadres temporels.

[1] En l'occurrence, le langage utilisé est « Groovy ».

CDML

```
Rule {type : "cadre"} :
    start({type : "introducteur"})
    end({type : "phrase"})
    homogeneity(comparator : portée)
    not presence(pattern : {type : "introducteur"}, amount : 2)
    size(mode : #LONGEST)

Comparator portée ({type : "verbe"} as $v1, {type : "verbe"} as $v2) :
    $v1/temps = $v2/temps

Comparator portée ({type : "introducteur"} as $i, {type : "tempo"} as $t) :
    ($t/debut >= $i/debut) and ($t/fin <= $i/fin)
```

FIG. 7.4 – Règle de détection des introducteurs de cadres temporels.

Il est toutefois peu satisfaisant que l'analyse de la portée soit réalisée par un script spécifique, où les contraintes effectivement appliquées sont diluées dans un code opaque, alors que toutes les autres étapes de l'analyse sont bien spécifiées sous forme de règles explicites. Pour toutes les raisons que nous discuterons dans la partie III, cette dernière façon de procéder est préférable à de nombreux égards, et il est souhaitable de pouvoir exprimer les contraintes régissant la portée de façon déclarative. Or c'est pour répondre à ce type de besoin qu'Antoine Widlöcher développe le formalisme CDML[2], particulièrement adapté à l'analyse de structures discursives. Ce système est encore en cours d'élaboration, mais le problème de l'analyse de la portée des cadres constitue l'un des cas d'utilisation choisis pour guider son développement, et une version purement déclarative a été développée à ce titre. Celle-ci ne reproduit pas encore la totalité des contraintes implémentées par la version précédente, mais évoluera à très court terme au fur et à mesure du développement de CDML.

La règle correspondante est donnée en figure 7.4. Elle décrit une unité textuelle composée de phrases complètes, commençant par un élément identifié comme introducteur et ne comportant pas d'autre élément de ce type, dont tous les verbes sont au même temps, et au sein de laquelle les expressions temporelles portent sur une plage comprise dans l'intervalle fixé par l'introducteur, en ne retenant que le plus long des candidats partageant un même introducteur.

L'agencement des différentes sous-tâches est bien sûr réalisé sous la forme d'une chaîne de traitement, reproduite par la figure 7.1. Elle comprend les éléments suivants :

- Les composants « Tokenizer » et « Tree Tagger » se chargent respectivement du découpage en mots et de leur analyse morphologique.
- Le composant « Sentences » se charge de délimiter les bornes de phrases.
- En exploitant les résultats de l'analyse morphologique, le composant « Tenses » se charge de classifier les verbes dans « les classes temporelles » choisies pour évaluer l'homogénéité des cadres de ce point de vue (cf. section 7.1).
- Le composant « Time » procède à l'analyse syntaxico-sémantique des expressions temporelles à partir d'une grammaire locale d'unification (cf. chapitre 6).
- Le composant « Connectors » projette sur le document un lexique de connecteurs de discours.
- Le composant « Introducers » correspond à la détection des introducteurs, en appliquant la règle MRE donnée en figure 7.2. Ce module exploite les résultats de plusieurs analyses antérieures : analyses morphologique, des expressions temporelles et des connecteurs de discours.
- Le composant « Frames » correspond à l'analyse de la portée des introducteurs, et délimite donc les cadres proprement dits. Dans la version la plus récente de l'analyseur, il s'agit de la grammaire CDML reproduite dans la figure 7.4. Ici, l'intégralité des informations préalablement

[2]Constraint-based Discourse Modeling Language, cf. section 12.1.3.

calculées sont utilisées simultanément.
– Les derniers composants (« Web Viewer » notamment) relèvent de la visualisation des résultats, et permettent de produire des vues telles reproduites dans la partie suivante.

7.3 Procédé d'évaluation

Le problème de l'évaluation d'un analyseur automatique tel que celui que nous venons de décrire est en soi un problème extrêmement complexe. Remarquons tout d'abord que deux grands types d'évaluation peuvent être envisagés dans ce contexte : l'évaluation dite *intrinsèque* qui vise à mesurer la qualité des résultats obtenus relativement à une annotation de référence réalisée manuellement, et l'évaluation dite *extrinsèque* qui vise à quantifier l'apport effectif des résultats obtenus dans un cadre applicatif donné, par exemple en évaluant le bénéfice apporté aux utilisateurs d'un moteur de recherche exploitant ces résultats, comme celui présenté dans la section 6.4. Il va sans dire que ce dernier type d'évaluation soulève de nombreux problèmes tant théoriques que pratiques, qui requièrent une approche largement pluridisciplinaire, notamment concernant les aspects psycholinguistiques ou même psychologiques, les problèmes logistiques, mais aussi les questions spécifiques au domaine de la base documentaire utilisée.

L'évaluation intrinsèque, à laquelle nous nous attacherons ici, est également complexe, et ce tout particulièrement dans le cas d'analyses d'ordre discursif comme celle des cadres du discours. Rappelons que même dans des domaines qui peuvent sembler assez contraints, comme celui de l'étiquetage morpho-syntaxique, ce problème est loin d'être trivial : les difficultés rencontrés concernent la constitution d'un jeu d'étiquettes assez largement admis, l'alignement entre les annotations manuelles et automatiques, la méthode de calcul de la qualité d'une annotation automatique, le coût de l'annotation manuelle et sa validité, etc.

Dans le cas des structures d'ordre discursif, d'autres problèmes viennent s'ajouter. Tout d'abord, les unités en jeu sont « élastiques » par essence : leurs dimensions sont diverses, et la quantification de leur taille pose problème (la portée d'un introducteur de cadre pourrait, par exemple, se mesurer en mots, en propositions, en phrases, etc.). D'autre part, certains types d'unités n'étant pas encore précisément caractérisés par la linguistique, on ne dispose pas toujours de critères suffisamment précis pour établir un protocole d'annotation susceptible de limiter les divergences d'interprétation entre annotateurs. C'est notamment le cas pour les cadres temporels qui nous intéressent ici, pour lesquels il n'existe pas à ce jour de critères de clôture établis, et dont le processus de « fermeture » (s'il existe) par un lecteur naïf est manifestement difficile à caractériser objectivement. Et même pour un annotateur isolé, il est parfois difficile de fixer avec certitude la borne droite d'un cadre. Ainsi, beaucoup des difficultés que nous avons rencontrées au cours du développement de la méthode d'analyse automatique elle-même se retrouvent dans le cadre de l'annotation manuelle.

D'autre part, la complexité des traitements nécessaires à l'analyse automatique de telles structure complique considérablement la tâche d'évaluation. Comme nous l'avons vu dans les sections précédentes, notre méthode s'appuie sur une chaîne de traitements complexe, où l'analyse de la portée des cadres constitue l'ultime phase, s'appuyant sur les différents résultats antérieurement obtenus. De ce fait, toute erreur dans une phase intermédiaire de l'analyse peut avoir des conséquences sur toutes les étapes qui suivent. Par exemple, la non-détection d'un cadre proviendra le plus souvent du fait que son introducteur n'a pas été reconnu comme tel, défaillance qui peut elle-même être imputable à l'analyseur temporel, etc. Bien-sûr, il s'agit là d'un phénomène très général d'accumulation des erreurs, qui touche de très nombreux systèmes, par exemple en extraction d'informations comme le souligne (Grishman, 2005), et qui n'est d'ailleurs pas spécifiquement lié à la discipline du TAL.

Ce problème n'est bien sûr pas une fatalité, et trouve des solutions par exemple en s'autorisant à chaque étape du traitement à réviser les choix préalablement effectués. Mais en tout état de cause,

il convient d'être particulièrement prudent quant aux conclusions que l'on peut tirer de l'évaluation d'une annotation ainsi obtenue. Car si l'on espère en retirer des éléments susceptibles d'améliorer le procédé d'analyse, il faut veiller à évaluer de la façon la plus indépendante possible les différentes étapes qui le composent. Ainsi, pour l'analyse des cadres, on cherchera à évaluer distinctement la qualité de la détection des introducteurs et celle de l'analyse de leur portée.

Enfin, devant la complexité des mécanismes interprétatifs en jeu dans l'instauration d'un cadre en discours, remarquons qu'il nous paraît également important de viser une évaluation qualitative, fût-elle partielle, en se donnant les moyens de constater de façon visuelle les différences entre l'analyse automatique et l'analyse de référence, sans se limiter aux données quantitatives qui ressortent de l'évaluation systématique. L'observation *in situ* au cas par cas reste en effet, comme le montre (Ferrari et Beust, 2003), un moyen très efficace de déceler les défauts d'un analyseur automatique, à condition que des outils de visualisation adéquats soient utilisés pour faire face à la quantité des données observables.

On voit donc que dans le cas qui nous occupe ici, le problème de l'évaluation intrinsèque soulève une problématique complexe, devant laquelle la communauté du TAL reste à l'heure actuelle très largement démunie. De fait, la question de l'analyse de discours dans la perspective « linguistique » que nous adoptons ici est relativement récente, et la question de son évaluation, pour essentielle qu'elle soit, reste encore très largement ouverte.

La méthode que nous allons détailler ici, développée en collaboration avec Stéphane Ferrari, Antoine Widlöcher et Marion Laignelet (Ferrari *et al.*, 2005), a autant pour vocation de faire apparaître de façon concrète certaines difficultés liées à la problématique que nous venons de poser, que de procéder effectivement à l'analyse des procédés décrits précédemment. Néanmoins, comme nous le verrons plus bas, nous pouvons d'ores et déjà en tirer des observations intéressantes quant à la qualité de nos annotations automatiques et surtout quant aux voies à explorer pour tenter de les améliorer.

7.3.1 Annotation de référence

L'annotation de référence, à laquelle nous comparerons les annotations produites par l'analyseur automatique, est bien sûr réalisée manuellement. Notre projet d'évaluation se base sur un corpus constitué de (HER), (BUL), ainsi que d'une série de numéros de (LM) publiés entre les années 1987 et 1989. Toutefois, l'annotation de cette dernière source est encore en cours de réalisation, et les résultats dont nous ferons état ici portent principalement sur l'annotation de (HER) réalisée par Marion Laignelet (Laignelet, 2003). Il s'agit néanmoins d'un ouvrage relativement conséquent (environ 56 000 mots), dont l'annotation manuelle a fait apparaître 156 cadres temporels.

Afin de tenir compte du caractère « flou » de la portée des introducteurs, le principe d'annotation que nous avons défini permet à l'observateur, en cas de doute, de spécifier plusieurs bornes de début ou de fin pour un même cadre (nous exploitons ici l'une des facilités offertes par le modèle d'annotation de LinguaStream, sur lequel nous reviendrons dans la partie III). Dans le cas particulier des cadres de discours, il est évident que cette possibilité sera peu utile dans le cas de la borne gauche, beaucoup moins ambiguë. En revanche, l'expérience de l'annotation manuelle montre qu'il est parfois difficile de prendre une décision ferme quant à l'emplacement de la borne droite. C'est pourquoi nous avons choisi de donner la possibilité à l'observateur de spécifier si nécessaire une borne droite « inférieure » ainsi qu'une borne droite « supérieure », permettant de signifier que la fin de la portée de l'introducteur se situe dans la zone ainsi marquée, sans prendre de décision ferme quant à son emplacement précis (ni même supposer que celui-ci existe). Dans ce cas, l'annotation produite signifie que le cadre ne peut pas raisonnablement être fermé avant la borne inférieure, ni après la borne supérieure, mais que toute clôture intervenant entre ces deux bornes est acceptable. C'est en ce sens que nous interpréterons, dans le protocole de comparaison avec l'annotation automatique présenté plus loin, la présence d'une

double borne droite.

Précisons qu'il reste bien sûr possible pour l'observateur de ne spécifier qu'une seule borne droite s'il considère que celle-ci peut être positionnée sans ambiguïté. Ajoutons également que nous avons délibérément écarté, mais seulement temporairement, la possibilité de spécifier non pas seulement une « zone de clôture » comprise entre deux bornes, mais un nombre arbitraire de bornes droites. Cela permettrait à l'observateur de positionner un nombre quelconque de points dans le texte où la clôture du cadre serait acceptable. Ce mode d'annotation semble de fait intéressant dans certains cas, mais ne peut se substituer totalement au mode préalablement présenté, dans la mesure où il existe effectivement d'autres cas où la spécification d'une zone reste la meilleure solution. De ce fait, afin de limiter la difficulté de notre tâche, tant en termes de coût de l'annotation manuelle qu'en termes de complexité du processus de comparaison lui-même, nous avons écarté cette possibilité dans le cadre de ce travail à vocation exploratoire, en la réservant pour de futures évolutions de la méthode.

On pourra par ailleurs s'étonner de la différence que nous avons d'introduite entre les modalités de l'annotation manuelle et celles de l'annotation automatique, qui ne pose jamais qu'une seule borne finale. On peut en effet considérer qu'il s'agit là d'un biais introduit dans le processus d'évaluation, et suggère que, si l'on accepte l'idée de la possible indétermination de la borne droite, nous devrions nous autoriser, lors de l'analyse automatique à procéder au même type d'annotation. Toutefois, il est important de garder à l'esprit que le choix de la possibilité de garder un certain flou dans l'annotation manuelle ne répond aucunement à des propriétés fixées par le modèle auquel nous nous référons, mais seulement à l'éventuelle indétermination des observateurs relativement à ce modèle. Ainsi, l'utilisation de bornes droites multiples n'est en rien une fin en soi, et la seule raison pour laquelle l'analyseur automatique devrait y avoir recours résiderait dans l'existence effective d'une indétermination entre les différentes règles appliquées lors de l'analyse automatique. Or, dans la version actuelle de l'analyseur, cette indétermination n'existe pas, et cela n'aurait aucun sens de l'introduire artificiellement. Pour ces raisons, la différence de principe entre les annotations comparées ne nous paraît pas problématique ici. En revanche, la constatation de ce type d'indétermination par plusieurs observateurs humains soulève bien entendu d'importantes questions quand au modèle de l'encadrement lui-même, sur lesquelles nous ne nous attarderons pas ici.

Remarquons enfin que l'annotation de référence de (HER) contient très peu de bornes droites doubles. Cela est tout d'abord dû au fait que la réalisation de cette annotation a été réalisée préalablement à l'instauration du mode d'annotation que nous venons de décrire (Laignelet, 2003). En conséquence, les seuls cas de bornes doubles dans cette partie du corpus ont été ajoutées dans un second temps, afin de résoudre quelques cas litigieux. On notera à ce propos que les ambiguïtés y semblent significativement moins fréquentes que dans la portion (LM), ce qui peut sans doute s'expliquer par la différence de genre.

L'annotation manuelle initiale de (HER) a été réalisée à l'aide du logiciel XXE[3], éditeur XML offrant une interface de type traitement de texte dite WYSIWYM[4], dont une copie d'écran est représentée en figure 7.5. À cette fin, nous avons développé une extension permettant d'ajouter facilement des balises LinguaStream dans un document XML, balises représentant ici les bornes des cadres de discours. Toutefois, en raison des limitations de cet outil, il n'a pas été possible d'améliorer cette extension pour la rendre capable de générer des annotations doubles ou de gérer d'éventuels chevauchements entre les cadres annotés et la structure logique du document. Plus généralement, on pourra regretter qu'à l'heure actuelle aucun outil d'édition XML soit à même de gérer de façon convaincante un modèle XML tel que celui utilisé par notre plate-forme[5], même si on peut raisonnablement espérer une amélioration de cette situation à relativement court terme.

[3] XMLmind XML Editor, cf. `http://www.xmlmind.com`.

[4] What You See Is What You Mean.

[5] Nous utilisons notamment les espaces de noms pour distinguer les annotations LinguaStream des autres balisages d'un document, ce qui n'est à ce jour correctement géré par aucun des éditeur WYSIWYM que nous avons pu tester.

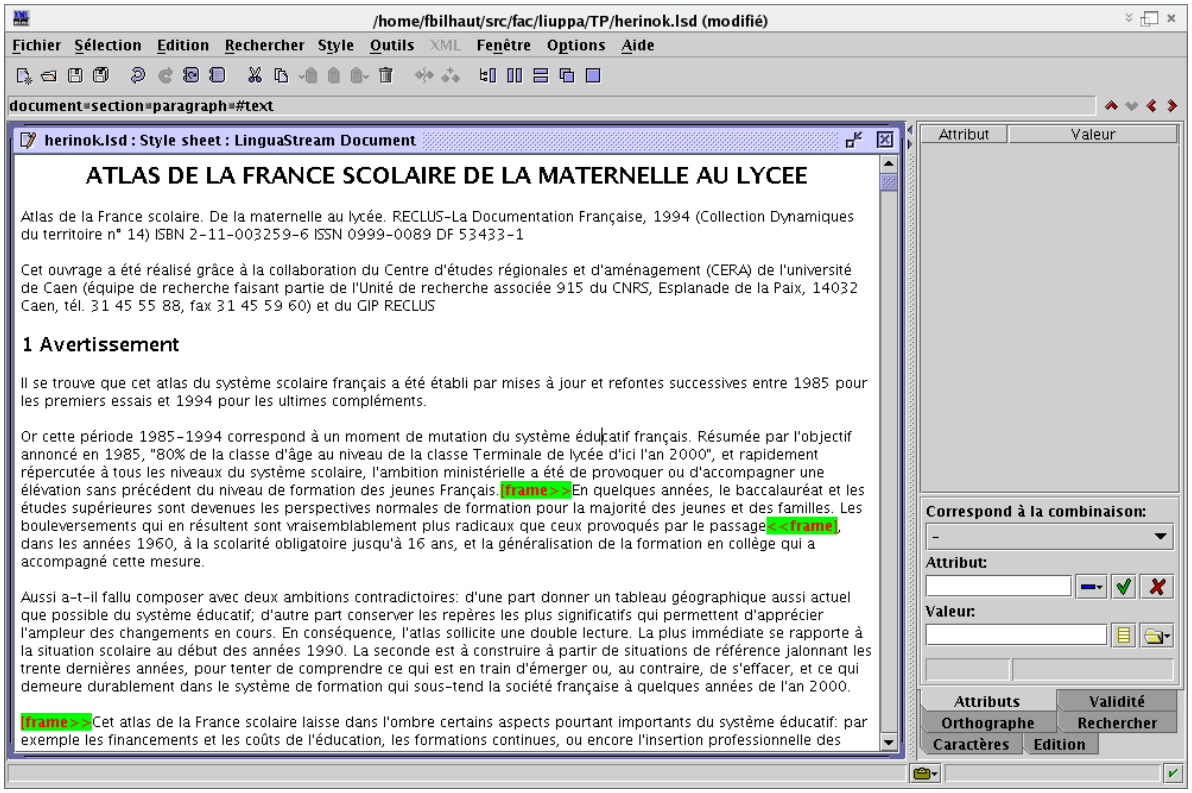


FIG. 7.5 – Annotation manuelle avec l'éditeur XXE.

Faute de mieux, nous utilisons donc des outils d'édition de texte « brut » pour procéder à l'annotation du reste du corpus d'évaluation, ce qui bien sûr ne facilite en rien cette tâche déjà laborieuse. Car même si nous pouvons ici nous aider des macro-commandes Emacs développées pour l'occasion par Antoine Widlöcher, cette approche nécessite une certaine « expertise » XML ainsi que le parcours en parallèle du code « source » et de son « rendu » dans un navigateur afin de ne pas perdre de vue la présentation originale du texte. Notons également que dans tous les cas, il faut également être capable de recouper les marquages produits par les différents annotateurs au sein d'un même document, ce qui est ici réalisé par un composant LinguaStream développé par Stéphane Ferrari.

Les balises relatives à l'annotation manuelle étant des annotations LinguaStream, tous les outils proposés par la plate-forme peuvent s'y appliquer. D'autre part, ces annotations peuvent cohabiter, au sein d'un même document, avec d'autres annotations produites manuellement ou automatiquement. Au delà des outils spécifiquement dédiés à la comparaison de ces différentes annotations dont nous reparlerons ci-après, on notera que les différents outils de visualisation offerts par la plate-forme offrent d'ores et déjà des moyens intéressants d'évaluer qualitativement l'annotation automatique. Il est ainsi possible, comme le montre la figure 7.6, de visualiser les différentes annotations (étiquette « frame » pour l'annotation automatique, et « cadreman » pour l'annotation manuelle) au sein même du document afin de procéder à des comparaisons au cas par cas pour prendre la mesure des défauts de l'annotation automatique.

Un autre mode de visualisation, plus synthétique, permet d'extraire les passages identifiés comme cadres par l'une ou l'autre des annotations, tout en n'affichant que certains éléments du texte contenu dans ces passages. Par exemple, la figure 7.7 montre un exemple de cette vue dite « macro-concordancier », où seuls les cadres annotés automatiquement ou manuellement sont affichés (le même texte étant répété s'il est annoté plusieurs fois). On remarquera que le texte contenu dans ces cadres n'est pas affiché

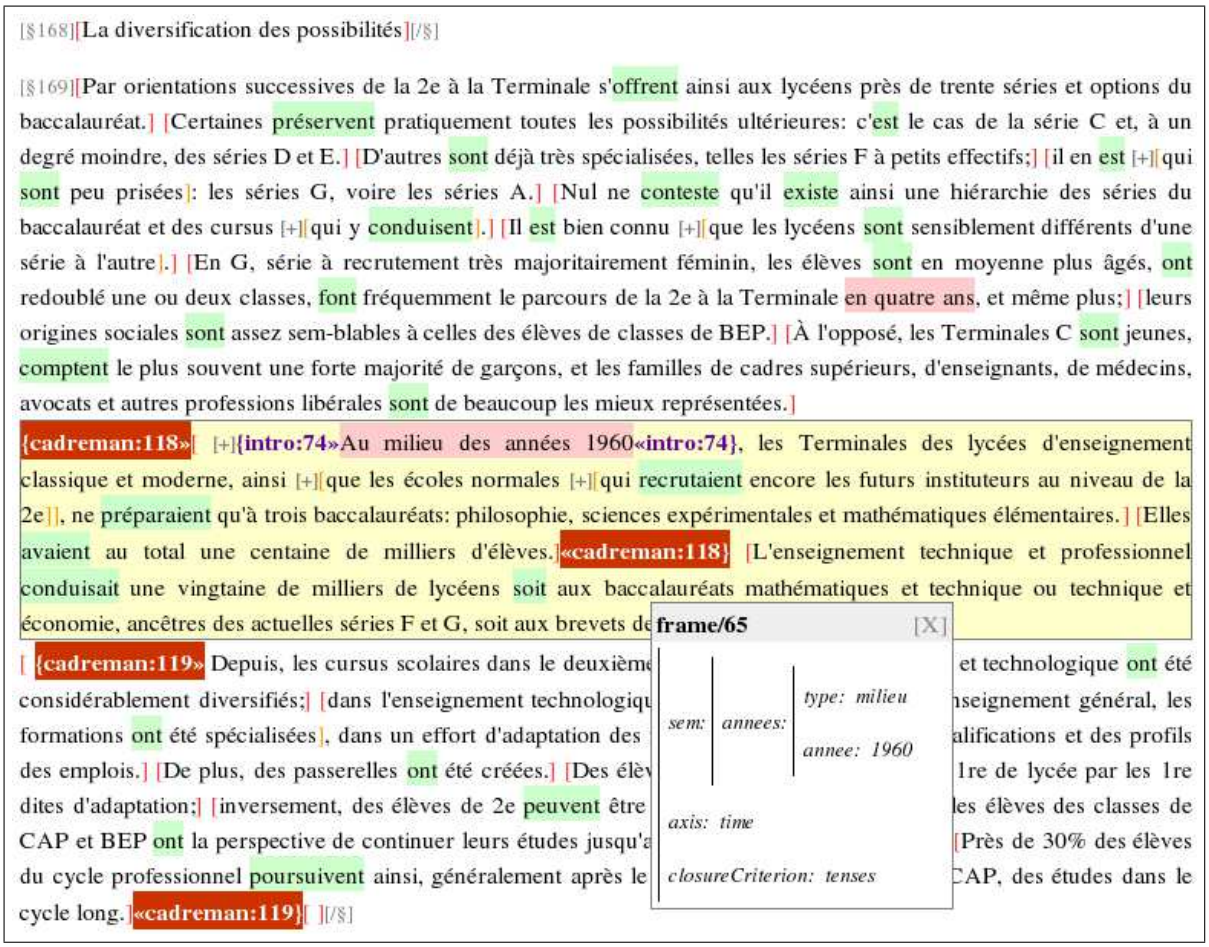


FIG. 7.6 – Comparaison visuelle des analyses manuelle et automatique des cadres temporels (en sortie de LinguaStream).

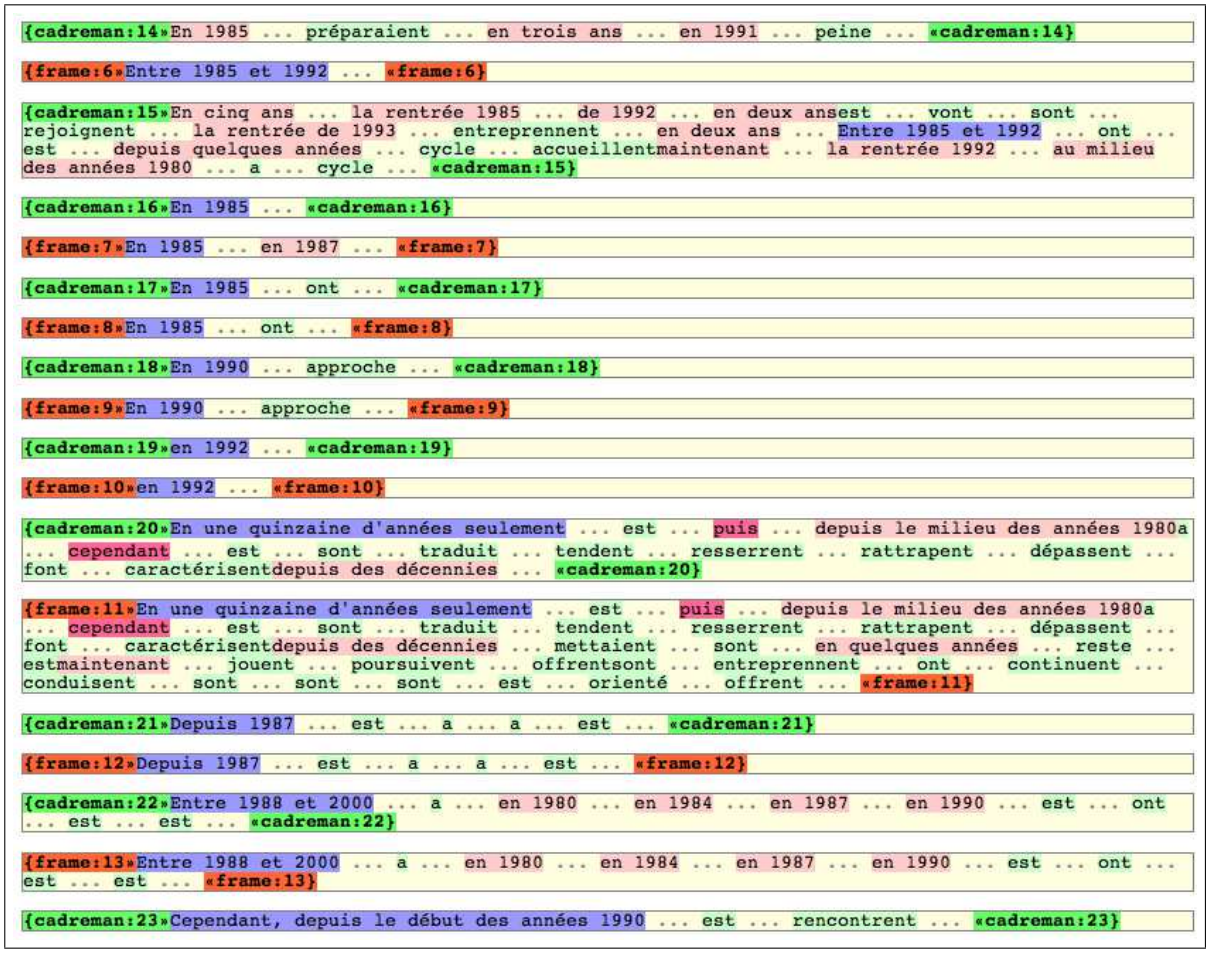


FIG. 7.7 – Comparaison visuelle des analyses manuelle et automatique des cadres temporels en mode « macro-concordancier » (en sortie de LinguaStream).

intégralement : en l'occurrence, seuls les introducteurs, les expressions temporelles et les verbes sont reproduits, dans des couleurs différentes (la présence de portions jugées non significatives est seulement indiquée par des points de suspension). Dans la mesure où des critères portant sur ces éléments jouent un rôle crucial dans la procédure d'analyse automatique que nous avons décrite plus haut dans ce chapitre, cela permet de juger rapidement, pour chaque occurrence de cadre, des éventuels défauts de l'analyse.

Enfin, comme nous le verrons plus loin, ces procédés de visualisation ne sont pas seulement applicables à la comparaison entre annotations automatique et manuelle, mais permettent également de visualiser en parallèle plusieurs annotations manuelles présentes dans un même document.

7.3.2 Méthode

Le procédé d'évaluation proprement dit se décompose en deux étapes : alignement et comparaison. Il a été réalisé sous la forme d'une chaîne LinguaStream, où chacune de ces deux étapes correspond, comme le montre la figure 7.8, à un composant spécifique. Comme dans toute chaîne LinguaStream, cela assure la totale indépendance des deux phases du processus, ce qui permettra éventuellement de modifier l'une d'entre elles sans remettre en cause le fonctionnement de l'autre.

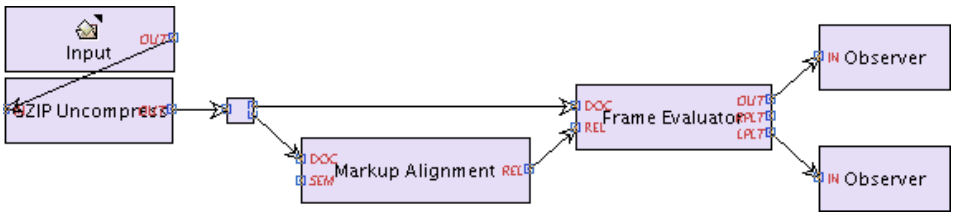


FIG. 7.8 – Chaîne LinguaStream permettant d'aligner les annotations manuelle et automatique et de procéder aux calculs de comparaison.

Alignement

L'étape d'alignement consiste à déterminer, étant donnés deux ensemble d'annotations sur un même texte, celles qui sont censées représenter un même objet linguistique, en l'occurrence un même cadre. Pris en toute généralité, il s'agit d'un problème notoirement complexe, y compris au niveau lexical ou syntaxique. Au niveau discursif, la tâche est d'autant plus ardue que les objets linguistiques en jeu sont parfois fluctuants et ambigus, et que l'expérience de la communauté reste réduite dans ce domaine (la plupart des campagnes d'évaluations portent sur des grains plus fins). De ce fait, il n'est pas toujours possible de l'automatiser, et une intervention humaine peut être nécessaire afin de garantir la qualité de l'alignement (faute de quoi l'évaluation risquerait d'être faussée).

Toutefois, dans le cas qui nous occupe ici, les structures à aligner possèdent des propriétés spécifiques, qui nous permettent d'envisager l'automatisation de l'alignement sans risquer un trop grand nombre d'erreurs. Nous pouvons en effet exploiter les introducteurs, qui sont systématiquement présents à gauche des cadres, en considérant que deux annotations qui partagent un même introducteur correspondent nécessairement à un même cadre. Plus précisément, la procédure que nous avons appliquée est la suivante : étant données deux annotations, on sélectionne l'introducteur de l'une d'entre elles ; si la même chaîne de caractères est présente au début de l'autre annotation avec une tolérance paramétrable relativement à la borne gauche[6], alors les deux annotations sont alignées.

Ce procédé présuppose bien sûr que les introducteurs soient annotés dans au moins l'un des deux documents à aligner. Deux options peuvent alors être considérées. On peut d'une part demander aux experts d'annoter explicitement les introducteurs, mais cela induit une surcharge conséquente (c'était toutefois le cas concernant l'annotation de (HER)). Dans le cas contraire, il sera possible dans certains cas d'exploiter l'annotation automatique, pour peu que celle-ci offre une précision satisfaisante (le rappel n'intervient pas ici). C'est par exemple possible dans le cas des introducteurs temporels, car l'analyseur que nous avons décrit précédemment offre justement une excellente précision au détriment du rappel, comme nous pourrons le constater plus loin en observant les résultats de l'évaluation sur (HER).

Ce procédé est réalisé par le composant « Markup Alignment » de la chaîne de la figure 7.6. Le résultat du traitement se présente dans LinguaStream sous la forme d'un ensemble de relations entre annotations, ce qui autorise leur manipulation avec n'importe quel outil de la plate-forme, et notamment leur visualisation au sein du document annoté (cf. partie III, figure 13.8).

Comparaison

La procédure de comparaison est bien sûr intimement liée aux modalités d'annotation décrites plus haut, mais soulève néanmoins de nombreuses questions. Il s'agit notamment de déterminer :

[6] Dans le cadre des expériences relatées ici, nous avons appliqué une tolérance de dix mots.

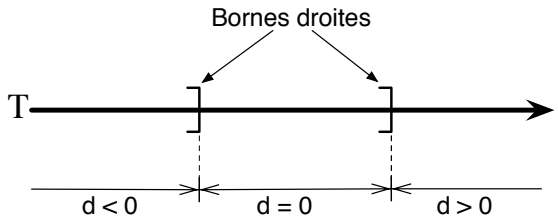


FIG. 7.9 – Calcul de la pseudo-distance entre la borne droite évaluée et les bornes droites de référence.

– quelles distances il est pertinent de mesurer ;
– la méthode à employer pour évaluer ces distances ;
– l'unité à utiliser pour les exprimer ;
– le caractère relatif de la mesure de distance, par rapport à la taille des objets étudiés.

Nous avons bien sûr été amenés à faire des choix concernant ces différents points pour pouvoir mener une première expérience d'évaluation. Toutefois, ces choix ne sont nullement irrévocables, et il s'agirait au contraire de les discuter à la lumière des enseignements qui seront apportés par de futures manipulations. Dans cette optique, les composants que nous avons développés pour l'évaluation sont largement paramétrables, et bénéficient de la flexibilité apportée par la notion de perspective d'analyse de LinguaStream (cf. section 10.3). De ce fait, il sera facile d'adapter leur comportement le cas échéant.

Le choix des distances mesurées pose peu de problème dans notre cas, et il s'agira notamment de mesurer les écarts entre les bornes finales de chaque couple aligné (évaluation de la portée). Le composant de comparaison permet également de mesurer les écarts entre les bornes initiales, bien que ce point soit, comme nous l'avons remarqué à plusieurs reprises, beaucoup moins épineux. Concernant l'évaluation de la portée, différents modes de calcul de distance sont envisageables, notamment en raison de la possible présence d'une double borne droite dans l'annotation de référence (cf. infra). Dans le cas où une seule borne droite est spécifiée, on évaluera simplement la distance en dénombrant une unité linguistique que nous définirons par ailleurs, afin d'obtenir une valeur qui sera (par convention) négative dans le cas où la borne évaluée précède la borne de référence, positive si elle lui succède, ou nulle si les deux bornes coïncident exactement. La même méthode de calcul sera bien sûr appliquée à la borne gauche, qui est toujours unique.

Dans le cas où deux bornes droites ont effectivement été spécifiées par l'expert, différentes options sont envisageables, selon le sens que l'on attribue à la présence de bornes multiples. Si l'on considère qu'elles désignent différents points de terminaison possibles, alors on cherchera par exemple à mesurer la distance de la borne évaluée relativement à la plus proche des bornes de référence. En revanche, si l'on considère que les deux bornes finales de référence délimitent plutôt une zone de « fermeture acceptable », il conviendra plutôt de considérer une pseudo-distance qui sera nulle dès lors que la borne évaluée se trouve entre les bornes de référence. Comme nous l'avons dit plus haut, chacun de ces deux types de bornes multiples peut avoir un intérêt, et il serait probablement intéressant de pouvoir les gérer parallèlement. Mais c'est la seconde solution, à l'exclusion de la première, que nous avons adoptée dans un premier temps dans un souci de simplicité. Ainsi, comme le montre la figure 7.9, la distance mesurée sera nulle si la bornée évaluée se situe entre les bornes de référence, négative si la borne évaluée précède la première borne de référence, et positive si elle succède à la seconde borne de référence.

Concernant le choix de l'unité de mesure, diverses possibilités sont là encore envisageables. Mise à part l'unité « caractère » qui ne trouverait ici que peu de justification linguistique, on pourra notamment considérer les unités « mot », « syntagme », « proposition » ou même « phrase ». Nous avons pour l'heure choisi la première option, *a priori* moins problématique que les autres dans la mesure

le choix de ces derniers grains demanderait également d'appréhender assez précisément leurs possibilités d'interaction avec le grain « cadre », et de résoudre le cas échéant d'éventuels problèmes de chevauchement entre ces unités. Toutefois, le choix d'une unité autre que le mot ne poserait pas de problème technique particulier, et il sera facile à l'avenir d'évaluer cette possibilité en exploitant les fonctionnalités de variabilité du grain de LinguaStream.

Un autre point de discussion concerne l'opportunité de rapporter les distances mesurées à différents ordres de grandeur observés sur les objets manipulés. Ici, il s'agirait tout d'abord de se donner les moyens de ramener les écarts mesurés à droite ou à gauche à la taille des cadres de l'annotation de référence. On peut en effet formuler l'hypothèse qu'un écart mesuré en distance « absolue » sera d'autant plus significatif que le cadre auquel il se rapporte est de faible longueur. Cette hypothèse reste toutefois à valider linguistiquement, et il s'agit également prendre garde à conserver des principes de mesures en rapport avec les applications visées[7].

Une autre possibilité (non exclusive de la précédente) consisterait à rapporter les écarts mesurés au « flou » consenti par l'expert quant à la borne finale. Dans ce cas, on pourrait soit considérer qu'un écart d'annotation est d'autant moins pénalisant que le flou est important, en admettant que cela indique un cas « difficile » car ambigu, soit au contraire qu'un écart d'annotation est d'autant plus pénalisant que le flou est important, partant du principe que dans ce cas la probabilité pour un analyseur automatique de positionner la borne finale dans la plage délimitée est plus grande.

Au cours de l'expérience relatée ici, nous avons simplement choisi d'exprimer les écarts mesurés en distance absolue, tout en calculant la taille moyenne des cadres des deux annotations comparées. Nous avons ainsi les moyens de rapporter les premières à la seconde, rapport dont l'intérêt nous a semblé le plus évident. En revanche, la possibilité d'exprimer chaque écart de façon relative à d'autres grandeurs (taille du cadre ou de la zone de flou) propre au cadre considéré reste à explorer.

7.4 Premiers résultats

Comme nous l'avons dit plus haut, les résultats d'évaluation obtenus à l'heure actuelle ne sont pas issus d'une masse de données suffisamment large pour être réellement significatifs. Nous souhaitons toutefois en faire état dans la mesure où ils peuvent d'ores et déjà apporter des enseignements intéressants. Ceux-ci portent sur trois points : les mesures de type précision/rappel, les mesures concernant la portée des introducteurs, et les différences entres les annotations de plusieurs experts sur un même texte.

7.4.1 Mesures de précision et de rappel

Les premiers résultats quantitatifs de l'évaluation portent sur le corpus (HER). 156 cadres qui y ont été annotés manuellement et 80 y ont été reconnus automatiquement. La procédure d'alignement a permis d'obtenir 72 couples de cadres. On voit donc immédiatement que le *rappel* est relativement faible (0,53). Cette mesure doit toutefois être interprétée avec précaution, car la détection d'un cadre est directement tributaire de la détection de l'expression temporelle qui l'introduit. Or l'analyseur d'expression temporelles ne vise aucunement l'exhaustivité, puisqu'elle est intentionnellement limitée à des expressions dont la valeur temporelle est directement calculable sous la forme d'une période explicite, et ce sans faire appel à des connaissances du domaine (expressions du type « à la rentrée » ou « après la première guerre ») ni procéder à des inférences complexes comme la résolution d'anaphores temporelles (expressions du type « depuis lors »). De ce fait, le spectre des expressions déclenchant la

[7] On peut par exemple se demander si le défaut d'une annotation tel que perçu par l'utilisateur du système qui l'exploite dépend réellement de la taille du cadre considéré. Tout cela dépend bien évidemment de l'application visée.

reconnaissance d'un cadre est significativement réduit par rapport aux à celui des expressions prises en compte par l'expert. Le système d'analyse des cadres en tant que tel n'est donc pas directement en cause sur ce point.

On remarquera d'autre part que le taux de précision est à l'inverse remarquablement bon (0,90). Il convient de rappeler à ce sujet que la procédure d'alignement ne tient compte que des introducteurs, ce qui signifie que cette mesure de précision n'est aucunement relative à l'analyse de la portée, mais reflète seulement le fait que la reconnaissance des introducteurs est très fiable. Ce résultat est peu étonnant dans la mesure où leur détection est liée à des critères positionnels très contraints, et que la précision de l'analyseur d'expressions temporelles est elle-même très bonne. Cette précision sera par d'ailleurs confirmée au niveau de la mesure de distance moyenne entre les bornes gauches que nous évoquerons plus loin.

On remarquera d'autre part qu'il existe 8 cas de cadres détectés par l'analyseur qui ne l'ont pas été par l'expert. Ceux-ci s'avèrent correspondre soit à de simples oublis soit à des introducteurs auxquels l'expert n'a pas attribué de portée extra-phrastique.

7.4.2 Évaluation du calcul de la portée

Concernant le calcul de la portée, on notera tout d'abord une longueur moyenne des cadres annotés manuellement de 120 mots, quand celle des cadres annotés automatiquement est de 175 mots. Cela nous indique immédiatement une forte propension de l'analyseur à délimiter des cadres trop grands, tendance confirmée comme nous allons le voir par la mesure des écarts à gauche et à droite. La méthode que nous avons précédemment décrite fournit en effet les résultats suivants, en nombre de mots (rappelons que pour chaque borne la mesure d'écart peut être soit nulle, soit positive si la borne mesurée est à gauche de la borne de référence, soit négative si elle est à sa droite) :

Écart	moyen en valeur absolue	moyen > 0	moyen < 0	minimum	maximum
À gauche	0,00	0,00	0,00	0,00	0,00
À droite	47,59	100,79	0,00	0,00	518

On remarque tout d'abord une erreur nulle au niveau de la borne initiale, ce qui n'est pas réellement surprenant puisque plusieurs facteurs y concourent :

- comme nous l'avons évoqué plus haut, l'analyseur d'introducteurs bénéficie d'une très bonne précision ;
- les bornes sont systématiquement posées en début de phrase (tant par l'expert que par l'analyseur automatique), et la détection automatique des limites des phrases n'est pas problématique dans le corpus ici étudié ;
- le procédé d'alignement fait intervenir une distance maximale entre les introducteurs des cadres alignés, ce qui écarte d'office les cas où il y aurait effectivement une erreur très grande au niveau de la borne droite. Ce cas reste toutefois très peu probable, et une vérification manuelle a d'ailleurs permis de vérifier qu'il ne s'était pas présenté dans les données ici traitées.

Le risque d'erreur à ce niveau est donc extrêmement réduit, même si on ne peut bien sûr pas s'attendre à une erreur nulle en toute circonstance. Les résultats obtenus au niveau de la borne droite sont en revanche plus intéressants. L'erreur moyenne en valeur absolue est un indicateur peu révélateur en règle générale, puisque sa grandeur n'est pas directement révélatrice de la distance moyenne effective (il même est nul si les écarts positifs et négatifs se compensent exactement). Il donne en revanche une indication sur une la tendance générale entre écarts positifs et négatifs, avec ici une nette prédominance positive qui confirme l'indication donnée par les tailles moyennes des cadres. De fait, on voit que les écarts mesurés dans le corpus ici étudié sont systématiquement positifs (la moyenne des écarts strictement négatifs est nulle), ce qui signifie que la borne droite mesurée se situe toujours à droite de la deuxième borne droite de référence. En d'autres termes, les cadres repérés automatiquement sont

systématiquement *trop longs* relativement aux cadres de référence. D'autre part, on remarque un écart positif moyen non négligeable (de l'ordre de 100 mots), tout comme l'écart maximum qui est de 518 mots.

Commentons tout d'abord ces valeurs en tant que telles avant d'en venir à la tendance générale. Au premier abord, l'écart moyen comparé à la taille moyenne des cadres de référence est assez important, les deux grandeurs étant du même ordre. L'ensemble des chiffres montre de fait une tendance de l'analyseur à détecter, en moyenne, des cadres « deux fois trop longs ». D'autre part, si l'on considère la taille moyenne des phrases dans ce même document (environ 29 mots), on obtient un écart à droite de l'ordre de 3 phrases. Mais il est nécessaire de relativiser ces chiffres. D'une part, il s'agit de moyennes, alors que l'écart-type associé dépasse les 95 mots. Il convient donc d'étudier plus précisément leur distribution, qui montre que les résultats obtenus restent encourageants. On voit en effet sur la figure 7.11 qu'une quantité importante de cadres sont reconnus de façon satisfaisante, puisque plus de 65% le sont avec une erreur inférieure à 10 mots.

D'autre part, les écarts observés sont à rapprocher de la « flexibilité » de la notion de cadre. On observe en effet, quand plusieurs experts sont amenés à analyser un même texte, des écarts parfois très importants entre les annotations produites. Nous avons pu vérifier ce fait concrètement en procédant à la triple annotation d'un numéro du journal *Le Monde*. Comme le montre la figure 7.10, ces différentes annotations ont été regroupées sous forme de marquage LinguaStream au sein d'un même document[8] et peuvent être facilement visualisées. Nous avons ainsi pu identifier un certain nombre de cas où, même après concertation entre les experts, plusieurs annotations différentes pouvaient être également acceptables, et qu'il faut donc accepter une marge d'erreur inhérente aux objets analyses eux-mêmes.

Mais dans un premier temps, l'intérêt des mesures obtenues réside essentiellement dans les indications qu'elles fournissent quant aux directions à suivre pour améliorer l'analyseur. En effet, même si ces résultats ont été obtenus sur un corpus encore trop restreint pour fournir des chiffres réellement significatifs, la tendance de l'analyseur à délimiter des cadres trop longs est si nettement marquée que nous devons en tenir compte dans notre réflexion sur les critères d'évaluation de la portée d'un introducteur. Et il est particulièrement intéressant de voir que cette propension de l'analyseur va à l'encontre de certaines intuitions que nous avions à ce sujet.

On pouvait en effet penser que les critères de fermeture exploités conduiraient plutôt à une clôture prématurée de nombreux cadres, notamment parce qu'ils sont appliqués séparément et que l'on sait qu'ils ne sont pas parfaitement fiables utilisés comme tels (cf. sections précédentes). Il est clair que les résultats de l'évaluation ne réfutent pas directement cette faiblesse des critères, qui est bien réelle et facilement démontrable par l'exemple. Toutefois, ils nous permettront d'orienter nos choix à court terme concernant l'amélioration des règles d'analyse de la portée : tant que la tendance à fermer les cadres « trop tard » persistera, il sera en effet inutile de chercher à relâcher les contraintes existantes. Au contraire, il nous faudra sous-doute chercher à identifier de nouveaux indices de rupture, ce à quoi le corpus doublement annoté utilisé pour l'évaluation pourra être particulièrement utile.

[8] À l'aide d'un outil de fusion développé par Stéphane Ferrari.

LISBONNE

de notre envoyé spécial
ETR

BRESSON HENRI DE

PORTUGAL : la révision de la Constitution
En avant toute pour l'Europe de 1992

La première vague de chaleur est venue tardivement cette année remplir du flot des touristes étrangers les terrasses d'Alfama, dominant le Tage. Malgré le bras de fer qui oppose le premier ministre aux socialistes sur la réforme de la Constitution, dernier avatar de la révolution des oeillets, cette irruption de l'été tombait à pic pour permettre à M. Cavaco Silva de fêter sereinement le premier anniversaire de la majorité absolue conquise par sa formation, le Parti social-démocrate, aux législatives du 19 juillet 1987.

{cadreSF:62» **{cadreFB:43»** **{cadreAW:72»**Après les années agitées de la révolution et d'une décolonisation qui a vu revenir au pays un million d'expatriés, la jeune démocratie portugaise s'est enfin stabilisée.**«cadreSF:62}** Cessant de s'abandonner aux délices des débats idéologiques, le Portugal de 1988 entend se consacrer tout entier à son développement **«cadreFB:43}**. **{cadreSF:63»** **{cadreFB:44»** **{cadreAW:73»**Après avoir mené avec succès les négociations sur l'entrée de son pays dans le Marché commun lors de son premier gouvernement minoritaire, M. Cavaco Silva regarde l'avenir avec confiance. Il dispose pour les prochaines années d'une majorité stable et peut compter sur d'importantes aides de la CEE pour mener à bien la modernisation du pays.**«cadreSF:63}** **«cadreSF:62}** **«cadreFB:44}** **«cadreAW:73}** **«cadreAW:72}**

L'ascension météorique de cet ancien professeur d'économie, surgi de l'anonymat de l'université pour prendre la tête, en 1985, du Parti social-démocrate et quelques mois plus tard celle du gouvernement, symbolise assez bien le nouveau paysage politique portugais. On ne peut qu'être frappé à Lisbonne par la jeunesse de tous les dirigeants actuels, aussi bien dans les partis que dans les milieux d'affaires. " La révolution a brûlé deux générations ", constate M. José Amaral, trente-deux ans, conseiller du président de la République, M. Mario Soares, pour les questions européennes, et membre du directoire d'une des nouvelles banques privées portugaises. " Celle de l'ancien régime a quitté le pays ou s'est retirée des affaires. Quant à la génération de 1974, elle a perdu sa place avec la stabilisation, quand on est passé aux choses sérieuses, à la gestion. "

Cette constatation vaut pour tous les bords. A droite, la prise de pouvoir de M. Silva au sein du PSD, principale formation conservatrice du pays, s'est accompagnée de la mise à l'écart des anciens caciques du parti. Parallèlement, une nouvelle classe de

FIG. 7.10 – Visualisation simultanée des annotations produites sur un même texte par des observateurs différents.

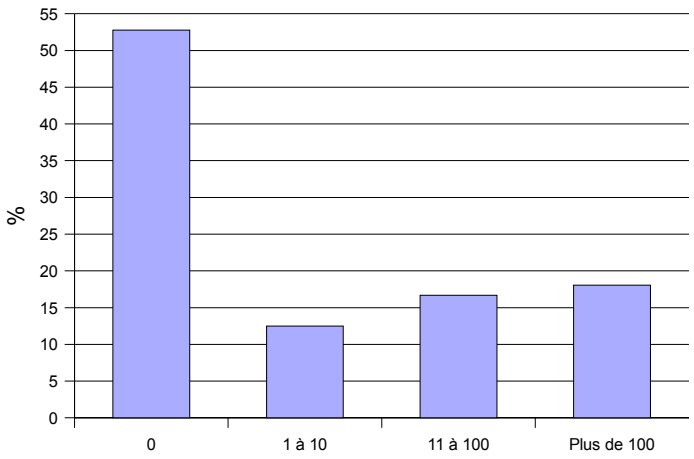


FIG. 7.11 – Évaluation du calcul de la portée : distribution des écarts relativement à l'annotation manuelle, en nombre de mots.

Chapitre 8

Thèmes discursifs composites

8.1 Introduction

Nous avons étudié dans les chapitres précédents le cas de la recherche d'information dans les documents géographiques, dans le cadre défini par le projet GeoSem. Nous avons alors présenté un ensemble de systèmes d'analyse de documents pouvant être mis en oeuvre pour produire une indexation sémantique fine des documents, directement applicable à la recherche d'informations sous la forme d'un moteur de recherche « multi-dimensionnel ».

Dans ce cas précis, la collaboration avec des géographes utilisateurs potentiels a permis de dégager un certain nombre de propriétés de l'information géographique que nous nous devions de prendre en compte dans l'analyse et l'indexation des documents du domaine. Cela concerne par exemple la méthodologie mise en oeuvre dans la rédaction même des documents, les problèmes de modélisation de l'espace et du temps, les formes graphiques et discursives à prendre en compte, et bien sûr la structure des requêtes formulées, avec ses trois composantes « phénomène », « espace », et « temps ». Dans le cadre précis de l'analyse automatique du discours, nous nous sommes ainsi intéressés de près aux liens en discours entre les composantes spatiales et temporelles avec les phénomènes décrits dans les textes, ainsi qu'à une catégorie particulière de structure discursive, les cadres de discours.

Plus généralement, l'exemple de l'information géographique montre très concrètement l'importance que peut revêtir la prise en compte des spécificités de l'information que l'on manipule dans le processus de recherche d'informations. Si cette approche n'est bien sûr pas toujours applicable (puisque certaines bases documentaires sont par essence « généralistes »), les gains potentiels pour les utilisateurs spécialistes sont loin d'être négligeables, et il nous semble donc intéressant de persévérer dans cette direction.

Le travail présenté dans cette partie a ainsi pour objectif de généraliser certains principes que nous avons développés dans le cadre de l'information géographique, selon les directions suivantes. Il apparaît d'une part que les dimensions de l'espace et du temps, même si elles sont quasi-universelles et dépassent largement le cas particulier de l'information géographique, peuvent être considérées d'un certain point de vue comme un cas particulier de ce que nous appellerons « axes sémantiques ». Nous nous appuierons sur cette notion pour généraliser la structure phénomène / espace / temps de l'information géographique et l'appliquer à d'autres domaines. Il apparaît d'autre part que si les cadres de discours au sens de Charolles (cf. chapitre 7) revêtent un rôle important dans les manifestations discursives du croisement entre des axes sémantiques, d'autres structures peuvent être envisagées. Plus généralement, nous envisagerons la notion de « thème composite », qui nous semble à même de caractériser certains phénomènes thématiques liés à différentes configuration discursives.

8.2 La notion de thème composite

Comme nous l'avons annoncé au début de ce chapitre, le travail présenté ici a pour vocation de généraliser les résultats obtenus dans le cadre du projet GeoSem. Cette étude a d'une part fait apparaître l'importance que peuvent revêtir certains groupes adverbiaux extra-prédicatifs, en l'occurrence spatiaux et temporels, pour une tâche d'indexation automatique. Pour le spécialiste, ces composantes font partie intégrante du contenu informationnel des textes géographiques, et doivent donc être prises en compte de façon explicite pour en produire une représentation pertinente. Nous avons donc été amenés à caractériser ce contenu par des triplets tels que (« le vieillissement de la population » ; « dans les années 80 » ; « en Europe »). Les n-uplets de ce type seront ici appelés *thèmes composites*.

Nous verrons à travers de nombreux exemples que ce fonctionnement n'est pas réservé aux adverbiaux spatio-temporels, mais peut au contraire être élargi à bien d'autres types d'expressions. Par exemple, tel passage d'un texte touchant à l'univers scolaire pourrait être représenté par (« le retard scolaire » ; « dans le secondaire » ; « depuis 1985 »). Il apparaît en effet que des ensembles de concepts tels que { PRÉ-ÉLÉMENTAIRE, ÉLÉMENTAIRE, SECONDAIRE, SUPÉRIEUR } peuvent fonctionner en discours de façon analogue aux axes du temps ou à l'espace. Ces ensembles, composés de notions à la fois interdépendantes et disjointes et apparaissant dans de nombreux domaines de spécialité, seront ici appelés *axes sémantiques*, notion sur laquelle nous nous attarderons dans la section 8.4. Autour de la notion de thème composite, nous allons pour l'heure nous intéresser à leur rôle du point de vue de l'organisation du discours et de la représentation de l'à propos de certains types de segments textuels.

Nous avons pu constater à plusieurs reprises dans le chapitre 6 que la mise en oeuvre du principe d'indexation par thèmes composites fait apparaître la nécessité d'adopter une approche discursive (point également discuté en annexe A). Nous avons par exemple vu que lorsqu'un fragment textuel décrit un phénomène socio-géographique relativement à une localisation dans l'espace et le temps, les références aux trois composantes impliquées n'apparaîtront pas nécessairement dans la même proposition, ni même dans la même phrase. Au contraire, elles seront généralement disséminées au sein de structures discursives complexes. Par exemple, il est nécessaire de procéder à une analyse globale de l'extrait ci-dessous pour pouvoir attribuer au segment noté S le thème composite (« la secondarisation des effectifs scolaires » ; « entre les années 1960 et la fin des années 1980 » ; « dans la France du Nord ») :

> Entre les années 1960 et la fin des années 1980, le nombre de collégiens et lycéens pour 100 élèves du primaire est passé de 45 à plus de 80, par le double effet de la croissance générale et souvent accélérée des effectifs des collèges et lycées et de la diminution ou, au mieux, de la croissance ralentie de ceux du primaire. { Mais cette secondarisation a été fort inégale. Elle est forte dans la France du Nord [...] : les effectifs du secondaire y ont fréquemment augmenté de plus des trois quarts en vingt ans, et le rapport secondaire/primaire y a souvent plus que doublé. }$_S$ Dans la France du Centre et du Sud-Ouest, [...]

Des éléments de réponse à ce problème a été donnée dans le chapitre 7 en se basant sur le modèle de l'encadrement du discours de Michel Charolles. Dans le cas de l'exemple précédent, l'analyse des cadres temporels permet effectivement de déterminer que l'expression temporelle donnée en initiale introduit un cadre englobant le segment S, et doit donc être prise en compte dans la description de son contenu informationnel. L'approche que nous adopterons dorénavant consistera à considérer le modèle de l'encadrement du discours comme un cas particulier d'un schéma discursif à la fois plus général, puisque l'encadrement n'en est qu'une instance possible, et plus restrictif, puisqu'il ne concerne que les structures liées aux thèmes composites. Nous chercherons donc à énumérer une certaine variété de configurations discursives permettant d'instaurer, au fil du texte, ce que nous appelons des thèmes composites.

8.2.1 Définition

La notion de thème composite se fonde sur l'idée déjà évoquée à plusieurs reprises que dans certaines circonstances, la description de l'à propos d'un texte ou d'un fragment textuel nécessite de prendre en compte non seulement ne concept ou le référent au sujet duquel on apporte une information, mais aussi le contexte au sein duquel doit être située cette information. En d'autres termes, selon les termes de Martin tels que repris par Charolles (cf. section 3.4.1), nous souhaitons adjoindre au thème ou topique (au sens classique) d'un segment l'univers de discours qui s'y rattache. On peut également y voir une application au niveau discursif du principe d'indexation par facettes (cf. section 2.3.3).

L'intérêt de cette approche est double. Dans la perspective particulière de la recherche d'information, elle nous permet de décrire de façon précise le contenu informationnel d'un texte ou d'un segment textuel, dont l'intérêt est évident quant à la précision des requêtes formulables et des résultats obtenus par l'utilisateur. Et plus généralement, dans la perspective de la sémantique du discours, il nous semble que la notion de thème composite associée à celle d'axe sémantique permet de rendre compte assez efficacement de la structure et du sens d'une certaine variété de configurations discursives où les approches classiques du thème seraient inadaptées.

Un thème composite sera constitué de deux éléments. Le premier, que nous appellerons *noyau thématique* ou simplement *noyau*, correspond à la définition traditionnelle du thème en tant que représentant de l'à propos d'un objet textuel. Le second est lui-même un ensemble d'éléments que nous appellerons *satellites*, qui définissent l'univers de discours au sein duquel se tient le noyau. Un thème composé d'un *noyau* n et d'un ensemble de *satellites* s_i, l'ensemble sera noté comme suit :

$$n \bullet\!\!-\!\!\circ (s_i, ..., s_n)$$

Précisons qu'à la manière de Lambrecht, nous considérerons que ces composantes d'un thème sont bien des objets purement conceptuels, distincts des unités linguistiques dont ils sont les référents. En cas d'ambiguïté, et en nous inspirant de la terminologie de ce même auteur, nous ferons usage des termes « expressions noyau » ou « expression satellite » pour désigner les unités linguistiques correspondantes. D'autre part, les entités conceptuelles seront par la suite notées en petites majuscules, sauf cas particuliers comme les périodes temporelles qui seront représentées sous forme d'intervalles. Considérons par exemple l'énoncé suivant :

> Dans l'Ouest, le taux de retard scolaire est en régression depuis une dizaine d'années.

Une représentation possible de son thème composite pourrait être :

LE TAUX DE RETARD SCOLAIRE $\bullet\!\!-\!\!\circ$ (DANS L'OUEST, DEPUIS UNE DIZAINE D'ANNÉES)

Bien-sûr, notre objectif est de décrire l'à propos de segments discursifs et non seulement de phrases. Nous serons amené à considérer, au cours de ce chapitre, différentes « configurations discursives » susceptibles de construire des thèmes composites en discours. Il existe en effet divers « outils discursifs » susceptibles de participer à l'installation progressive des différentes facettes qui participent de « ce dont on parle » dans un segment textuel. Toutefois, la construction canonique en la matière pourrait être représentée par les cadres de discours, dont voici un nouvel exemple[1] :

[1] L'acronyme VFR signifie Visual Flying Rules (règles de vol à vue), et est toujours employé sous cette forme en français.

> En VFR, l'évitement des autres aéronefs et des obstacles s'effectue généralement en appliquant la règle "voir et éviter". Il s'agit de détecter assez tôt l'éventuel danger extérieur pour concevoir et exécuter une manoeuvre d'évitement. Il peut arriver que le regard des occupants en place avant soit orienté longuement vers les instruments. Ainsi, la surveillance extérieure devient défaillante.
>
> Source : BEA

Le thème composite associé pourrait alors être le suivant :

$$\text{PROCÉDURES D'ÉVITEMENT} \bullet\!\!-\!\!\circ \text{ (EN VFR)}$$

Au niveau discursif, les thèmes composites seront souvent réalisés sous forme de structure hiérarchiques. Dans ce cas, bien que la notation que nous venons d'introduire soit à même de représenter les feuilles de la structure résultante, nous devons introduire une notation applicable aux autres éléments de l'arbre. Envisageons pour cela l'exemple suivant (où nous pouvons par ailleurs observer le rôle du titre dans l'établissement du noyau thématique) :

> θ **L'explosion des effectifs scolaires**
>
> § { {Dans l'enseignement public, elle s'accélère en Île-de-France, en Picardie, dans le Centre, ainsi qu'en Provence ; elle reste modérée dans l'Ouest et le Nord. [...] }$_{S1}$ {L'enseignement privé enregistre des baisses d'effectifs en Bretagne, où il est fortement implanté, ainsi que dans les académies de la diagonale Pyrénées-Lorraine, où son audience est par contre traditionnellement réduite [...] }$_{S2}$ }$_{S0}$

Dans ce cas, puisque les thèmes composites de S_1 et S_2 peuvent être représentés par EFFECTIFS SCOLAIRES $\bullet\!\!-\!\!\circ$ (PUBLIC) et EFFECTIFS SCOLAIRES $\bullet\!\!-\!\!\circ$ (PRIVÉ), nous représenterons le thème du segment englobant S_0 en utilisant la notation suivante :

$$\mathcal{T}(S_0) = \text{EFFECTIFS SCOLAIRES} \bullet\!\!-\!\!\circ (\langle\langle \text{STATUT} \rangle\rangle)$$

Cette notation nous permet de spécifier que le noyau thématique de S_0 est décrit en relation avec plusieurs entités d'un « axe sémantique » (forme de classe sémantique sur laquelle nous reviendrons par la suite) ici nommé « statut ». De la même façon, le thème composite de S_1 pourrait être noté comme suit :

$$\mathcal{T}(S_1) = \text{EFFECTIFS SCOLAIRES} \bullet\!\!-\!\!\circ (\text{PUBLIC}, \langle\langle \text{STATUT} \rangle\rangle)$$

Nous définissons également une notation graphique permettant de représenter l'ensemble de la structure, dont un exemple est donné dans la figure 8.1. Les différents types de flèches sont utilisés pour distinguer les relations sémantiques entre les différents objets considérés.

On remarquera que la définition que nous avons donnée plus haut du noyau d'un thème composite reste relativement floue, dans la mesure où des approches très diverses se cachent, comme nous l'avons vu dans la partie I, derrière ce que nous avons appelé « définition traditionnelle ». La raison en est que même s'il nous sera par la suite nécessaire, pour en produire une implémentation, de statuer sur la nature précise de ce noyau, nous nous garderons pour l'instant de chercher à en donner une définition plus précise. En effet, si la méthode d'analyse automatique dont nous ferons état dans la section 8.3 sera « termino-quantitative » (de façon similaire à celle présentée en annexe A), cela n'est qu'une possibilité parmi d'autres, et le modèle des thèmes composites est quant à lui fait totalement indépendant de tout choix d'analyse ou de représentation des noyaux. Nous verrons en effet que l'analyse du discours en

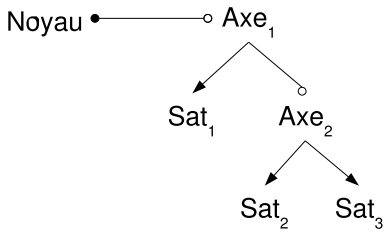


FIG. 8.1 – Représentation graphique arborescente d'un thème composite.

termes de thèmes composites repose avant tout sur l'analyse des satellites, les noyaux étant analysés *a posteriori* au sein des segments ainsi délimités.

Ainsi, notre propos est de proposer une *structure* susceptible de représenter efficacement l'à propos de certains fragments textuels dans certains types de textes, sans bien sûr chercher à résoudre en toute généralité les problèmes posés par l'émergence d'un thème en discours. En d'autres termes, nous cherchons à étudier un phénomène particulier de cohésion et de cohérence fondé sur un phénomène sémantique de « positionnement », phénomène sur lequel nous reviendrons en section 8.4.

Notre démarche pourrait sembler se rapprocher des théories basées sur la notion de macro-structure que nous avons envisagée à plusieurs reprises. De fait, il nous semble indispensable, si l'on souhaite décrire l'à propos de segments discursifs d'une certaine ampleur, d'avoir recours à des structures sémantiques plus ou moins complexes, et non pas seulement de chercher à « sélectionner » dans le texte une ou plusieurs unités, qu'elles soient sémantiques ou de surface, point que nous avons déjà discuté dans la section 3.2. Mais là s'arrête le parallèle, puisque nous suivrons ici les prescriptions de Hutchins (cf. section 2.1), en postulant qu'il est préférable de ne pas chercher à reproduire par ces structures une représentation réduite de l'intégralité de la valeur sémantique d'un texte, mais plutôt de viser l'extraction des éléments qui constituent le « support » de l'information nouvelle, et qui forment ainsi le pivot naturel du processus de recherche d'information. Les structures sémantiques que sont les thèmes composites ne visent donc aucunement à représenter le contenu informationnel d'un fragment textuel dans sa totalité, mais seulement le socle sur lequel s'appuie ce contenu. De ce point de vue, nous rejoignons donc la dualité donné-nouveau (ou thème/rhème) que nous avons détaillée dans la partie I, avec le parti-pris de représenter le thème par des objets sémantiques et structurés.

8.2.2 Thèmes composites et structure informationnelle

À la lecture de la présentation qui précède, on verra immédiatement apparaître un lien entre la notion de thème composite et certaines questions liées à la structure informationnelle dont nous avons discuté dans le chapitre 3. Pour préciser ce point, reprenons l'exemple suivant, que nous avions alors commenté :

> At eight o'clock this morning$_{P1}$ the President$_{P2}$ left from Barajas to attend the international conference to be held in Rome.

Nous avions mentionné à ce propos différents modèles linguistiques conduisant à distinguer dans cette phrase deux constituants potentiellement topicaux, ici marqués P_1 et P_2. En s'autorisant à jouer entre les dénominations proposées par les différents auteurs, le premier pourra être appelé « scene-setting topic », et le second « sujet » ou simplement « topique ». Conformément à ce que nous avions alors annoncé, nous considérerons ici les référents associés comme deux composantes d'un même thème composite : THE PRESIDENT •—○ (AT EIGHT THIS MORNING).

Ce choix est tout d'abord justifié par le fait que la notion de thème relève pour nous directement de l'à propos, et que l'on peut effectivement considérer que cette phrase concerne « ce qu'a fait le Président ce matin à huit heures ». Mais il est d'autre part justifié par le fait que nous prenons en considération le discours dans son ensemble et non pas des phrases isolées. Nous adoptons ainsi une approche pragmatique telle que défendue par Lambrecht, où l'on considère que le rôle topical d'un constituant ne peut être déterminé sans tenir compte du contexte, ou tout du moins que la structure propre à une phrase donnée ne donne que des indications « par défaut ». Il nous paraît opportun de reproduire ici à nouveau ces propos de Charolles que nous avions déjà commentés dans la section 3.4.1 :

> Il n'est pas possible de statuer sur le topique (*aboutness*) d'une phrase isolée. Par défaut de contexte ultérieur, les adverbiaux détachés en tête de phrase, n'indiquent pas ce à propos de quoi est la phrase. Mais rien n'empêche que la suite oblige à leur restituer ce statut. Ce constat, s'il est bien fondé, milite en faveur d'approches dynamiques des phénomènes de topicalisation et conduit à réintégrer dans la discussion de ces phénomènes la notion de topique de discours. (Charolles, 2003, p. 47)

Pour ces raisons, nous décrirons dorénavant toujours les thèmes composites d'un segment *relativement au contexte considéré* : nous considérerons le thème composite d'un segment donné non pas comme une donnée absolue, mais relative à un autre segment englobant que nous dirons « de référence ». Selon le segment de référence choisi, on pourra ainsi obtenir des thèmes différents pour un même segment[2]. Pour les distinguer, nous noterons $\mathcal{T}(S_1, S_0)$ le thème composite d'un segment S_1 considéré relativement à un autre segment S_0. Pour illustrer ce point, considérons les trois exemples suivants :

(1) { { Dans l'enseignement primaire, on assiste à une forte diminution du taux de retard scolaire.}$_{U1}$ Cette baisse est en partie attribuable à [...] }$_{S1}$

(2) { { Dans l'enseignement primaire, on assiste à une forte diminution du taux de retard scolaire.}$_{U2}$ Cette section de notre système scolaire est alors celui qui [...] }$_{S2}$

(3) { { L'enseignement primaire a connu une forte diminution du taux de retard scolaire.}$_{U3}$ Cette baisse est en partie attribuable à [...] }$_{S3}$

Dans le premier cas, les règles « par défaut » nous invitent à considérer que le thème de la première phrase considérée isolément est $\mathcal{T}(U_1, U_1)$ = RETARD SCOLAIRE $\bullet\!\!-\!\!\circ$ (DANS LE PRIMAIRE). La phrase suivante confirme cet fait, et le thème composite de cette phrase considérée relativement à l'ensemble du segment est identique : $\mathcal{T}(U_1, S_1) = \mathcal{T}(U_1, U_1)$. Dans le second cas, la première phrase est identique mais la seconde vient remettre en cause les choix faits *a priori*. Ou plutôt, le thème de U_2 considérée relativement à S_2 est différent du thème de cette même phrase considérée isolément, puisque $\mathcal{T}(U_2, S_2)$ = ENSEIGNEMENT PRIMAIRE $\bullet\!\!-\!\!\circ$ $\emptyset$. Dans le troisième cas enfin, la situation est inverse, puisque $\mathcal{T}(U_3, U_3)$ = ENSEIGNEMENT PRIMAIRE $\bullet\!\!-\!\!\circ$ $\emptyset$ alors que $\mathcal{T}(U_3, S_3)$ = RETARD SCOLAIRE $\bullet\!\!-\!\!\circ$ (DANS LE PRIMAIRE) exactement comme dans le premier cas.

Nous reviendrons plus loin sur les configurations discursives faisant intervenir la dimension relative des thèmes composites. Mais pour l'heure, nous allons considérer un certain nombre de configurations discursives sans-doute moins problématiques.

[2]Notons que cela ne nous conduira pas nécessairement à considérer que l'un ou l'autre est plus « valide » que les autres, puisque dans notre perspective applicative on pourra effectivement être amené à considérer un segment à l'exclusion d'une part de son contexte.

8.2.3 Manifestations discursives des thèmes composites

Un premier exemple de thème composite discursif peut être observé dans l'extrait suivant, déjà discuté à plusieurs reprises :

> { { { Dans l'enseignement primaire$_{P1}$, on assiste à une forte diminution du taux de retard scolaire dans les années 80.}$_{U1}$ Cette baisse est en partie attribuable à la réduction du nombre d'élèves par classe, qui [...] }$_{S1}$ { Dans le secondaire$_{P2}$, on assiste au contraire à une augmentation sensible du taux de retard. Celle-ci est principalement imputable à [...] }$_{S2}$ }$_{S0}$

Dans ce cas, le passage est structuré par deux cadres de discours, introduits par P_1 et P_2. Cette configuration peut être vue comme un « cas d'école », puisque le mode de persistance de chaque composant topical est « typique » de sa fonction thématique : le noyau fait l'objet d'une chaîne de référence, alors que les satellites bénéficient de la portée propre aux adverbiaux détachés en initiale. Dans cette situation, les configurations phrastiques et discursives sont analogues, et le rôle de chaque constituant topical est identique selon que l'on le considère seulement dans la phrase qui l'héberge ou relativement à un segment plus grand. Par exemple, nous avons ici $\mathcal{T}(U_1, U_1) = \mathcal{T}(U_1, S_1) = \mathcal{T}(U_1, S_0)$.

Pour cette raison, le modèle de l'encadrement joue un rôle central dans notre approche des thèmes discursifs. Au sein de la typologie établie par Charolles, les cadres qui nous intéresseront particulièrement sont dits « véridictionnels », et correspondent plus particulièrement aux cadres temporels, spatiaux, représentationnels, ou encore praxéologiques. Mais les cadres dits « thématiques » (introduits par des constructions de type « as-for »), peuvent également jouer un rôle comparable dans certains cas, comme on peut l'observer dans l'exemple suivant :

> Le Conseil de sécurité des Nations Unies a créé en 1991 une Commission spéciale (UNSCOM) en charge du désarmement de l'Irak, tout en confiant le volet nucléaire à l'AIEA. [...] Huit années plus tard, quel bilan peut-on dresser ? [...] { En ce qui concerne le nucléaire$_{PA}$, l'AIEA estime avoir épuisé dès 1995 sa mission de destruction et d'enlèvement des matières prohibées [...]. }$_{S1}$ { Dans le domaine biologique$_{P2}$, l'UNSCOM a procédé à l'élimination de [...] }$_{S2}$

Cet exemple illustre la fonction thématique de satellite qui peut parfois être attribuée à un introducteurs thématique : dans ce cas, cette construction participe à l'élaboration du thème composite $\mathcal{T}(S_1)$ = LE DÉSARMEMENT DE L'IRAK $\bullet\!\!-\!\!\circ$ (NUCLÉAIRE). On notera pourtant que Charolles ne considère pas les cadres thématiques comme véridictionnels. Il est de fait très clair que le statut de satellite ne peut pas toujours être attribué à ce type d'introducteurs : dans certains cas, et tout particulièrement en langue peu soutenue, ils peuvent effectivement être explicitement repris, et ainsi introduire un noyau thématique. Par exemple :

> En ce qui concerne le nucléaire, il pollue tout autant que [...]

Mais il apparaît cependant que dans de nombreux autres cas, cette construction est utilisée pour attribuer le statut de satellites à des expressions à partir desquelles on ne peut construire d'expression adverbiale (par exemple *« dans le nucléaire » or *« en nucléaire » ne sont pas acceptables). Nous serons donc amené à considérer certaines introducteurs en « as-for » comme véridictionnels et non thématiques, c'est-à-dire comme définissant un satellite thématique et non un noyau. Dans d'autres cas en revanche, une reprise explicite nous invitera à les considérer comme introduisant un cadre proprement « thématique », et donc comme introduisant un noyau (il nous semble toutefois que ce cas reste peu fréquent à l'écrit).

Terminons enfin, pour ce qui est des cadres de discours, avec la possibilité de trouver des configurations caractéristiques d'un domaine, non pas seulement au niveau des concepts évoqués mais bien au niveau de la construction elle-même. C'est par exemple le cas dans l'extrait suivant, du domaine médical, dont le thème composite pourrait être ADMINISTRATION INTRAVEINEUSE D'OCYTOCIQUES $\multimap$ (CHEZ LES PATIENTES À RISQUE HÉMORRAGIQUE) :

> Chez les patientes à risque hémorragique, l'administration intraveineuse d'ocytociques avant ou après la sortie du placenta n'a pas réduit l'incidence des hémorragies de la délivrance ni la durée de la délivrance du placenta. [...]
>
> Source : MED

Même s'ils revêtent une importance capitale dans notre approche, nous considérons les cadres de discours comme une configuration parmi d'autres relativement au problème des thèmes composites, et nous considérons qu'il existe une variété de configurations susceptibles produisant des phénomènes équivalents. L'extrait suivant fait par exemple intervenir une construction clivée :

> { En théorie, le district de Lüchow-Dannenberg dispose dans tous les domaines d'un potentiel suffisant pour couvrir la totalité de son approvisionnement énergétique à partir de ressources renouvelables. { C'est dans le secteur de l'électricité$_{P1}$ que cet objectif est le plus facile à réaliser. [...] }$_{S1}$ { Un inconvénient pour le secteur thermique$_{P2}$ tient au fait que le district ne dispose de pratiquement aucun réseau de chauffage urbain. }$_{S2}$ { Les gros clients qui seraient à même de rentabiliser l'exploitation de la géothermie$_{P3}$ font eux aussi défaut. }$_{S3}$ }$_{S0}$

Voici l'analyse en thème composite que l'on pourrait donner pour ce passage, en notant E_0 le noyau thématique L'APPROVISIONNEMENT ÉNERGÉTIQUE DU DISTRICT DE LÜCHOW-DANNENBERG :

$$\begin{cases} \mathcal{T}(S_0, S_0) = E_0 \multimap (\langle\langle \text{MODES D'APPROVISIONNEMENT ÉNERGÉTIQUE} \rangle\rangle) \\ \mathcal{T}(S_1, S_0) = E_0 \multimap (\text{ÉLECTRIQUE}) \\ \mathcal{T}(S_2, S_0) = E_0 \multimap (\text{THERMIQUE}) \\ \mathcal{T}(S_3, S_0) = E_0 \multimap (\text{GÉOTHERMIQUE}) \end{cases}$$

Ici comme dans le précédent exemple, le noyau thématique est introduit en amont des sous-segments régis par les satellites, pour former une structure quasi-énumérative. Le premier satellite apparaît sous la forme d'une construction clivée qui, renforcé par un superlatif, annonce l'occurrence d'entités sémantiquement comparables à celle réalisée par P_1 (entité dorénavant notée E_1). C'est effectivement le cas avec P_2 et P_3, même si la saillance de ces expressions est remarquablement faible. Il est particulièrement intéressant d'observer que la saillance de P_1 (vraisemblablement attribuable à la construction clivée), et donc de l'axe sémantique auquel il appartient, semble rendre acceptable l'absence de marque explicite comme « au contraire » ou « de même » pour introduire S_1 et S_2 : le lecteur initié sera probablement dans l'expectative d'une ou plusieurs entités sémantiquement comparables à E_1, et leur occurrence effective paraît suffire à faire apparaître la structure du discours. En d'autres termes, la saillance d'un axe sémantique semble pouvoir participer très activement à l'émergence de la structure discursive.

On voit donc que si le processus de satellisation passe souvent par différentes formes de détachement, cela n'est pas systématiquement le cas. On peut en outre observer d'autres configurations discursives au sein desquelles ces satellites non détachés bénéficient d'une réelle portée, comme nous allons le voir dans la section suivante.

8.2.4 Détachement, satellisation et portée

Considérons maintenant ce cas, qui est une version légèrement modifiée du premier extrait de la section précédente :

> § { { L'enseignement primaire$_{P1}$ a connu une forte diminution du taux de retard scolaire ces dernières années. }$_{U1}$ Cette baisse est en partie attribuable à la réduction du nombre d'élèves par classe, qui [...] }$_{S1}$ { Dans le secondaire$_{P2}$, on assiste au contraire à une augmentation sensible du taux de retard. Celle-ci est principalement imputable à [...] }$_{S2}$

Dans cette version, S_1 n'est plus introduit par un introducteur de cadre *stricto sensu* : « l'enseignement primaire » apparaît ici comme sujet de la prédication, et n'est donc plus syntaxiquement détaché. Il est toutefois évident que, tout comme dans l'exemple précédent, P_1 fait ici écho à P_2, et que l'ensemble du passage reste organisé pour opposer ces deux niveaux du système scolaires. La fonction discursive de P_1 paraît analogue à celle du premier introducteur de la version précédente, dans la mesure où il spécifie bien un critère d'interprétation s'appliquant au propos central du discours (« le retard scolaire »), et que ce critère vaut pour plusieurs propositions sans être explicitement repris.

Tout se passe donc comme si P_1 bénéficiait d'une portée comparable[3] à celle d'un introducteur syntaxiquement détaché, et nous le considérons donc ici comme une forme spécifique d'introducteur d'univers, que nous dirons « intra-prédicatif » (dorénavant IU$_{IP}$). Notre hypothèse est que nous sommes ici en présence d'une structure discursive fonctionnellement équivalente à la précédente, et que P_1 y joue bien un rôle analogue à celui d'un introducteur. Différents facteurs semblent pouvoir expliquer ce phénomène.

i) En premier lieu, il convient ici de considérer avec attention l'antécédent du syntagme pronominal « cette baisse » : il est clair dans ce cas qu'il ne reprend pas seulement le référent de la « forte diminution du retard scolaire », mais bien l'ensemble du contenu propositionnel de l'énoncé qui précède (U_1), qui pourrait s'exprimer par « la diminution du retard scolaire dans le primaire ». De ce fait, on peut considérer que l'objet sémantique auquel se rapporte la chaîne de référence du segment S_1 est bien un structure complexe, centrée sur la « forte diminution », mais emportant avec lui « le primaire ».

ii) En second lieu, la forme même du contenu propositionnel de U_1 est particulière. En effet, l'acception ici employée du verbe « connaître » correspond ici à un *méta-prédicat* dont le second argument est lui-même un prédicat. Celui-ci est exprimé par la nominalisation du verbe « diminuer », et son argument est spécifié sous la forme du complément du nom « le taux de retard scolaire ». Or, ce méta-prédicat est neutre, et la structure sémantique résultante peut être « réduite » sans perte d'information, comme nous l'avons représenté dans la figure 8.2, en une autre structure dont la formulation la plus immédiate serait « le taux de retard scolaire a diminué dans le primaire ».

Très probablement, le choix par le scripteur d'une construction du type « X a connu Y » dans un cas comme celui-ci vise la topicalisation de X, qui apparaît ainsi en initiale. Toutefois, du fait de son rôle sémantique qui demeure « périphérique », il semble que cette topicalisation ne suffise pas ici à définir l'à propos de l'énoncé, tout comme un adverbial détaché ne définit pas nécessairement le thème au sens de l'à propos, comme le remarque Charolles dans (Charolles, 2003) (cf. section 3.4.1). Ainsi, dans notre exemple, « l'enseignement primaire » est topicalisé sans pour autant constituer le noyau thématique du segment S_1, ce qui participe à son accès à la fonction de satellite. On peut donc voir ici une forme de pseudo-détachement que l'on pourrait qualifier de « détachement sémantique ».

iii) Il est également possible de faire apparaître ce phénomène de pseudo-détachement à l'aide de la théorie du centrage agrémentée de la notion de « coût » introduite par Strube et Hahn (cf. section 3.4.2). Pour argumenter ce point, annotons ainsi le début de notre exemple :

[3] Sans qu'elle soit nécessairement équivalente, c'est-à-dire de même « force » cadrative.

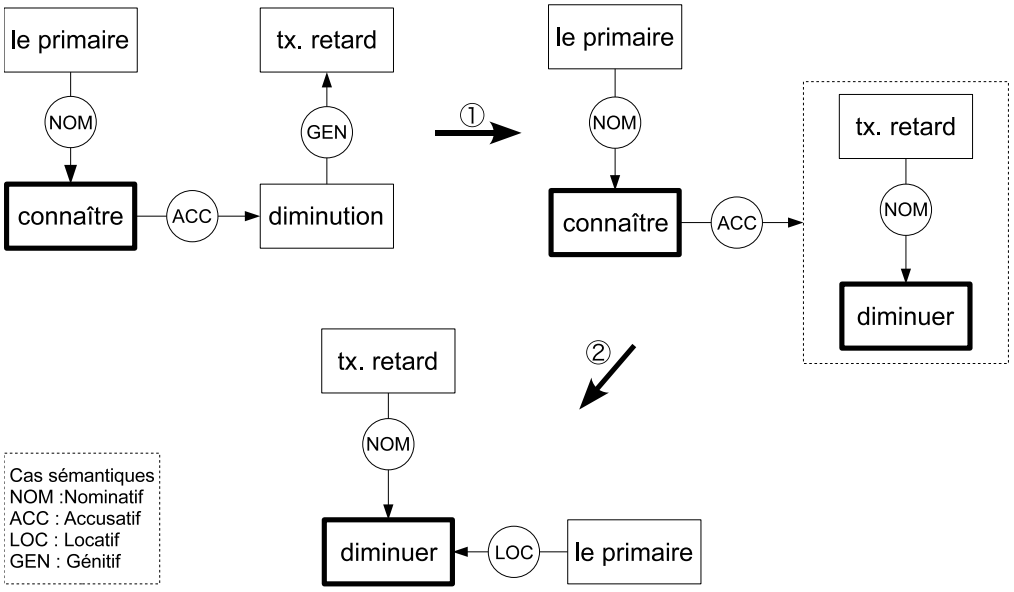


FIG. 8.2 – Transformations de la structure sémantique associée au méta-prédicat « X a connu P ».

§ { L'enseignement primaire$_{P1}$ a connu une forte diminution du taux de retard scolaire ces dernières années. }$_{U1}$ Cette baisse$_{P2}$ est en partie attribuable à la réduction du nombre d'élèves par classe, qui [...] Dans le secondaire$_{P3}$, on assiste au contraire à [...]

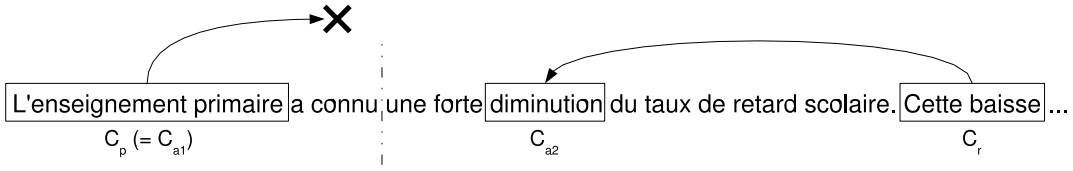
Soient E_1 l'entité réalisée par le syntagme P_1, E_2 l'entité réalisée par les syntagmes P_2 et P_3 (E_2 est bien l'élément le plus central de l'antécédent de P_3, même si, comme nous l'avons remarqué en (i), il ne s'y réduit pas). Dans les termes posés par la théorie du centrage, l'ensemble des centres anticipateurs de l'énoncé U_1 est $C_a(U_1) = \{E_1, E_2\}$. D'autre part, si l'on s'autorise à appliquer au Français la règle de d'ordonnancement donnée dans (Grosz *et al.*, 1995) pour déterminer le centre préféré (sujet > objet(s) > autres)[4], le centre préféré de U_1 est $C_p(U_1) = E_1$. Enfin, puisque l'énoncé U_1 n'est précédé d'aucun autre (il est ici en initiale de paragraphe), son centre rétroactif est indéterminé : $C_r(U_1) = \varnothing$. Pour l'énoncé U_2, le centre rétroactif est $C_r(U_2) = E_2$, et pour les même raisons que précédemment, son centre préféré est $C_p(U_2) = E_2$.

Nous nous trouvons donc dans la situation suivante : $C_p(U_2) = C_r(U_2)$ et $C_r(U_1) = \varnothing$. Selon les critères donnés dans (Walker *et al.*, 1998), il s'agit d'un cas de continuation, qui correspond en l'occurrence à l'instauration de E_2 comme noyau thématique. Toutefois, on ne pourra que convenir du statut particulier de l'entité E_1 qui constitue, du point de vue du centrage d'attention, le centre préféré de U_1. Or un centre préféré constitue par définition « une prédiction sur le centre rétroactif de l'énoncé suivant » (*ibid.*), prédiction ici contrariée par le fait que E_1 n'est pas repris directement dans l'énoncé U_2 : $C_r(U_2) \neq C_p(U_1)$.

Cette configuration, qui nous intéresse ici tout particulièrement, n'est pas explicitement discutée dans (Grosz *et al.*, 1995), mais est en revanche examinée dans (Strube et Hahn, 1999) autour de la notion de *coût* attribuée aux transitions entre énoncés. Selon les critères fournis par ces auteurs, que nous avons déjà détaillés dans la section 3.4.2, nous observons ici une continuation sur E_2 dont la particularité est justement d'être initiée par une relation $\langle U_1, U_2 \rangle$ coûteuse. Ceci implique que l'interprétation de cette transition nécessite un effort cognitif particulier, qui nous semble renforcer l'impression de dé-

[4]Cette règle n'est pas donnée comme complète ni universelle, mais paraît suffisante dans le cas simple qui nous occupe ici, et les autres règles formulées par la suite aboutiraient ici à la même conclusion.

tachement de P_1, comme le figure le schéma suivant, détachement que l'on pourrait cette fois qualifier de « référentiel ».



iv) Enfin, la portée de P_1 nous semble également explicable en recourant à la notion d'univers « virtuel » proposée par Charolles (cf. section 3.4.1). Il se trouve en effet que le syntagme « dans le secondaire », dont on ne peut douter du statut d'introducteur, projette un univers parent implicite lié à l'ensemble des niveaux scolaires du système éducatif français (pré-élémentaire, primaire, secondaire, supérieur). Or il se trouve que « le primaire » est bien un univers dérivé de cet univers parent, ce qui nous incite probablement à le considérer de façon équivalente à l'univers du « secondaire ». Et le fait que l'introducteur au sens strict apparaisse après celui que nous qualifions d'intra-prédicatif ne semble pas problématique si l'on adopte comme Charolles une approche incrémentielle, qui prévoit « des possibilités de réanalyse a posteriori avec mise à jour des interprétations construites » (1997, p. 3). Nous rejoignons ici la notion d'écho entre ces deux univers, déjà mentionnée plus haut. Les connaissances de domaine semblent ici jouer un rôle important, puisque la familiarité supposée du lecteur avec les niveaux du système scolaire interviennent dans la reconnaissance de la structure discursive du passage comme quasi-énumérative.

Considérons maintenant deux autres exemples qui nous semblent relever du même phénomène. Le premier est en tout point semblable au précédent, mais fait cette fois intervenir des univers temporels :

> À la fin des années 80, Ullman estimait que ces deux modèles étaient même incompatibles, ce que confirmaient les faits puisque jusqu'alors les bases de données avaient été soit déclaratives mais orientées-valeur, soit orientées-objet mais non-déclaratives. Cependant, les années 90 ont vu apparaître plusieurs tentatives de conciliations, que nous présentons brièvement dans cette synthèse avant d'en développer deux plus longuement dans le reste de ce chapitre. Ces tentatives peuvent être réparties en deux domaines : les bases de données (monde système) et les bases de connaissances (monde IA). Au risque d'une simplification excessive, on dira que ces domaines se distinguent par le fait que le premier privilégie les aspects pratiques et l'efficacité, et le deuxième les aspects théoriques et l'expressivité.
>
> Source : SIL

Le second se distingue des précédents par le fait qu'il ne recours pas au procédé du meta-prédicat et que l'IU$_{IP}$ apparaît ici sous la forme d'une extension prépositionnelle dans le syntagme sujet :

§ Pour ce qui est du transport ferroviaire, la législation en matière de transport de marchandises dangereuses par rail a été renforcée et, à la suite de la scission de la société nationale des chemins de fer en 5 sociétés au début de 1999, la restructuration du secteur ferroviaire a été poursuivie en 2000. [...]

§ Les activités dans le domaine de la navigation intérieure ont fort souffert du blocage du Danube dû à la crise du Kosovo, ce qui a eu pour conséquence de priver ce secteur des ressources financières nécessaires à son adaptation à l'acquis de l'UE. Les aspects pratiques concernant la conformité des navires roumains aux normes de l'UE pourraient poser problème pour des motifs d'ordre économique, eu égard à l'objectif des autorités roumaines d'accès au Rhin. Un décret ministériel a été adopté afin de transposer les règles de l'UE relatives à l'accès à la profession de transporteur de marchandises par voie navigable. [...]

Source : ROU

Il est remarquable ici que malgré sa faible saillance au niveau phrastique, le syntagme « le domaine de la navigation intérieure » ait bien une portée significative. Alors que dans les précédents exemples le phénomène de pseudo-détachement semblait jouer un rôle non négligeable dans la perception du rôle des syntagmes concernés, cet IU_{IP} apparaît ici dans une position « syntaxiquement profonde », a priori peu encline à lui conférer une portée. Il semble pourtant que cette portée soit bien réelle, puisque si le champ lexical lié à la navigation est significativement présent dans texte qui suit, il n'y a aucune reprise du qualificatif « intérieure », qui est pourtant persistant. D'autre part, il est clair que cet IU_{IP} possède bien une fonction d'indexation au même titre que « le transport ferroviaire » auquel il répond. On peut raisonnablement supposer que dans ce cas l'apparition en initiale de paragraphe joue un rôle important, mais là encore la relation sémantique entre plusieurs introducteurs successifs et comparables au sein d'une structure plus globale semble à prendre sérieusement en considération.

De fait, nous formulons l'hypothèse que c'est essentiellement la relation sémantique forte et supposée connue entre ces introducteurs qui autorise l'un (ou même plusieurs) d'entre eux à apparaître dans une position qui n'est pas explicitement détachée. Et comme nous le verrons dans la section suivante, c'est ce dernier critère qui agira de façon prédominante dans la détection automatique de ces structures discursives particulières.

Plus généralement, nous défendons ici l'hypothèse que des constituants non détachés syntaxiquement peuvent dans certains cas constituer des introducteurs d'univers dotés d'une réelle portée, et spécifier à ce titre des critères d'interprétation portant sur plusieurs propositions sans faire appel aux mécanismes référentiels. Nous avons montré à travers plusieurs exemples que des mécanismes de pseudo-détachement peuvent intervenir (détachement « sémantique » ou « référentiel »), mais aussi que la saillance présupposée de certaines relations sémantiques avec un autre introducteur semble autoriser une absence de marque de détachement explicite.

8.3 Analyse automatique de structures en thèmes composites

Nous allons maintenant présenter la méthode d'analyse automatique que nous avons élaborée autour de la notion de thème composite. Précisons dès à présent que tous les critères considérés dans les parties précédentes ne trouvent pas nécessairement de pendant immédiat dans cette méthode. Cela est dû pour une part au fait que nous n'avons pas encore implémenté l'intégralité des règles qui découlent de nos observations, mais aussi et surtout parce que tous les phénomènes linguistiques susceptibles de jouer un rôle ne peuvent pas être directement mis en oeuvre automatiquement.

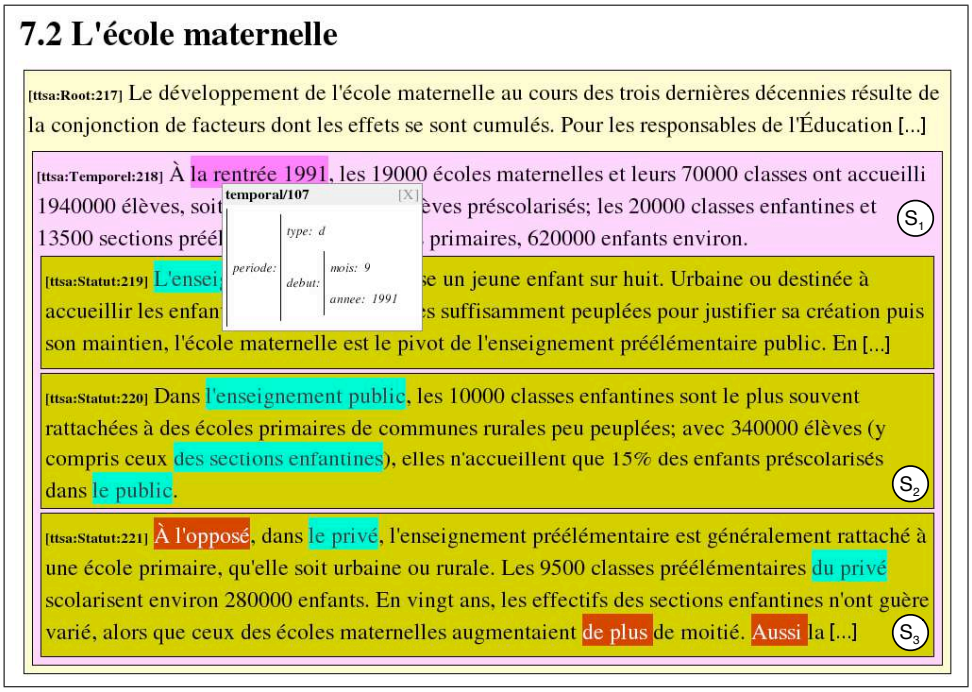
Considérons par exemple le cas des expressions satellites qualifiées, dans la section 8.2.4, d'introducteurs intra-prédicatifs, et qui posent un problème particulier à l'analyse automatique dans la

mesure où aucune marque de surface ne semble les caractériser. Parmi les différents critères que nous avons évoqués à ce sujet, se pose la question de ceux qui sont à la fois suffisamment généraux et analysables automatiquement. Sur ce point précis, on remarquera notamment que les détachements que nous avons qualifiés de « sémantique » et « référentiel » ne sont pas nécessairement de bons candidats, d'une part parce qu'ils sont difficiles à reconnaître automatiquement (car impliquant respectivement une analyse sémantique profonde et une détection fiable des chaînes de référence), mais surtout parce qu'ils ne semblent pas systématiquement associés aux IU_{IP} même si on les observe fréquemment. Il nous semble préférable ici de considérer que la présence d'une relation sémantique avec d'autres constituants comparables dans le discours environnant constitue un indice à la fois plus fiable et plus facilement repérable automatiquement. C'est d'ailleurs bien ce phénomène qui se produit (certes le plus souvent en conjonction avec d'autres) dans les différents exemples que nous avons envisagés jusqu'ici : chacun des satellites que nous avons décrits fait bien écho à une ou plusieurs autres entités du même ordre apparaissant dans le co-texte (droit ou gauche). On remarquera également que dans la plupart des cas, si certaines expressions satellites ne sont pas immédiatement identifiables comme telles, l'une des expressions auxquelles elles font alors écho est bien dans une position caractéristique.

Il semble que dans ce cas, la fonction de satellite d'un référent du discours puisse apparaître très clairement sans que sa textualisation fasse l'objet d'aucun détachement explicite, pourvu que l'on considère leur contexte et de potentielles relations sémantiques entre plusieurs satellites. Cette hypothèse est effectivement mise en oeuvre dans notre système d'analyse automatique, qui permet quand cela est nécessaire de tenir compte à la fois de schémas discursifs globaux et de connaissances d'ordre ontologique. Ces dernières sont formalisées sous forme d'axes sémantiques, qui correspondent à des espaces notionnels susceptibles de participer à l'indexation de l'information dans les textes considérés. Il pourra s'agir d'axes génériques comme le temps ou espace, ou d'axes plus spécifiques à un domaine ou à une pratique (axe des niveaux scolaires, des types de transports, etc.).

La méthode que nous avons expérimentée a pour particularité de débuter par une segmentation reflétant la structure satellitaire, avant de procéder dans un second temps à l'identification des noyaux au sein des segments obtenus. La première phase du traitement consiste à baliser les unités lexicales ou syntagmatiques susceptibles de jouer un rôle dans l'organisation du discours, sous la forme de connecteurs discursifs ou de termes apparaissant dans les axes sémantiques connus du système (les axes sont donnés sous la forme de liste d'entrées lexicales ou de grammaires sémantiques dans le cas du temps ou de l'espace). La segmentation proprement dite est fondée sur un catalogue de configurations discursives prédéterminées, ces dernières pouvant être réparties en trois catégories selon la nature plutôt formelle ou plutôt sémantique des critères qui les déterminent :

- Les configurations *explicites* décrivent des structures intégralement marquées par des indices accessibles par leur forme de surface. Il s'agira par exemple de séquences de cadres de discours (si la question de la portée des cadres est très complexe, leurs introducteurs n'en sont pas moins aisément détectables).
- Les configurations *mixtes* semblent les plus fréquentes, probablement pour des raisons stylistiques. Il s'agit de situations où la structure n'est pas intégralement marquée, et dont l'analyse requiert un recours aux connaissances du domaine, comme dans la plupart des exemples présentés dans la section précédente. Dans ce cas, le système cherche un terme apparaissant dans une position discursive caractéristique (introducteur, clivée, connecteur discursif, etc.) et appartenant à un axe sémantique connu. La détection effective d'une telle expression déclenchera la recherche dans son cotexte de la réalisation d'autres entités du même axe, pour procéder le cas échéant à la segmentation proprement dite.
- Enfin, les configurations implicites regroupent les cas où aucune marque discursive n'a pu être repérée. Ces passages pourront être segmentés si différents éléments d'un même axe sémantique y apparaissent simultanément. Cette situation reste cependant délicate, et il nous reste à déterminer précisément combien d'éléments d'un même axe devraient être présents au mini-



$$\begin{cases} \mathcal{T}(S_1) = \text{ÉCOLE MATERNELLE} \multimap (\text{À LA RENTRÉE } 1991, \langle\langle \text{STATUT} \rangle\rangle) \\ \mathcal{T}(S_2) = \text{ÉCOLE MATERNELLE} \multimap (\text{À LA RENTRÉE } 1991, \text{DANS LE PUBLIC}) \\ \mathcal{T}(S_3) = \text{ÉCOLE MATERNELLE} \multimap (\text{À LA RENTRÉE } 1991, \text{DANS LE PRIVÉ}) \end{cases}$$

FIG. 8.3 – Résultat d'une analyse automatique en thèmes composites.

mum, comment évaluer le bon équilibre des structures candidates, et comment traiter le cas où plusieurs axes sont simultanément représentés.

La structure que nous définissons étant par nature hiérarchique, le processus d'analyse doit inclure une forme de récursivité permettant de détecter les configurations enchâssées. Pour l'heure, le système considère un premier grain d'analyse choisi *a priori* (par exemple le paragraphe), au sein desquelles les configurations ci-dessus seront recherchées. Par la suite, le même processus est appliqué aux segments obtenus lors de cette première passe, et sera répété récursivement jusqu'à une profondeur préalablement choisie ou jusqu'à ce qu'aucune configuration ne soit plus reconnue.

Une fois cette segmentation obtenue, le système procède à l'identification des noyaux pour chaque segment afin d'obtenir une représentation complète de leur thème composite. Ceux-ci sont obtenus par une méthode quantitative basée sur le principe déjà évoqué dans la section 6.3 (et détaillé en annexe A), permettant d'obtenir pour chaque segment une liste de syntagmes distributionnellement saillants. Cet aspect combine ainsi la modélisation linguistique et une méthode numérique, mais il serait possible d'employer un procédé différent, qui pourrait par exemple se fonder sur une analyse plus fine de la structure informationnelle. Notons que si notre représentation des thèmes composites n'admet en principe qu'un seul noyau pour chaque segment, il semble intéressant dans la perspective de la recherche d'information de conserver plusieurs noyaux potentiels pour chaque segment. En effet, il ne s'agit pas tant d'obtenir une représentation idéale sur le plan linguistique que de produire une indexation susceptible de donner les meilleurs résultats possibles à l'utilisateur. Dans cette application, nous représentons donc chaque noyau par une liste de termes triée par ordre de pertinence.

La méthode a pour l'heure été testée sur différents corpus de 10 000 à 100 000 mots, et produit

d'ores et déjà des résultats intéressants bien que toutes les configurations discursives que nous avons identifiées n'aient pas encore été implémentées. Un exemple de texte analysé par le système est présenté en figure 8.3, où l'analyseur a détecté une structure arborescente basée sur le noyau thématique « l'école maternelle », décliné selon les axes du temps et du statut (public ou privé) des établissements scolaires. Précisons qu'il s'agit ici d'une sortie telle que visualisée dans un navigateur Web à des fins d'expérimentation, et non pas destinée à un utilisateur final. En revanche, les données ainsi visualisées sont directement applicables dans le contexte de l'accès assisté à l'information. Le système produit à cette fin un index intra-documentaire au format décrit dans la section 6.4 et exploitable par le moteur de recherche décrit dans cette même section.

Le procédé est réalisé sous la forme d'une chaîne LinguaStream, reproduite par la figure 8.4. À la suite des composants habituels qui procèdent aux découpages initiaux (mots, phrases) et à l'analyse morpho-syntaxique, le composant « SN Marker » se charge d'identifier les syntagmes nominaux à l'aide de la grammaire locale déjà mentionnée à plusieurs reprises (reproduite en annexe E.1). Par la suite, le composant « Axis Mapper » se charge de reconnaître les termes appartenant à un axe sémantique connu, étant donné un ensemble d'axes constituant les ressources propres au domaine (ici dans le fichier « scolaire.xml »). On notera que ce composant ne se charge que des axes qui sont spécifiés par des listes d'entrées lexicales. Dans le cas où une analyse spécifique est nécessaire (comme pour le temps ou l'espace, qui constituent des axes pertinents dans de nombreux cas), il est nécessaire d'insérer dans la chaîne des composant adaptés (non reproduits ici).

Dans un second temps, un lexique de connecteurs de discours est projeté sur le texte, et un ensemble de patrons donnés sous forme de règles MRE permettent de reconnaître les termes apparaissant dans une position caractéristique de la fonction satellite (introducteurs, clivées, initiale de paragraphe, etc.). À ce stade, sont donc marqués dans le texte d'une part les termes appartenant à un axe sémantique connu, et d'autre part ceux dont la fonction est marquée à la surface du texte.

La dernière phase du traitement, réalisée par le composant « Segmenter » procède à la segmentation proprement dite, en exploitant les indices préalablement relevés et en cherchant à reconnaître les instances des configurations évoquées plus haut. Il s'agit pour l'heure d'un composant *ad-hoc*, chaque configuration particulière étant « implémentée » par une classe Java spécifique. De façon à exprimer ces configurations par des règles déclaratives, nous chercherons à l'avenir à mettre à l'oeuvre le langage CDML (déjà évoqué au sujet de l'analyse des cadres, et présenté dans la partie suivante), qui permettra dans une version ultérieure d'exprimer sous forme de « grammaire de discours » l'ensemble des configurations qui nous intéressent ici. Nous nous baserons alors sur les mêmes traitements en amont, seul le composant « Segmenter » étant à remplacer.

On notera que pour l'heure, nous ne disposons concernant la portée des expression satellites « que » du procédé d'analyse des cadres temporels discuté dans le chapitre 7. Nous ne disposons donc pas de méthode générique pour évaluer ces portées, ce qui se traduit pour l'heure par le fait que la borne finale du dernier segment de chaque configuration se termine nécessairement avec le segment englobant. Nous évoquerons toutefois plus loin la possibilité d'exploiter les résultats d'une analyse rhétorique dans ce contexte, en utilisant par exemple les techniques proposées par (Widlöcher, 2004).

On remarquera également que la chaîne ici reproduite ne gère pas, sous cette forme précise, la récursivité des structures à analyser. En effet, il est pour l'heure nécessaire de répéter le composant « Segmenter » pour obtenir une structure telle que celle présentée en figure 8.3 : la première passe s'attachera à un grain maximal fixé à l'avance (typiquement le paragraphe ou la section), alors que les suivantes analyseront les segments établis lors des passes précédentes. Cette solution est bien sûr assez peu satisfaisante, et nous chercherons à l'avenir à donner les moyens à notre plate-forme de gérer par elle-même cette récursivité, soit au niveau des chaînes de traitement, soit au sein du langage CDML.

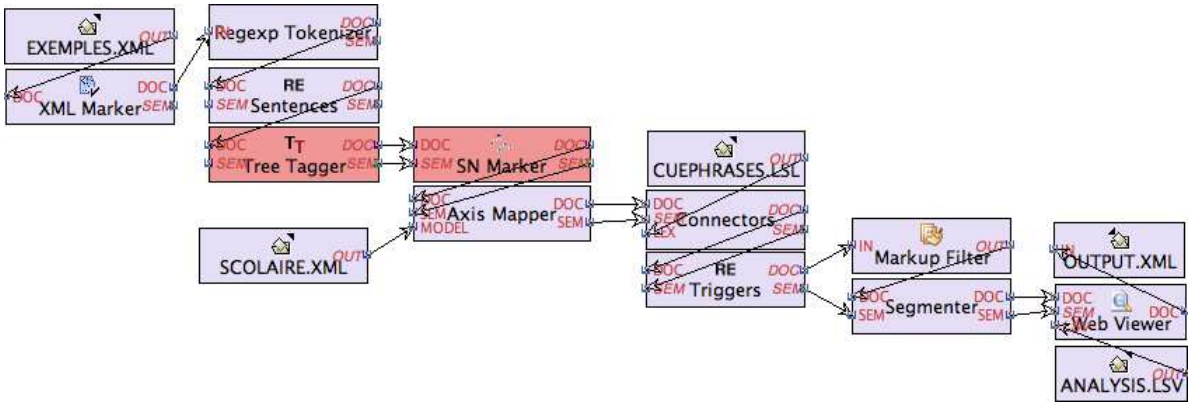


FIG. 8.4 – Chaîne de traitement de l'analyseur

8.4 La notion d'axe sémantique

> Un domaine (ou une sous-section de domaine) n'est accessible mentalement que si le champ notionnel est structuré, c'est-à-dire s'il constitue ce que l'on appelle un système de notions. Dans cet ensemble, chaque notion révèle ses rapports avec les autres notions. (Felber, 1987)

Notre démarche de généralisation consiste comme on l'a déjà évoqué à décrire une certaine forme de *structure* de l'information propre à chaque un domaine. Plus précisément, il s'agit de prendre en compte le fait que pour un domaine donné, certains concepts peuvent jouer un rôle tout particulier dans l'organisation des connaissances, rôle que nous pourrions qualifier de « structurant » ou de « positionnant ». Il s'agirait d'un ensemble, lui-même structuré, de concepts susceptibles d'être couramment utilisés pour situer le « contexte » ou le « cadre » d'un phénomène, et ce de façon plus ou moins spécifique à un domaine. Cette relation de « positionnement » s'inscrit dans une perspective proche de l'ingénierie des connaissances, et peut être vu comme un type particulier de relation conceptuelle qui participerait à la structure globale du *système notionnel* propre à un domaine.

De toute évidence, ce sont d'abord les composantes spatiales et temporelles qui sont utilisées pour « situer » toutes sortes de faits, phénomènes ou événements, qu'ils soient géographiques ou non. L'idée de situation dans l'espace et le temps est d'ailleurs intrinsèquement liée aux notions d'événement et de phénomène (Bachta, 2002). Depuis Newton, un événement au sens physique du terme ne peut être conçu que relativement à un certain point de l'espace et du temps, et la notion de phénomène physique est intimement liée aux notions de mouvement et de durée, puisqu'elle implique une succession d'événements et donc un déplacement dans l'espace et/ou du temps. Même dans une perspective plus « subjectiviste », la notion de phénomène est également dépendante l'espace et au temps : chez Kant, ces derniers sont considérés comme « formes pures de l'intuition »[5], qui conditionnent et rendent possible l'expérience des objets et donc l'existence même des phénomènes qui désignent alors le produit de notre perception et notre entendement de ces objets. Dans tous les cas, la notion de phénomène appelle donc à une certaine forme de « positionnement », au moins par rapport au temps et à l'espace.

L'idée que nous défendons est que selon les domaines de connaissance, on pourra trouver d'autres entités conceptuelles capables de jouer ce rôle de « positionnement ». Prenons l'exemple du domaine de la scolarité en France. Le phénomène de « retard scolaire » pourra y être discuté relativement à différents lieux géographiques (selon les académies, par exemple) comme à différentes plages de temps, mais aussi relativement à différents niveaux scolaires : on pourra par exemple parler du retard scolaire

[5]Bien que Kant conçoive par ailleurs une différence de nature entre temps et espace, cf. (Kant, 1781).

θ **Public-privé, des rôles stabilisés ?**

§ Du début des années 1960 à la fin des années 1980, les effectifs de l'enseignement privé ont augmenté de 25 à 30%, proportion qui est voisine de celle de l'enseignement public. En maternelle, le privé paraît en recul, alors que dans le primaire il se renforce. Dans le second degré, le rôle des établissements privés a peu varié en importance relative dans les lycées du second cycle long ; il s'est nettement réduit dans le second cycle professionnel. On remarque que, dans l'ensemble, le privé a progressé dans les grandes villes, y compris en Méditerranée et tout autour de Paris, alors qu'il a nettement régressé dans les régions les plus rurales.

Source : HER

FIG. 8.5 – Exemple de structure discursive mêlant différents axes sémantiques.

« dans l'académie de Caen » ou « depuis les années quatre-vingt », mais aussi « dans le secondaire » (*vs.* « dans le primaire » ou « dans le supérieur »), ou « dans le privé » (*vs.* « dans le public »). Il est évident que ces positionnements « notionnels » ne sont ni universels ni liés à l'expérience individuelle au même titre que les positionnements spatiaux et temporels. Toutefois, ils nous paraissent jouer un rôle tout à fait similaire dans l'organisation des connaissances et dans la façon dont ils servent de support à la description des phénomènes. On pourra s'en convaincre en observant l'extrait reproduit dans la figure 8.5, où une localisation temporelle est spécifiée conjointement à différentes « positionnements » propres au domaine scolaire.

8.4.1 Hypothèse : la notion de concept structurant

Le parallèle entre positionnement notionnel et positionnement spatio-temporel nous semble principalement justifié par le fait qu'ils partagent la propriété d'être inhérents à la description de tout phénomène, fut-ce de façon latente. Il est certes possible de formuler une proposition au sujet d'un phénomène qui se veuille valable en toute généralité, indépendamment de tout contexte, mais il semble que pour la grande majorité des concepts que nous manipulons, ce contexte existe nécessairement, même s'il est implicite. On peut par exemple donner une définition du « retard scolaire » sans spécifier aucune référence spatiale ni temporelle, mais on ne pourra pas pour autant la considérer comme universelle dans la mesure où elle sera nécessairement, comme pour tout phénomène social, liée à un certaine région géopolitique (elle ne vaut que pour certains modèles sociaux) et à une certaine plage de temps (au moins contrainte par la nécessité de l'existence même du système scolaire auquel on se réfère). Il est clair que certaines notions plus « abstraites » pourront se soustraire à ce raisonnement, mais ce qui nous importe ici est qu'il existe bien une relation sémantique *latente* entre certains concepts et les dimensions du temps et de l'espace, idée que nous avons déjà abordé précédemment, notamment en citant Charolles qui considère que « le temps et le lieu sont des traits inhérents à tous les états de choses » (Charolles, 2003, p. 25).

C'est ici que nous rejoignons la notion de positionnement notionnel : de même que toute instance d'un phénomène est liée à certain lieu spatio-temporel, fut-ce implicitement, nous faisons ici l'hypothèse que certains concepts, considérés dans un contexte bien précis, sont liés *a priori* à d'autres concepts, ces derniers participant d'un « espace notionnel » dont les premiers ne peuvent s'extraire. Par exemple, si l'on se place à nouveau dans le contexte particulier du système éducatif français, la notion de « retard scolaire » est tout aussi indissociable de l'ensemble des niveaux scolaires que des dimensions spatiale et temporelle. De fait, toute description de ce phénomène sera soit explicitement relative à un certain niveau (primaire, secondaire, sixième, cinquième, etc.), soit valable pour une cer-

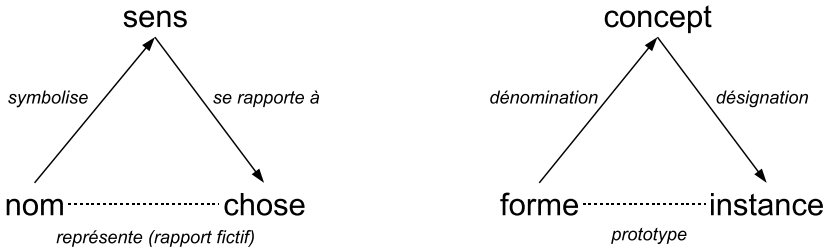


FIG. 8.6 – À gauche : triangle sémiotique d'Ullmann (1952). À droite : notre version ré-étiquetée.

taine « plage » implicite[6], soit valable par défaut pour l'ensemble des niveaux scolaires.

On remarquera à ce sujet, comme nous l'avons déjà fait dans la section 3.4.1, que l'on peut très bien trouver en initiale d'un texte un adverbial comme « dans le secondaire », au même titre qu'une expression temporelle ou spatiale, pour peu que le texte soit *a priori* rattaché au domaine scolaire. C'est par exemple le cas dans l'extrait ci-dessous[7]. Ainsi, en reprenant l'argument de Charolles qui remarquait que le statut particulier des traits de l'espace et du temps se traduit par la possibilité de trouver les adverbiaux du même nom en initiale de texte (cf. section 3.4.1), nous pouvons par exemple considérer que dans le trait « niveau scolaire » est inhérent aux concepts propres au domaine scolaire.

> θ **Coopération secondaire**
>
> § Dans le secondaire, la politique du BCF s'organise autour de trois axes : la formation continue, la promotion des Technologies de l'Information et de la Communication (TIC ou TICE), et la promotion du français. [. . .]
>
> Source : AFT

Pour tenter d'étayer ce point, nous allons nous reposer sur une forme ré-étiquettée du triangle sémiotique d'Ullmann (1952), les deux versions étant représentées dans la figure 8.6. Le triangle d'Ullmann a pour propriété de considérer comme fictif le rapport (dit de représentation) entre une unité signifiante et son référent objectal, en adoptant l'hypothèse saussurienne voulant que « tout se passe entre l'image auditive et le concept » (Saussure, 1916). Nous nous rangeons ici volontiers à cette vision du triangle sémiotique, considérant que, hormis le cas particulier de noms propres, la fonction communicationnelle première du signe linguistique réside dans l'évocation du concept et non dans la dénomination de l'objet, même si les individus établissent vraisemblablement des liens direct entre signes et instances prototypique, en fonction leur expérience individuelle. En d'autres termes, si l'on cherche à décrire le processus permettant à deux individus de s'accorder sur un même référent, il nous paraît nécessaire de flécher le triangle à la façon de Ullman, partant du signe vers le sens puis du sens vers la chose.

Nous utiliserons en fait une variante de ce triangle, qui reprend le même principe mais étiquette ses pôles de façon différente. Précisons qu'il n'est pas question de proposer ainsi un n-ième schéma sémiotique généraliste, mais seulement d'utiliser un vocabulaire plus à même de rendre compte des mécanismes en jeu dans le cas particulier qui nous occupe ici. Pour cela, nous avons d'abord remplacé l'étiquette « nom » par « forme », moins nominaliste et susceptible de recouvrir des unités qui ne soient pas seulement d'ordre lexical. D'autre part, nous avons utilisé pour les deux pôles restants les étiquettes « concept » et « instance », en référence aux mécanismes classificatoires qui jouent

[6] Le terme de « retard scolaire » est par exemple employé préférentiellement au sujet du primaire ou du secondaire, et non de l'enseignement supérieur

[7] BCF : Bureau de Coopération pour le Français.

selon toute vraisemblance un rôle important dans l'organisation des connaissances et des langues. Le concept constitue donc la notion centrale de ce triangle, en tant que pivot entre la forme langagière et les instances de ces concepts, instances qui constitueront pour nous les référents du discours. Sur cette base, nous supposerons que (au moins dans certains cas), le processus permettant d'instaurer un référent dans le discours à partir des formes linguistiques inclut les deux étapes suivantes :

Dénomination : évocation d'un concept par un ou plusieurs signes, un concept étant représenté par l'ensemble des traits partagés par ses instances.

Désignation : identification d'une instance particulière (ou d'une spécialisation) de ce concept, par l'énumération de traits spécifiques, susceptibles de distinguer cette instance (ou cette classe d'instances) des autres instances du même concept.

Par exemple, dans la locution « cette chaise », on pourra distinguer l'étape de dénomination qui consiste à évoquer chez l'allocutaire le concept de chaise, de l'étape consistant en la reconnaissance d'une instance particulière que l'on désigne. C'est ce dernier mécanisme qui nous intéressera ici tout particulièrement. Il est clair que l'on pourra rencontrer une grande variété de modes de désignation, qui seront tout d'abord dépendants du contexte de l'énonciation. À l'oral, par exemple, nous ferons par exemple intervenir des gestes en complément de pronoms démonstratifs. Ou encore, comme dans « la petite chaise à côté de la table », différents traits susceptibles de conduire à l'identification de l'objet relativement au contexte. Dans tous les cas, il semble que ce processus de désignation procède par ajouts successif de traits discriminants, permettant de passer progressivement de la catégorie à l'objet, avec bien sûr tous les intermédiaires possibles. Nous nous intéresserons ici plutôt au cas de l'écrit, et le plus souvent à des concepts plus « abstraits », comme c'est le cas dans l'extrait précédent.

Observons dans cet extrait le processus d'instauration du référent partiellement identifié par « la politique ». Durant la phase de dénomination, différentes entités conceptuelles sont introduites, dont notamment « le secondaire », « la politique » et « le BCF ». Il nous paraît clair qu'à l'issue de cette première phase, aucun référent de discours n'est encore installé. Il est pour cela nécessaire de procéder à la phase de désignation, qui permet de situer le concept « la politique » relativement aux concepts qui l'entourent. Une première étape agit pour cela au niveau syntagmatique, sous la forme d'un complément du nom. Dans un second temps, l'adverbial détaché vient préciser cette désignation, aboutissant finalement à l'identification du référent du discours « la politique du BCF dans le secondaire ». On remarquera de façon très claire dans cet exemple que ce phénomène de désignation dépasse largement le plan lexical, puisque les niveaux syntagmatique et même phrastique doivent être pris en compte. Dans d'autres cas, il sera même nécessaire de considérer le niveau discursif, notamment pour tenir compte des « concepts structurants » introduits dans des introducteurs de cadres de discours.

Comme pour les dimensions de l'espace et du temps, mais cette fois au sein d'un domaine donné, nous considérerons donc qu'il existe un lien latent entre les concepts de ce domaine et un certain nombre d'autres concepts *structurants*, indépendamment du fait que ce lien soit rendu explicite ou non par des propositions. Ce lien latent interviendra dans le processus de désignation, qui permet de passer des concepts au référent du discours.

8.4.2 Des concepts structurants aux axes sémantiques

Une propriété essentielle que nous attribuerons à ces concepts structurants est l'appartenance à un *ensemble* de concepts comparables, ensemble que l'on pourrait appeler *paradigme* ou *classe sémantique*. En effet, si l'on considère que la fonction sémiotique de ces concepts est de « situer » d'autres concepts dans l'espace notionnel qu'ils définissent, cela implique qu'ils soient en relation avec d'autres concepts du même ordre : un concept structurant ne peut positionner un autre concept que s'il s'oppose à d'autres concepts structurants qui forment avec lui un ensemble cohérent d'éléments mutuellement

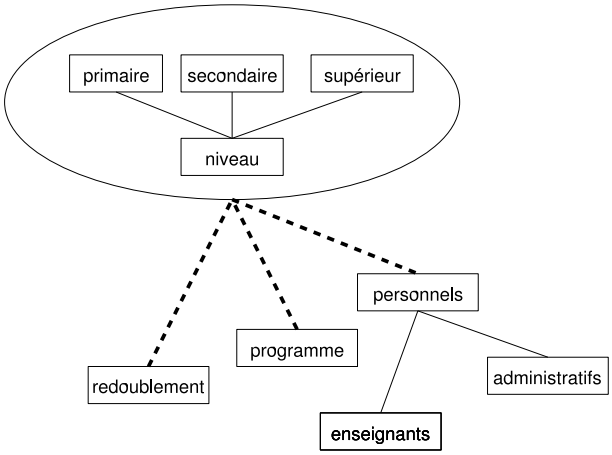


FIG. 8.7 – Exemple d'axe sémantique (inclus dans l'ovale) dans le réseau conceptuel du domaine scolaire. On observe en effet au sein du réseau, un sous-ensemble de concepts qui sont susceptibles de « situer » les autres.

exclusifs. Cela n'aurait par exemple aucun sens de considérer que « le secondaire » permet de *situer* des notions propres au domaine scolaire s'il ne s'opposait aux autres concepts d'un même paradigme, en l'occurrence « le primaire » et « le supérieur ». Par analogie avec la notion mathématique d'espace vectoriel, nous appellerons *axes sémantiques* ces ensembles conceptuels, considérant que le choix d'un concept au sein d'un axe sémantique est analogue au choix d'une coordonnée au sein de l'une des dimensions d'un espace où viendraient se positionner les concepts d'un domaine.

Précisons à toutes fins utiles que dans cette perspective, les « coordonnées » d'un concept dans un tel espace ne constitueront en rien une définition de ce concept, comme c'est le cas pour des approches purement vectorielles où les dimensions correspondent à des composantes sémantiques et où un concept est défini uniquement par ses coordonnées dans cet espace. Dans notre cas, nous considérerons que les concepts d'un domaine sont définis *en dehors* de l'espace défini par ses axes sémantiques, ce dernier servant seulement à positionner les différentes instances de ces concepts en discours, comme un objet physique vient se positionner dans l'espace, ou un événement dans le temps.

Un exemple d'ensemble de concepts structurants est donné dans la figure 8.7, qui représente une fraction d'un réseau conceptuel lié au domaine scolaire. La nature précise des relations qui lient les différents concepts n'est pas étiquetée, mais nous avons figuré en pointillés gras les relations latentes qui lient les différents concepts du domaine à l'axe sémantique des niveaux scolaires, et dont l'instanciation conduira au choix d'une ou plusieurs composantes de chaque axe.

On remarquera qu'un axe sémantique n'est autre, d'un point de vue « statique », qu'un cas particulier de classe sémantique. En tant que tel, il s'agit donc d'un objet largement étudié dans les différentes disciplines touchant à la sémantique, la terminologie et l'ingénierie des connaissances. Dans le cadre de la sémantique interprétative, où la notion de classe sémantique est définie de façon assez précise, on pourra par exemple considérer qu'un axe sémantique correspond à un *taxème* (cf. section 3.3.2).

Mais la notion d'axe sémantique apparaît également au sein des relations communément admises en ingénierie des connaissances, qui nous semblent tout aussi éclairantes dans ce cas particulier. Il s'agira notamment des relations issues d'une certaine approche de la terminologie qui s'intéresse plus particulièrement aux relations qui se placent sur un plan sémantique, en décrivant les rapports entre les notions (ou concepts) attachées aux termes, relations dites *notionnelles* par l'École de Vienne. Certaines sont applicables en toute généralité, comme les relations dites logiques ou taxinomiques (hyperonymie, hyponymie, co-hyponymie), ou celles dites ontologiques, notamment partitives (méronymie,

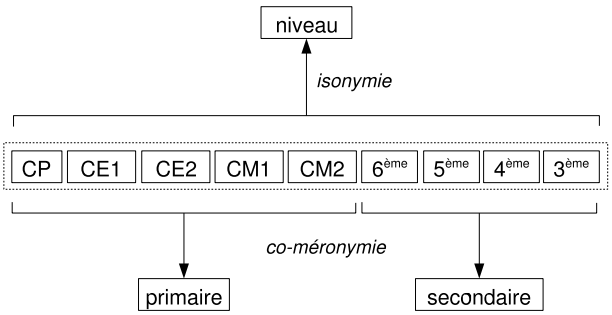


FIG. 8.8 – Axes sémantiques et relations terminologiques de coordination.

holonymie, co-méronymie). D'autres peuvent être définies relativement à un domaine de spécialité, et sont idéalement spécifiées de façon formelle au sein d'un système terminologique complet[8]. À partir des relations hiérarchiques telles que l'hyper/hyponymie et la méro/holonymie, on peut définir également des relations agissant entre des notions situées à un même niveau de l'arborescence (c'est-à-dire subordonnées à une même notion par une même relation). On parlera alors de co-hyponymie[9] ou de co-méronymie, et plus généralement de *relations coordonnées*. On trouve dans (Wüster, 1974) ou (Cruse, 1986) des éléments de caractérisation des relations entre notions coordonnées, comme le chaînage (ordonnancement séquentiel et ses dérivés : cycle, hélice, échelles), ou encore l'antonymie quand deux éléments sont coordonnés tout en étant liés par un trait distinctif.

La notion de coordination est généralement discutée dans la littérature relativement aux relations hiérarchiques classiques que sont l'hyperonymie et la méronymie. On peut toutefois l'étendre à toute relation terminologique, en considérant que deux concepts sont coordonnés dès lors qu'ils sont liés à un concept « pivot » par une même relation notionnelle. On pourrait alors donner une définition de la notion d'axe sémantique d'un point de vue terminologique : un axe correspondrait à un ensemble de concepts coordonnés (même si tout ensemble coordonné ne correspond pas nécessairement à un axe), dont un exemple est donné dans la figure 8.8.

Plus généralement, il est clair que différents modèles du sens pourront conduire à différentes caractérisations des relations qui unissent les composantes d'un axe sémantique, dont font partie les approches issues de la sémantique componentielle ou de la terminologie. Pour notre part, nous considérerons la notion d'axe sémantique comme largement indépendante de tout modèle sémantique particulier. En effet, ce qui nous semble primordial dans la caractérisation de cette notion ne réside pas tant dans les propriétés des relations entre composantes d'un même axe, que dans celles de la relation entre un axe et les autres concepts d'un domaine, que nous avons évoquées plus haut. Ce qui importe de notre point de vue réside donc plutôt dans la relation latente de « positionnement » entre les axes et les autres concepts que dans la nature précise des relations entre les composantes de ces axes qui est pour sa part fortement dépendante du modèle sémantique adopté.

Nous adoptons ainsi une démarche que l'on pourrait qualifier de fonctionnelle, tant au sujet des entités sémantiques qui nous intéressent ici que des structures linguistiques que nous décrirons plus tard. C'est cette démarche qui nous conduit à définir la notion d'axe sémantique par sa fonction au sein du réseau conceptuel auquel il appartient plutôt que par ses propriétés intrinsèques.

Corrélativement, nous considérerons que la fonction d'une entité sémantique doit être évaluée en contexte : il ne s'agit pas de postuler qu'une entité donnée remplit systématiquement une fonction structurante dans un domaine, mais plutôt de considérer qu'il existe un ensemble d'entités qui joueront *préférentiellement* ce rôle. Plus exactement, il s'agit d'établir une *échelle* permettant de situer le

[8] Voir par exemple le projet UMLS, « Unified Medical Language System ».
[9] Également appelée isonymie.

caractère plus ou moins structurant de chaque entité. Ainsi, si certains concepts apparaissent comme préférentiellement structurants pour un domaine donné, ils pourront néanmoins abandonner cette fonction pour constituer l'objet même d'un discours. Par exemple, bien que les syntagmes « le public » et « le privé » apparaissent le plus souvent sous forme prépositionnelle pour situer d'autres concept (comme dans « le retard scolaire dans le privé/public »), il pourront également constituer l'objet même d'une proposition ou d'un discours en tant que « phénomène », comme c'était le cas dans l'extrait de la figure 8.5.

Par opposition à la dimension « statique » que nous venons d'évoquer il nous faut donc insister sur la dimension « dynamique » des axes sémantiques. La dimension statique, déjà évoquée plus haut, permet de définir la fonction d'un axe au sein du réseau conceptuel d'un domaine. La dimension dynamique relève au contraire de la fonction des axes au sein de la structure du discours. En particulier, comme nous l'avons vu avec la notion de thème composite et de ses manifestations en discours, il s'agit avant tout de s'intéresser aux conséquences de l'apparition d'un axe sémantique en discours sur la structure de ce dernier. Nous avons notamment constaté que dans certains cas, la saillance d'un axe sémantique peut réduire le besoin d'un scripteur d'avoir recours à des connecteurs de discours explicites. Nous pouvons finalement caractériser ainsi les deux facettes qui fondent notre définition des axes sémantiques :

Dimension statique : un axe sémantique correspond à une catégorie sémantique jouant le rôle particulier de « positionnement » au sein d'un réseau conceptuel plus ou moins spécifique.

Dimension dynamique : cette catégorie sémantique est susceptible de jouer un rôle particulier dans la structuration du discours.

Ces deux facettes sont de notre point de vue d'importance équivalente, et c'est bien leur combinaison qui nous permet de caractériser l'objet « axe sémantique ». Un concept que nous disons « structurant » l'est donc tant pour le réseau conceptuel auquel il appartient que pour la structure du discours qui s'y rapporte.

8.4.3 Structure des axes sémantiques et système notionnel

Parmi les axes sémantiques que nous avons rencontrés jusqu'ici, on pourra distinguer deux catégories. La première regroupe les axes qui sont les plus proches du concept de « dimension » au sens mathématique, et que l'on pourrait qualifier de « continus ». Il s'agit d'axes qui correspondent effectivement à un espace continu, de dimension quelconque. C'est notamment le cas des axes du temps et de l'espace, le premier de dimension 1 et l'autre de dimension 2 ou 3 selon les cas. Un tel axe correspond donc plus exactement à un sous-espace, au sein duquel on peut situer une infinité de points ou de zones en utilisant un système de coordonnées. On notera que ce fonctionnement n'est pas réservé aux composantes temporelles et spatiales, d'autres espaces continus pouvant intervenir dans des domaines spécialisés. En voici un exemple, lié à la mesure de l'intensité des phénomènes sonores :

> Entre 85 et 130 décibels, le traumatisme peut entraîner une détérioration de l'audition. Une menace d'autant plus dangereuse qu'elle peut passer inaperçue. Suivant les individus, le risque est effectif dès 60 décibels. [...] Surtout, il est essentiel de savoir que la destruction des cellules de l'oreille interne par le bruit est irrémédiable. Ainsi, entre 130 et 150 décibels, le traumatisme peut entraîner une perte totale et définitive de l'audition.
>
> Source : YR

La seconde catégorie comprend les axes composés de concepts clairement distincts et non nécessairement ordonnés, axes que l'on pourrait qualifier de « discrets » par opposition aux axes « continus ».

En voici quelques exemples :

- Domaine politique :
  - Types d'élections : municipales, législatives, présidentielles, etc.
  - Tendances politiques : de l'extrême gauche à l'extrême droite.
- Domaine scolaire :
  - Niveaux scolaires : pré-élémentaire, élémentaire, secondaire, supérieur.
  - Statut des établissements : privé ou public.
  - Filières : générale, technique, apprentissage, etc.
- Domaine logistique :
  - Modes de transport : routier, ferroviaire, aérien, fluvial, maritime, etc.
  - Types d'inventaire : physique, cyclique, permanent, temps-réel, etc.

On remarquera que les relations sémantiques entre les concepts constitutifs de ces différents exemples d'axes sont loin d'être homogènes. Dans certains cas il s'agira de concepts totalement indépendants, éventuellement en opposition binaire. Dans d'autres cas, on pourra remarquer que les différentes notions ne sons pas réellement exclusives (ou pourrait par exemple ajouter le « ferroutage » dans l'axe des modes de transport). Dans d'autres cas, enfin, il existe une relation d'ordre (niveaux scolaires) voire même une quasi-continuité (tendances politiques). Néanmoins, et c'est comme nous l'avons explicité plus haut, ce qui nous intéresse ici, il apparaît que ces différents ensembles conceptuels revêtent bien une fonction similaire au sein de leurs systèmes notionnels respectifs, fonction que nos avons qualifiée plus haut de structurante.

Nous ne chercherons donc pas à formuler d'hypothèse trop restrictive quant aux liens sémantiques qu'entretiennent les différents concepts d'un même axe. Il est toutefois une relation importante dont nous ne pouvons faire l'économie dans le cadre de l'analyse automatique des thèmes composites, qui concerne un potentiel recouvrement (intersection ou inclusion) entre différentes notions appartenant à un même axe. Il est en effet nécessaire dans ce contexte d'être capable de déterminer si deux éléments d'un même axe sont ou non mutuellement exclusifs. Dans le cas des axes continus, cette relation est implicite, puisque tout système de coordonnées permet de mesurer le degré d'intersection entre deux régions (par exemple deux plages temporelles ou deux zones géographiques). Mais dans le cas des axes discrets, il faut expliciter ces relations quand elles existent. Cela sera par exemple le cas en ce qui concerne les différents niveaux scolaires, que nous avons représentés plus haut dans la figure 8.8 : on ne peut considérer comme deux axes indépendants l'ensemble des niveaux scolaires (pré-élémentaire, primaire, secondaire, etc.) et l'ensemble des classes qui les subdivisent (CP, CE1, CE2, 6e, 5e, etc.). Il s'agit au contraire d'une même échelle envisagée à deux niveaux de grain différents.

Pour cette raison, les axes sémantiques discrets devront donc être structurés de façon à pouvoir mesurer le degré d'intersection ou de recouvrement entre les éléments qui le composent. Dans la majorité des cas, nous pourrons pour cela utiliser une simple représentation arborescente basée sur la relation partie-tout, comme dans la figure 8.9. Nous nous sommes de fait limité à cette structure particulière dans l'implémentation que nous avons précédemment décrite, pour juger de l'exclusivité mutuelle entre deux satellites. Mais en toute généralité, la structure effective du sous-réseau conceptuel formant un axe sémantique ne devrait pas entrer en considération, pour peu qu'elle permette de mesurer le degré d'intersection entre ses composantes.

Précisons que dans le cas où un axe est représenté sous cette forme arborescente, la structure obtenue se rapprochera le plus souvent d'une *taxinomie*. Il nous faut cependant insister sur fait qu'il le ne s'agit là que d'un cas particulier, et que nous ne considérerons en aucun cas qu'une telle arborescence définit l'organisation des connaissances d'un domaine : il ne s'agit que d'un sous-ensemble structuré formant un seul axe. C'est seulement en discours, comme nous l'avons vu à travers plusieurs exemples, que le croisement des axes produira une structure arborescente complète.

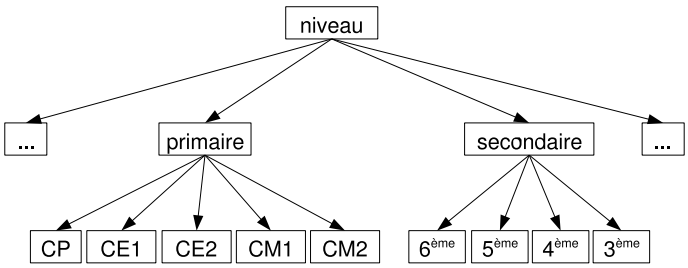


FIG. 8.9 – Représentation arborescente d'un axe discret.

C'est donc par un ensemble d'axes que nous serons amenés à caractériser un domaine, les uns étant continus et correspondant à un modèle sémantique spécifique (cf. section 6 pour ce qui est des expressions spatiales et temporelles), les autres étant discrets et généralement définis sous forme d'arborescente. Précisons que bien qu'il soit plaisant de considérer que les axes ainsi définis forment une *base* de l'espace notionnel dont ils définissent les *dimensions*, il ne s'agit que d'une acception très approximative du terme au regard de sa définition mathématique. En particulier, une propriété définitoire d'une base d'un espace vectoriel est d'être libre, c'est-à-dire qu'aucune de ses composantes ne doit pouvoir s'exprimer sous la forme d'une combinaison linéaire des autres composantes. En d'autres termes, aucun vecteur d'une base ne doit pouvoir s'exprimer en fonction des autres, puisque cela le rendrait superflu. Or, cette contrainte (ou plutôt son équivalent) ne sera pas nécessairement respectée dans le cas des axes sémantiques. En effet, il est tout à fait possible que le choix d'une composante ou d'une région au sein d'un axe se traduise par la sélection d'une composante ou d'une région au sein d'un autre axe. Par exemple, la sélection de la filière dite « technique » dans l'axe correspondant (cf. exemples ci-dessus) conduit nécessairement à sélectionner une région particulière de l'axe des niveaux, en l'occurrence le secondaire. De façon plus subtile, ce dernier axe pourrait également être lié à l'axe temporel, par exemple si l'on décrit le parcours scolaire d'un individu particulier.

8.4.4 Assistance à la constitution d'axes sémantiques

L'intervention de ressources d'ordre ontologique au sein d'un processus informatique est éminemment problématique, notamment en raison de leur non-universalité, et de la difficulté inhérente à leur constitution même. Pour palier à ces problèmes, nous avons envisagé la possibilité de concevoir un système de gestion dynamique des connaissances fondé sur la notion d'axe sémantique. Celui-ci garantira tout d'abord la flexibilité des ressources fournies : elles sont non seulement considérées comme relatives à un domaine, mais aussi à un utilisateur et même à une utilisation. En effet, pour différentes tâches de recherche documentaire, l'utilisateur pourra souhaiter adopter des points de vue différents sur un même domaine. Il disposera pour cela d'interfaces adaptées, permettant d'adapter rapidement les axes sémantiques considérés comme pertinents pour une utilisation donnée. Cette interface spécifique n'a pas encore été développée, mais nous évaluons pour l'heure la possibilité d'utiliser des outils génériques de manipulation d'ontologies tels que Protégé[10] qui permettent comme le montre la figure 8.10 de concevoir rapidement des interfaces d'édition *ad-hoc* (une simple feuille de style permet ensuite de transformer les fichiers RDF obtenus dans le format XML que nous utilisons pour représenter les axes sémantiques).

Mais aussi et surtout, la constitution des axes pourra être amorcée par un système d'extraction automatique d'axes sémantiques à partir d'un corpus homogène, afin de limiter les écueils inhérents à la constitution *ex-nihilo* de ces ressources. Nous avons pour cela développé un système d'extraction dont

[10]http ://protege.stanford.edu

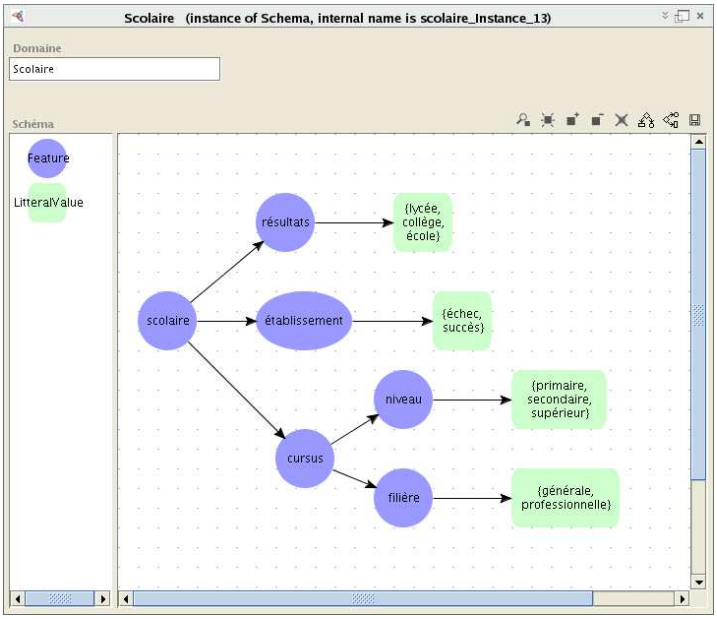


FIG. 8.10 – Application Protégé permettant de constituer des axes sémantiques.

l'originalité principale réside dans la prise en compte de la fonction discursive des termes rencontrés dans le corpus d'apprentissage. Ce procédé est par certains aspects très proche des systèmes d'extraction terminologique, mais s'attache cependant à un objectif sensiblement différent : il ne s'agit pas ici d'obtenir la couverture la plus large possible de l'espace terminologique du domaine, mais plutôt d'isoler les termes qui le structurent, par leur faculté à situer les autres termes dans un espace conceptuel. D'autre part, nous sommes ici particulièrement intéressé par les relations entre les termes extraits, puisque nous souhaitons *in fine* les regrouper en axes cohérents, ce qui nous rapproche plus particulièrement des méthodes d'extraction de terminologies structurées (Sta et Yildiz, 1997; Bourigault et Fabre, 2000), ainsi que de certaines méthodes de constitution automatique de classes sémantiques (Grefenstette, 1994).

De façon analogue à de nombreuses méthodes proposées dans ce domaine, notre méthode d'extraction peut se décomposer en trois phases : (i) identification des unités textuelles susceptibles de constituer des termes pertinents ; (ii) évaluation de la pertinence des unités candidates ; (iii) établissement d'éventuelles relations sémantiques entre les termes retenus.

L'ensemble du procédé a là encore été implémenté à l'aide de LinguaStream. La phase (i) opère à la suite d'une étape d'étiquetage morphologique et de lemmatisation, et procède au marquage des syntagmes nominaux et prépositionnels à l'aide d'une grammaire d'unification simple. Insistons sur le fait qu'il ne s'agit à ce stade que d'un balisage et non d'une extraction, puisque le contexte discursif des candidats sera déterminant lors des phases suivantes. Notons également que les syntagmes prépositionnels revêtent ici une importance particulière, puisque les termes agissant comme satellites apparaissent régulièrement sous cette forme (par exemple « dans le primaire », « dans le secteur thermique », etc.). Les deux phases suivantes font intervenir des critères portant sur la fonction discursive des candidats ainsi identifiés, afin d'évaluer leur caractère structurant d'une part, et pour les regrouper en axes sémantiques d'autre part.

Pour extraire les termes structurants – phase (ii) – nous nous appuyons sur l'hypothèse suivante : bien que les termes constitutifs des axes structurants n'apparaissent pas systématiquement dans des positions caractéristiques, ils y apparaissent plus fréquemment que les autres. De ce fait, l'analyse globale d'un corpus doit permettre d'extraire les termes susceptibles d'appartenir à des axes du domaine

étudié. Pour cela, notre procédé évalue la fonction discursive de chaque terme candidat, en combinant une certaine variété d'indices (leur pondération relative restant à déterminer plus précisément, nous les considérons pour l'heure comme équivalents). Là encore, nous nous reposons sur des analyseurs réalisés sous LinguaStream pour repérer dans le texte les différentes constructions mises en jeu. La valeur structurante d'un terme sera d'autant plus élevée que celui-ci apparaît fréquemment dans les positions suivantes :

- introducteurs de cadres véridictionnels (ou pseudo-thématiques) ;
- constructions clivées ;
- initiales de paragraphe ;
- extensions prépositionnelles de syntagmes nominaux ;
- titres de chapitres, sections ou sous-sections ;
- saillance distributionnelle locale.

Le système comptabilise les apparitions de chaque syntagme candidat dans ces positions caractéristiques. À l'issue de cette étape, il attribue ainsi à chacun d'entre eux une valeur numérique reflétant sa valeur structurante au vu du corpus analysé.

La dernière phase (iii) vise le regroupement en axes sémantiques des termes possédant les plus fortes valeurs structurantes. Les critères utilisés pour cela sont de deux ordres. Les premiers agissent au niveau syntagmatique : de façon (sur ce point) comparable à (Bourigault et Fabre, 2000), le système regroupe les termes apparaissant de façon fréquente en tant qu'extension prépositionnelle d'une même tête (par exemple « le primaire » et « le secondaire » s'ils apparaissent fréquemment en extension de la tête « le retard scolaire (dans X) »).

D'autre part, interviennent des critères d'ordre phrastique ou discursif, basés sur la notion de co-énumérabilité, communément admise comme indice de co-hyponymie (Sta et Yildiz, 1997). Là encore, la particularité de notre approche et de prendre en compte la dimension discursive de ce phénomène, et non pas seulement syntaxique. Nous évaluons ainsi les configurations suivantes :

- relations de coordination aux niveaux syntagmatique et phrastique ;
- énumérations sous forme de listes marquées typographiquement ;
- appositions uniformes de constructions discursives caractéristiques (séquences d'introducteurs de cadres de discours, de clivées, etc.) ;
- titres parallèles (de même niveau et appartenant à une même section englobante).

Finalement, les axes obtenus sont eux-mêmes pondérés par la moyenne des poids attribués à leurs composantes, ce qui permet de présenter à l'utilisateur des résultats triés par ordre de pertinence. La méthode (implémentée sous la forme d'une chaîne LinguaStream, cf. figure 8.11) – reste pour l'heure très expérimentale, mais produit d'ores et déjà des résultats encourageants. Le tableau ci-dessous présente des axes obtenus à partir de l'analyse de (HER) :

Rang	Termes
1	2e, 4e, CAP, 5e, âge, BEP, enseignement spécial, 4e de collège, 2e de lycée, 1ere S, section enfantine, terminale de lycée, terminale
2	2e, année, 5e, 3e, école élémentaire
4	enseignement privé, public
9	collège, primaire, communale, maternelle

8.4.5 Conclusion

Les notions de fonction structurante et d'axe sémantique sont donc intimement liées à celle de thème composite, et ce à plusieurs titres. En premier lieu, elles interviennent relativement à la notion

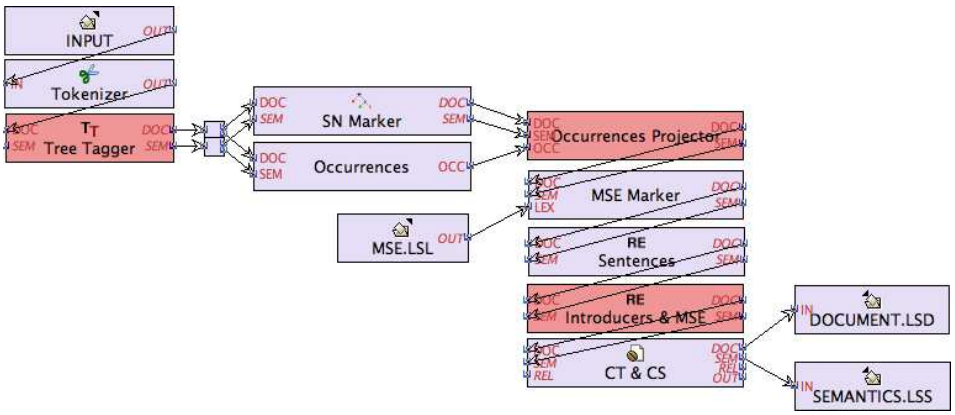


FIG. 8.11 – Chaîne de traitement utilisée pour l'extraction de termes structurants.

d'à propos : comme dans le cas du projet GeoSem, ont peut considérer que leurs constituants font partie intégrante de la description du contenu informationnel d'un texte ou d'un passage lorsqu'ils se prêtent à en description en thèmes composites. D'autre part, nous avons vu que les axes sémantiques d'un domaine influencent la structure du discours qui s'y rapporte, et interviennent de fait dans notre méthode d'analyse automatique de la structure thématique du discours.

Enfin, tout comme dans le cas particulier de l'information géographique, la notion d'axe sémantique peut intervenir dans le processus même de recherche d'information, le moteur de recherche que nous avons présenté dans la section 6.4 permettant de spécifier des critères propres à chaque axe du domaine de la base documentaire consultée.

8.5 Thèmes composites et structures rhétoriques

Comme nous l'avons dit plus haut, le modèle des thèmes composites vise à décrire la structure thématique de certaines configurations discursives, en se basant sur la notion d'*aboutness*. Notre analyse s'attache ainsi à la description d'une certaine part du contenu informationnel des textes, et à la répartition de ce contenu au fil du texte. D'autre part, les constructions discursives que nous décrivons s'insèrent également dans un schéma rhétorique qui n'est bien sûr pas déconnecté de ce que nous qualifions de structure thématique. Mais s'il est évident que ces deux « pans » de l'analyse du discours sont en interaction, les modalités régissant leur collaboration en discours n'en demeurent pas moins difficiles à cerner.

Déjà faut-il admettre l'idée qu'il s'agit bien de deux aspects distincts, même s'ils participent de concert à la cohérence du discours. C'est le point de vue que nous adopterons ici, en considérant que :

- L'analyse thématique vise à décrire la répartition du contenu informationnel d'un texte. Elle délimite des segments auxquels elle attribue une certaine représentation de leur thème (en termes d'à propos), et décrit d'éventuelles relations existant entre ces thèmes. Les objets obtenus sont donc des structures sémantiques ancrées sur les référents du discours, distinctes *a priori* des structures textuelles auxquelles elles se rapportent.
- L'analyse rhétorique vise à décrire l'organisation argumentative d'un texte. Elle délimite des segments auxquels elle attribue une certaine fonction de discours relativement à d'autres segments. Son produit est une structuration du texte, généralement arborescente, explicitant les relations de cohérence existant entre différents segments, généralement adjacents. Si ces relations sont par essence sémantiques, l'analyse ne vise pas à expliciter la valeur informationnelle des segments qu'elles régissent.

Nous ne prétendons évidemment pas élucider ici en toute généralité le problème de l'interaction entre ces deux facettes de l'organisation du discours. Il s'agit plutôt d'éclairer les rapports que peuvent entretenir nos thèmes composites avec des structures habituellement qualifiées de rhétoriques. Concernant ces dernières, et afin de donner un cadre précis à ces réflexions, nous nous en tiendrons ici à la Rhetorical Structure Theory (Mann et Thompson, 1987), déjà présentée en section 3.4.4. Nous allons ainsi comparer leur capacité à rendre compte des phénomènes qui nous intéressent. Il ne s'agit bien évidemment pas de les mettre en concurrence, mais de montrer en quoi les phénomènes représentés par les thèmes composites sont différents de ceux décrits par une analyse rhétorique de type RST.

Écartons tout d'abord la divergence qui concerne le degré de généralité visé. La RST est en principe capable de représenter l'organisation rhétorique de l'ensemble d'un texte, et ce pour une très grande variété de genres discursifs. Les thèmes composites, au contraire, ne sont applicables qu'à certaines configurations discursives particulières, qui apparaissent principalement dans les textes dits informatifs ou expositifs. Il serait donc nécessaire de les inclure au sein d'un modèle plus général pour obtenir la couverture complète de la structure thématique d'un texte. Pour cette raison, nous nous restreindrons le plus souvent dans cette partie à des structures discursives susceptibles d'être décrites à la fois en termes de thèmes composites et de relations rhétoriques. Nous pourrons alors expliciter d'autres différences, plus fondamentales, qui distinguent les deux approches.

Dans ce contexte, nous allons chercher à répondre aux questions suivantes :

- Nous nous interrogerons tout d'abord sur la représentativité d'une analyse RST relativement à la structuration en thèmes composites d'un même texte (quand il s'y prête). Nous verrons que les deux modèles décrivent des objets de nature différente, puisque les thèmes composites sont des structures sémantiques quand les structures RST sont des structures textuelles. Nous nous demanderons toutefois si, en dépit de cette différence de nature, il peut y avoir isomorphie entre les deux structures, et/ou si l'une peut se déduire de l'autre.
- Nous envisagerons par la suite le problème de la segmentation : même si les thèmes composites sont des objets sémantiques, ils n'en sont pas moins liés à une segmentation particulière du texte. Nous nous demanderons donc dans quelle mesure, du point de vue de la segmentation, les structures obtenues par une analyse thématique et celles obtenues par une analyse rhétorique sont superposables.
- Dans un troisième temps, nous chercherons à savoir si certaines structures rhétoriques pourraient être caractéristiques de structures en thèmes composites. Car même si les structures rhétoriques sous-jacentes aux thèmes composites sont variées, on ne peut exclure que certaines structures rhétoriques produisent systématiquement des thèmes composites.
- Enfin, nous verrons quel peut être l'intérêt de disposer d'une analyse rhétorique pour mener à bien une analyse thématique, dans le cas particulier des structures en thèmes composites.

8.5.1 Question de l'équivalence structurelle

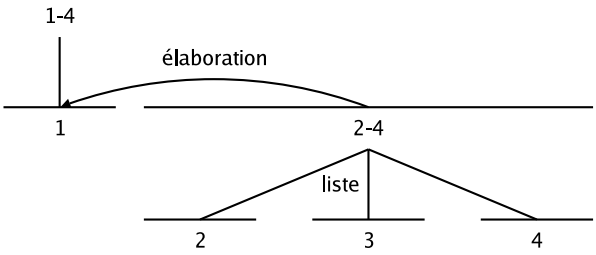
Dans cette partie, nous allons envisager la possibilité d'un isomorphisme entre la structure *sémantique* reflétant un thème composite et la structure rhétorique du texte correspondant. Bien qu'il s'agisse d'objets différents par nature, on ne peut exclure qu'ils prennent systématiquement des formes analogues (i.e. superposables), ou que l'on puisse déduire l'une de l'autre. Nous verrons toutefois que cela n'est pas le cas.

Considérons tout d'abord l'extrait suivant, déjà présenté en section 8.2 :

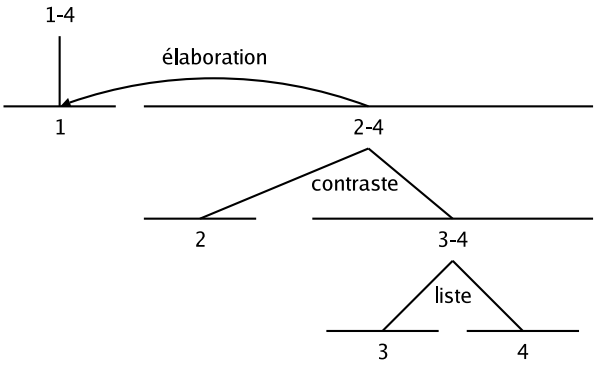
(1) En théorie, le district de Lüchow-Dannenberg dispose dans tous les domaines d'un potentiel suffisant pour couvrir la totalité de son approvisionnement énergétique à partir de ressources renouvelables. **(2)** C'est dans le secteur de l'électricité que cet objectif est le plus facile à réaliser. [...] **(3)** Un inconvénient pour le secteur thermique tient au fait que le district ne dispose de pratiquement aucun réseau de chauffage urbain. **(4)** Les gros clients qui seraient à même de rentabiliser l'exploitation de la géothermie font eux aussi défaut.

Source : EC

En se basant sur le catalogue de relations de la RST dite « classique » (cf. figures 3.12 et 3.13), un observateur pourrait en donner l'analyse suivante :



Ce diagramme exprime le fait que le segment 2-4 apporte des détails supplémentaires sur un fait plus général introduit par le segment 1 (relation d'élaboration). Le segment 2-4 est lui-même scindé en trois sous-segments organisés par la relation multi-nucléaire dite de liste, qui réunit trois items similaires. Une représentation plus fine des relations existant entre ces trois derniers segments est donnée dans cette seconde analyse, qui suppose une relation de contraste entre les segments 2 et 3-4 :

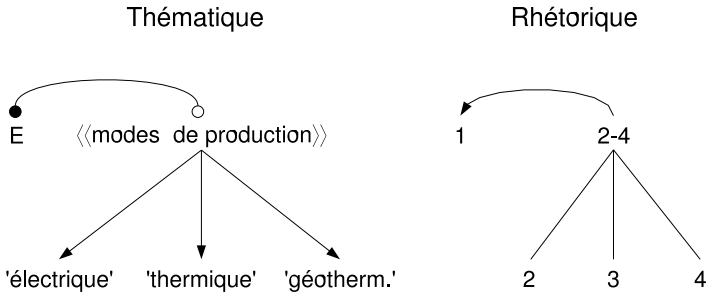


Par ailleurs, une analyse en thèmes composites du même passage (cf. section 8.2) pourrait être la suivante (avec E = La production énergétique du district de Lüchow-Dannenberg)[11] :

$$\begin{cases} \mathcal{T}(1\text{-}4, 1\text{-}4) = E \multimap (\langle\langle \text{modes de production énergétique} \rangle\rangle) \\ \mathcal{T}(2, 1\text{-}4) = E \multimap (\text{électrique}) \\ \mathcal{T}(3, 1\text{-}4) = E \multimap (\text{thermique}) \\ \mathcal{T}(4, 1\text{-}4) = E \multimap (\text{géothermique}) \end{cases}$$

Il faut bien remarquer que dans ce cas précis, les structures obtenues sont parfaitement analogues. Voici en effet ce que l'on obtient si l'on représente ces thèmes composites sous leur forme graphique, et la structure RST du même texte sous une forme légèrement épurée :

[11] Si nécessaire, on se rapportera à la section sus-citée pour obtenir les détails de la notation ici utilisée.

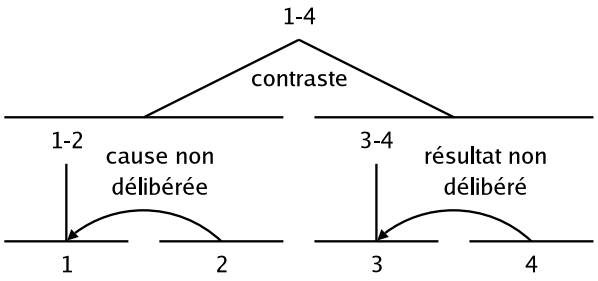


Pourtant, il apparaît tout aussi clairement que les informations exprimées dans chaque cas sont foncièrement différentes, conformément aux attendus de chaque analyse. La structure thématique, pour sa part, propose une représentation du contenu informationnel des segments délimités, qui n'apparaît aucunement dans l'analyse RST. Inversement, l'analyse rhétorique fournit des informations sur les relations qui lient les segments entre eux, y compris entre ceux qui constituent dans notre modèle des segments « satellitaires ». Par exemple, le fait que 2 et 3-4 soient en relation de contraste, explicite dans la structure rhétorique, n'apparaît aucunement dans notre analyse thématique[12]. Inversement, les critères sémantiques qui autorisent à placer ces segments sur un même plan ne sont pas explicites dans le schéma RST, alors qu'ils apparaissent dans l'analyse en thèmes composites, sous la forme d'un axe sémantique.

De fait il existe bien une différence fondamentale de nature entre une structure en thèmes composites et une structure RST : comme nous l'avons défini dans la section 8.2, et comme nous pouvons le constater en observant l'analyse ci-dessus, les thèmes composites explicitent des relations entre des *objets sémantiques*. De façon tout à fait différente, l'analyse RST établit des relations entre des *segments discursifs*. Et si ces deux approches mênent bien à des arborescences équivalentes pour l'exemple précédent, elles conduiront dans d'autres cas à produire des structures foncièrement différentes. C'est par exemple le cas pour le passage suivant :

> **(1)** L'enseignement primaire a connu une forte diminution du taux de retard scolaire ces dernières années. **(2)** Cette baisse est en partie attribuable à la réduction du nombre d'élèves par classe, qui [...]. **(3)** Dans le secondaire, on assiste au contraire à une augmentation sensible du taux de retard, **(4)** ce qui a eu pour principal effet de [...].

Une analyse RST acceptable pour ce passage pourrait être :

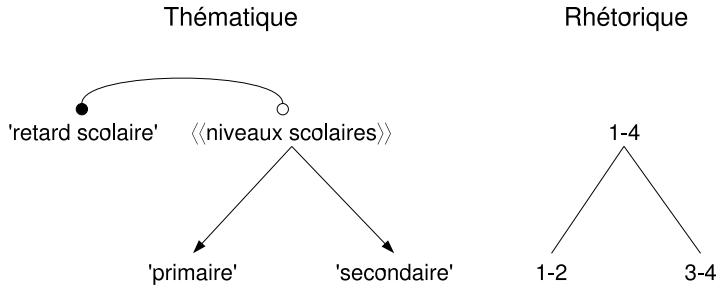


[12]Notons que l'observateur RST doit, en contrepartie, formuler des hypothèses sur la dimension intentionnelle du texte, alors que l'analyse thématique peut semble-t-il en faire l'économie. Ce point revêt une importance non négligeable dans le cadre de l'analyse automatique, les intentions du scripteurs étant par essence difficiles à objectiver. Rappelons que la RST a initialement été conçue pour des tâches de génération, qui présupposent une certaine modélisation des intentions du « scripteur artificiel », et ne posent donc pas ce problème.

D'autre part, l'analyse en thèmes composites que nous avions donnée dans une section précédente pour ce passage est la suivante :

$$\begin{cases} \mathcal{T}(1\text{-}4, 1\text{-}4) = \text{RETARD SCOLAIRE} \bullet\!\!-\!\!\circ (\langle\langle \text{NIVEAUX SCOLAIRES} \rangle\rangle) \\ \mathcal{T}(1\text{-}2, 1\text{-}4) = \text{RETARD SCOLAIRE} \bullet\!\!-\!\!\circ (\text{PRIMAIRE}) \\ \mathcal{T}(3\text{-}4, 1\text{-}4) = \text{RETARD SCOLAIRE} \bullet\!\!-\!\!\circ (\text{SECONDAIRE}) \end{cases}$$

On constate cette fois que les arborescences obtenues ne sont pas équivalentes, comme le montre le schéma ci-dessous. En particulier, il est intéressant de constater que la relation qui lie, dans la structure thématique, le noyau RETARD SCOLAIRE avec l'axe sémantique $\langle\langle$NIVEAUX SCOLAIRES$\rangle\rangle$, n'a plus d'équivalent dans la structure RST :

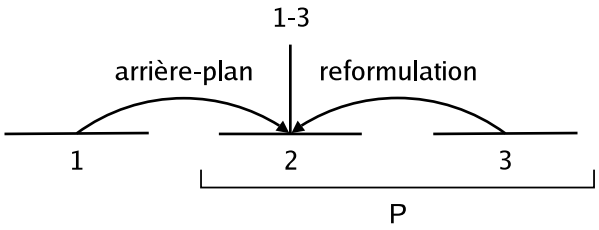


En comparant les deux derniers exemples, on constate donc que pour deux passages qui se prêtent à des représentations thématiques parfaitement isomorphes, on peut produire deux analyses rhétoriques totalement différentes. En outre, dans ce dernier cas, une relation essentielle sur le plan thématique ne trouve aucun équivalent dans la structure rhétorique. Tout cela tend à confirmer l'hypothèse qu'il s'agit de deux points de vue visant à représenter des réalités différentes, et que pour un texte donné, les structures obtenues ne peuvent simplement être considérées comme deux reflets d'un même phénomène.

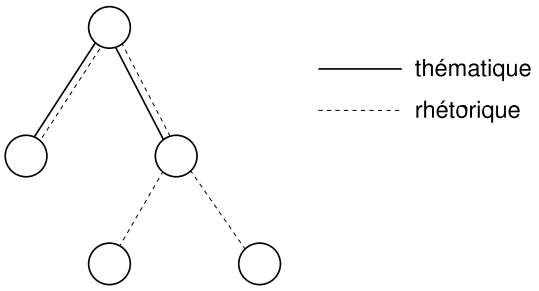
8.5.2 Question de l'équivalence des segmentations

Nous avons montré dans la section précédente des structures thématiques et rhétoriques représentant des phénomènes discursifs et sémantiques bien distincts. En revanche, pour tous les exemples proposés jusqu'ici, y compris dans le cas où les structures obtenues n'étaient pas équivalentes, les *segmentations* correspondantes demeuraient identiques. Nous souhaitons montrer dans cette section que cela n'est pas toujours le cas, les deux analyses pouvant parfois mener à des segmentations différentes.

En premier lieu, certaines relations rhétoriques définies par la RST semblent imposer une certaine continuité thématique. En d'autres termes, il s'agirait de relations « purement » rhétoriques, sans incidence sur l'organisation thématique. Par exemple, on peut raisonnablement considérer que les relations dites de reformulation ou de résumé, pour ne citer qu'elles, présupposent une certaine stabilité thématique. Une structure rhétorique pourra ainsi segmenter un passage P en deux segments reliés par une telle relation, alors que la structure thématique du même passage ne fera apparaître qu'un seul segment plus global. Cela pourrait par exemple être le cas dans la situation suivante :



Force est de reconnaître que le cas inverse paraît moins vraisemblable : on peine à imaginer deux segments thématiquement disjoints qui ne formeraient qu'un seul bloc du point de vue rhétorique. Cela pourrait nous conduire à la conclusion que la structuration rhétorique relèverait seulement d'un grain plus fin, et que les deux modes organisations, thématique et rhétorique, formeraient en fait une seule et même arborescence, dont l'une serait plus « profonde » que l'autre, comme l'illustrerait le schéma suivant :



Cette hypothèse ne nous semble toutefois pas vérifiée. Nous pouvons en effet trouver des contre-exemples, où les deux structures ne sont pas compatibles, c'est-à-dire qu'elles forment des arborescences non superposables. Observons pour cela l'exemple de la figure 8.12, que nous avons choisi parmi la base d'analyses RST disponibles sur le site consacré à cette théorie[13], afin d'éviter tout biais relatif à l'analyse rhétorique elle-même.

Le schéma de la figure 8.13 représente deux segmentations obtenues à partir de ce texte, l'une d'un point de vue rhétorique, et l'autre d'un point de vue thématique. Dans le premier cas, la segmentation reprend point pour point celle de l'analyse proposée par Mann et Thomson eux-mêmes, en omettant certaines subdivisions qui ne sont pas utiles à notre propos (l'analyse complète est reproduite en annexe F.2). L'analyse thématique est quant à elle de notre propre fait, et sera commentée ci-dessous. Dans les deux cas, le typage des relations entre segments a été omis puisque nous nous attachons ici uniquement à la segmentation du texte.

La structure thématique de l'extrait est manifestement liée à sa nature : en tant que résumé d'un article traitant de différents cadres théoriques linguistiques, le discours alterne entre la description de l'article lui-même et celle des différentes théories considérées. Ainsi, les segments 2-9 et 11-15 traitent des approches fonctionnelles en linguistique, alors que les segments 10-11 et 16-18 ont trait à la problématique de l'article résumé. Plus précisément, les « thèmes » que nous pouvons attribuer aux différents segments délimités sont les suivants :

- segment 2-3 : problématique linguistique ;
- segment 4-9 : approches fonctionnelles ;
- segment 10-11 : positionnement de l'article ;
- segment 12 : la Rhetorical Structure Theory ;
- segment 13-15 : linguistique systémique ;
- segment 16-18 : problématique de l'article.

[13]Cf. www.sfu.ca/rst. L'exemple choisi est appelé « the Two Frameworks Text ».

θ **(1) Functions of Language in Two Frameworks (Abstract)**

§ **(2)** Some of the most central problems in linguistics concern how language fills its characteristic roles : how it is useful, the nature and extent of its translatability, and the nature of the integrity of texts. **(3)** Within linguistics there are many kinds of description that bear on such questions, **(4)** one kind being the description of language in terms of its functions. **(5)** Comparing these functional descriptions, **(6)** the various descriptions do not all cover the same ground. **(7)** Rather, each is quite partial, **(8)** and appropriate ways to combine them into a more comprehensive account are not evident. **(9)** It is hard to know wherein they conflict, wherein they agree, and where they simply speak of different things.

§ **(10)** This paper is part of an effort to relate various accounts. **(11)** It is the first in a pair of papers that compare two particular accounts : Rhetorical Structure Theory and Systemic Linguistics.

§ **(12)** Rhetorical Structure Theory, initially formulated in 1983, describes texts in terms of functionally-defined relations that hold between their parts. **(13)** Systemic Linguistics is a much more comprehensive view of language initiated in the early 1960s. **(14)** Where the two approaches are comparable, **(15)** Systemic Linguistics describes texts in terms of categories of processes which the texts perform.

§ **(16)** The paper focus on correlating the relations used in rhetoric al structure theory with the categories of function found in systemic linguistics. **(17)** The correlation employs descriptions of speakers' intentions in an essential way. **(18)** A surprisingly strong correlation results.

FIG. 8.12 – Extrait communément appelé « the Two Frameworks Text »

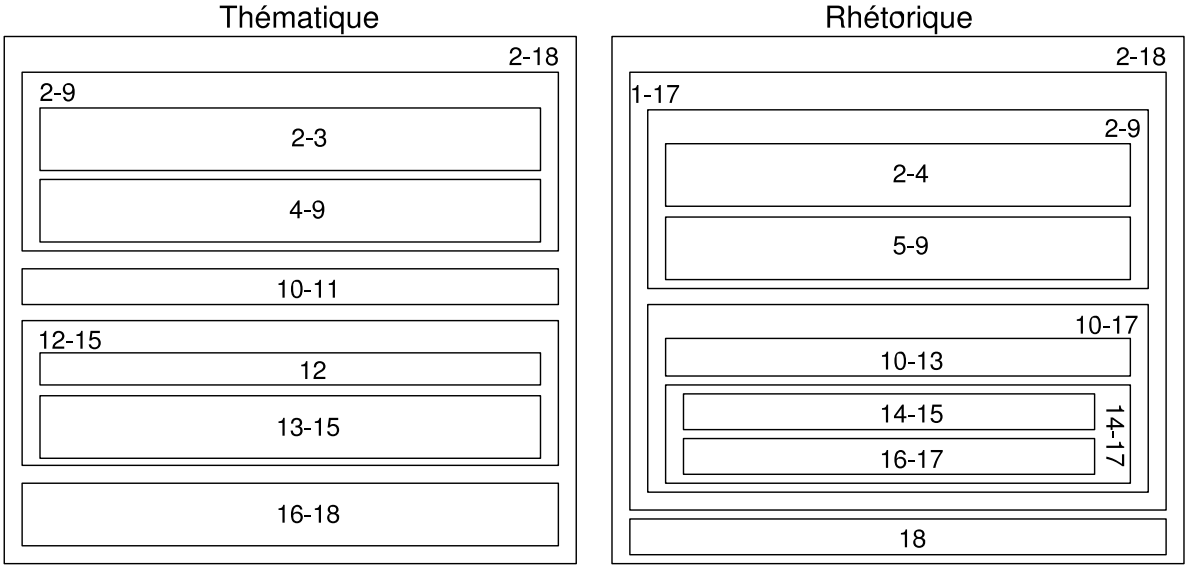


FIG. 8.13 – Segmentations thématique et rhétorique du « Two Frameworks text »

En comparant les deux segmentations, on constate effectivement qu'elles ne sont pas superposables. Une première différence réside dans le traitement du segment 18 : alors que l'analyse rhétorique la place au même niveau que le segment 2-17 en posant une relation de type « interprétation »[14], rien ne nous autorise à faire de même du point de vue thématique. Il semble en effet délicat de réfuter son appartenance au segment 16-18, sa continuité thématique avec 16-17 étant évidente. Cette divergence nous paraît ici facilement explicable en raison des fonctions très différentes que l'on peut ici attribuer à 18 selon que l'on se place sur le plan thématique ou rhétorique : la phrase est peu significative dans le premier cas, alors que sa portée sur le plan argumentatif est particulièrement importante.

Une autre divergence apparaît au niveau du segment 2-9, dont la subdivision n'est pas la même dans les deux cas. Cette différence d'appréciation est relativement peu significative dans la mesure où elle ne concerne que le rattachement de 4, soit à 2-3 soit à 5-9, au sujet duquel il peut sembler difficile de trancher. Mais cette ambiguïté nous semble là encore corroborer la coexistence des deux facettes. Car si 4 constitue manifestement une transition entre 2-3 et 5-9, son rattachement effectif dépendra du point de vue adopté : il s'agit bien d'une élaboration de 3 sur le plan rhétorique, mais aussi du premier élément d'un bloc thématiquement homogène qui s'étend jusqu'à 9.

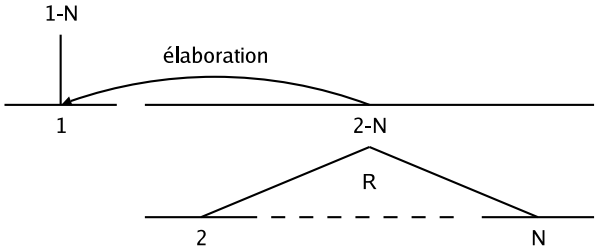
D'autres incompatibilités de segmentation apparaissent dans la seconde moitié du texte (portion 10-17). Notamment, l'analyse rhétorique fait apparaitre un segment 10-13 en relation d'arrière plan avec 14-17. Là encore, cette segmentation ne paraît pas appropriée sur le plan thématique : si l'on raisonne en termes d'à propos, le segment 10-11 ne peut pas être considéré comme homogène avec celui qui suit, car le premier décrit le contenu de l'article, alors que le second traite de modèles linguistique en toute généralité (la même remarque s'applique pour ce qui est de 14-15 et 16-17, où se produit le basculement opposé). Inversement, la segmentation obtenue sur le plan thématique ne représente pas de façon adéquate le cheminement argumentatif de ce passage. Là encore, chacune des analyses paraît donc pleinement justifiée selon que l'on adopte l'un ou l'autre des deux points de vue.

Il est évident que, chaque analyse résultant d'un processus interprétatif subjectif, leur confrontation ne constitue pas une preuve de l'existence d'un discours dont les structures thématique et rhétorique seraient objectivement distinctes. Les différences constatées ici pourraient par exemple exister du simple fait d'une divergence d'appréciation entre les observateurs. Toutefois, en l'absence d'un modèle plus général capable de rendre compte des conflits ici constatés, et compte tenu des remarques formulées dans les section précédentes, nous ne pouvons nous autoriser à établir, en toute généralité, une bijection entre des structures rhétoriques de type RST avec des structures thématiques en général, ni *a fortiori* avec des structures en thèmes composites. Il nous semble donc raisonnable de considérer, si l'on s'en tient aux définitions que nous en avons données plus haut, que les dimensions thématique et rhétorique constituent bien deux facettes distinctes de l'organisation du discours.

8.5.3 Question de la détermination de la structure thématique par la structure rhétorique

Nous avons vu dans les précédentes sections que des structures thématiques équivalentes peuvent se matérialiser par des structures rhétoriques sensiblement différentes, dont certaines ne leur sont pas isomorphes. Toutefois, rien ne nous interdit *a priori* de penser que certaines structures rhétoriques particulières pourraient se traduire systématiquement par une structure thématique en TC. En particulier, en observant les exemples donnés dans la section 8.5.1, on pourrait supposer que cela est le cas pour toute structure du type suivant, où R est une relation multi-nucléaire quelconque (ou une combinaison de telles relations), et N un nombre de segments quelconque :

[14]Précisons que cette analyse semble difficile à remettre en cause, dans la mesure où : (i) dans ce cas précis les observateurs sont également les auteurs du texte ; (ii) le modèle RST est par essence fortement lié aux intentions du locuteur.

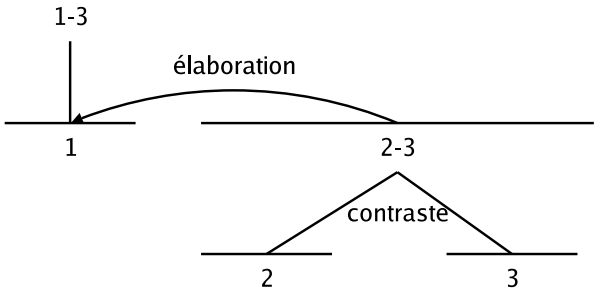


Il semble effectivement qu'il s'agisse d'une situation fréquente. Toutefois, même dans ce cas simple, il est difficile de statuer sur ce point en toute généralité, et il semble hasardeux de chercher à déduire la structure thématique d'un segment (du moins au sens où nous l'entendons) à partir de sa structure rhétorique. En effet, nous allons voir à travers deux exemples que des structures rhétoriques qui sont des instances du schéma général présentée ci-dessus peuvent décrire des textes dont les structures thématiques ne se prêtent que très difficilement à une représentation en thèmes composites. Voici un premier exemple :

§ **(1)** À cause de son aspect porteur, le mot est parfois galvaudé par les services de marketing des entreprises. **(2)** Ainsi, ClearType est présentée comme une technologie alors que ce n'est qu'une technique [. . .]. **(3)** En revanche le Wi-Fi est bien aujourd'hui pour sa part une technologie.

Source : WP1

L'analyse de la structure rhétorique de ce paragraphe pourrait être la suivante. On constatera qu'il s'agit bien d'une instance du schéma donné plus haut :

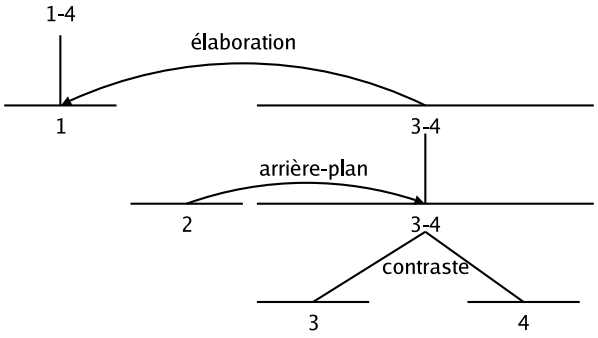


Pourtant, il n'est pas évident de voir dans ce texte un thème composite, puisque cela reviendrait à considérer que le concept de « technologie » est envisagé relativement à un axe sémantique qui contiendrait « ClearType » et « WiFi ». Ce point pose en effet problème dans la mesure où le texte vise précisément à démontrer que ces dernières n'appartiennent pas à une même classe, et que seul le « WiFi » doit être considéré comme une instance du concept de « technologie ». Il est certes possible de considérer que ces deux notions forment bien un axe sémantique, installé temporairement pour mieux être invalidé, mais on se heurte alors à la difficulté de voir en cet exemple une application du principe de « positionnement », qui fait partie intégrante de la notion de thème composite. Il s'agirait donc au mieux d'un thème composite très imparfait.

Observons maintenant ce second cas, qui paraît s'éloigner encore plus nettement de la notion de thème composite :

> [...] **(1)** Des mutations peuvent aussi, en l'absence d'agoniste, déplacer l'équilibre vers l'état R* et déclencher spontanément une réponse physiologique du récepteur. **(2)** Cette activité du récepteur est appelée activité constitutive. **(3)** On connaît aujourd'hui des pathologies liées à ce type de mutations. **(4)** En revanche, on n'avait jamais observé jusqu'alors de cas d'activité constitutive réversible des RCPG, dans des conditions physiologiques.
>
> Source : REC

Notre analyse rhétorique de ce paragraphe est donnée ci-dessous. On s'apercevra qu'il s'agit bien là encore d'une instance de notre schéma, si l'on veut bien faire abstraction de la phrase 2 et de sa relation d'arrière-plan avec 3-4, qui n'interfèrent en rien avec notre propos :



Pourtant, la structure thématique de ce passage se prête encore plus difficilement que la précédente à une description en termes de TC. En effet, même si la séquence 3-4 constitue bien une élaboration relativement à un concept C introduit en 1 (un certain type d'activité d'un récepteur dit RCPG), il semble difficile de réunir au sein d'un même axe sémantique des référents introduits dans 3 et 4 : la première fait référence à des pathologies liées à C, alors que la seconde évoque l'observation de C dans des circonstances particulières.

L'analyse du passage sur le plan sémantique laisse donc à penser qu'une structure de type TC n'est pas adaptée à la description de son contenu informationnel, ce qui témoigne en faveur de la non détermination de la structure thématique par la structure rhétorique, puisque deux textes représentés par des schémas RST analogues ne se prêtent pas à la même analyse thématique.

8.5.4 Intérêt de l'analyse rhétorique pour améliorer l'analyse thématique

Nous souhaitons terminer cette section en mentionnant rapidement que le fait de considérer les structures rhétorique et thématique comme bien distinctes n'interdit en rien d'envisager des points de contact entre elles, que l'on pourrait exploiter dans le cadre de l'analyse automatique de l'une ou de l'autre.

Dans le cas de l'analyse thématique, on pourrait par exemple bénéficier de l'analyse de segments conclusifs pour améliorer la segmentation par thèmes composites. Car s'il est envisageable, comme nous l'avons vu dans le chapitre 7, de procéder automatiquement à l'analyse de la portée de certains introducteurs particuliers (en l'occurrence spatio-temporels), ce problème est évidemment très loin d'être résolu en toute généralité. On voit alors que la détection des segments conclusifs pourraient aider à reconnaître la fin du dernier segment d'une configuration en TC, qui se traduit par la « remontée » à un niveau supérieur comme on peut le voir sur le schéma de la figure 8.14. Il ne s'agit bien sûr que d'une possibilité parmi d'autres, et il est probable que divers points de jonctions de cet ordre entre les deux modes d'organisation puissent être exploités en analyse automatique.

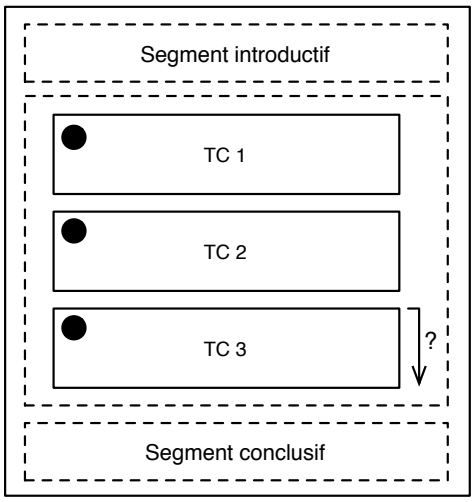


FIG. 8.14 – Exemple d'interaction possible entre les analyses rhétorique et thématique. Les puces noires représentent les expressions satellites caractérisant les différents sous-segments d'une configuration en TC. On voit comment la détection d'une marque conclusive peut permettre de clôturer le dernier de ces segments.

Troisième partie

La plate-forme LinguaStream

Nous allons présenter dans cette partie la plate-forme LinguaStream, dont nous avons pu voir plusieurs exemples d'utilisation dans la partie précédente. Elle a été conçue parallèlement à ces travaux pour faciliter leur élaboration, et est devenue au fil du temps une plate-forme générique pour le traitement automatique des langues. Elle a pour ambition de simplifier la réalisation d'expériences non triviales sur corpus, ainsi que le cycle d'évaluation / ajustements qui en découle. Sans outil adapté, le coût de développement induit par chaque nouvelle expérience devient en effet un frein considérable à l'approche expérimentale, et pour répondre à cette problématique, LinguaStream facilite la mise en oeuvre de procédés complexes tout en impliquant un investissement technique minimal. Elle est aujourd'hui utilisée à des fins de recherche ou d'enseignement dans différents laboratoires, et son développement est aujourd'hui assuré au sein d'une équipe à laquelle prennent notamment part Antoine Widlöcher et des membres de l'ERSS.

Nous allons dans un premier temps en donner une **présentation générale**, où nous présenterons la plate-forme comme un environnement d'expérimentation intégré fondée sur la notion de chaîne de traitement, assemblage de modules d'analyse de types et de niveaux variés. Nous détaillerons ensuite ses **fondements méthodologiques**, à l'aide des notions de *perspective d'analyse* ou encore de *complémentarité des modèles d'analyse*. Nous décrirons par la suite le **modèle documentaire** sur lequel elle s'appuie pour représenter les annotations et procéder à leur marquage effectif dans les documents. Le modèle sera d'abord décrit dans sa forme abstraite, avant d'exposer brièvement les modalités concrètes du balisage XML sous-jacent et une API élaborée pour faciliter l'exploitation de documents conformes à notre modèle. Le chapitre suivant concernera les différents **modèles d'analyse** qui ont été intégrés à la plate-forme et qui permettent, en utilisant différents langages déclaratifs, de formaliser la connaissance linguistique que l'on souhaite projeter sur des corpus. Nous donnerons enfin un aperçu concret de l'environnement intégré proposé par la plate-forme, avant de mentionner en conclusion un certain nombre de cas d'utilisation ainsi que quelques détails techniques relatifs à sa réalisation logicielle.

Chapitre 9

Présentation générale

LinguaStream est une plate-forme générique pour le traitement automatique des langues naturelles, fondée sur le principe de l'enrichissement incrémental des documents électroniques. Elle permet la conception et l'évaluation de chaînes de traitement complexes, par assemblage de modules d'analyse de types et de niveaux variés : morphologique, syntaxique, sémantique, discursif ou encore statistique. Ainsi, chaque palier de la chaîne de traitement se traduit par la découverte et le marquage de nouvelles informations, sur lesquelles pourront s'appuyer les analyseurs subséquents. En fin de chaîne, différents outils permettent de visualiser les documents analysés et leurs annotations.

La plate-forme propose différents mécanismes d'élaboration des composants de traitement : règles morphologiques, grammaires d'unification, expressions rationnelles, lexiques sémantiques, grammaires de contraintes, moteurs d'inférences, etc. La plupart d'entre eux s'appuient sur des formalismes déclaratifs, certains étant couramment utilisés en TAL. Chaque composant d'analyse est réutilisable immédiatement dans d'autres chaînes de traitement, et peut être remplacé par un autre composant fonctionnellement équivalent. Une interface graphique prend en charge les différents aspects de l'élaboration d'une chaîne de traitement complète.

LinguaStream a pour ambition de faciliter la réalisation d'expériences non triviales sur corpus, ainsi que le cycle d'évaluation/ajustements qui en découle. Sans outil adapté, le coût de développement induit par chaque nouvelle expérience devient en effet un frein considérable à l'approche expérimentale. Pour répondre à cette problématique, LinguaStream facilite la mise en oeuvre de procédés complexes tout en requérant des compétences informatiques minimales.

La notion de chaîne de traitement constitue le fondement de la plate-forme, et vise à répondre au besoin d'articulation de traitements qui émerge massivement en TAL. Ces chaînes de traitements peuvent être éditées visuellement, les divers composants apparaissant sous forme de « boîtes » reliées entre elles. Une représentation uniforme des annotations permet à chaque composant d'exploiter les résultats produits par les analyses précédentes. La plate-forme dispose en standard d'un jeu de composants extensible, qu'il suffit de sélectionner dans une « palette ». Chaque composant dispose d'un ou plusieurs points d'entrée et/ou de sortie que l'utilisateur peut relier à sa guise pour obtenir la chaîne voulue. Les composants sont généralement caractérisés par un certain nombre de propriétés qui permettent d'adapter leur comportement à chaque cas particulier. On notera que, le modèle sous-jacent à une chaîne de traitement étant un graphe orienté acyclique, les chaînes ne sont pas limitées à de simples « pipelines ». Par exemple, les différents points de sortie d'un composant peuvent être connectés à des sous-chaînes différentes, dont les résultats pourront à leur tour être fusionnés par un composant acceptant plusieurs points d'entrée. À l'exécution, la plate-forme se chargera de l'ordonnancement des sous-tâches.

LinguaStream incorpore nativement plus d'une cinquantaine de composants différents (ensemble facilement extensible grâce une API Java, un système de macro-composants, et des templates). Cer-

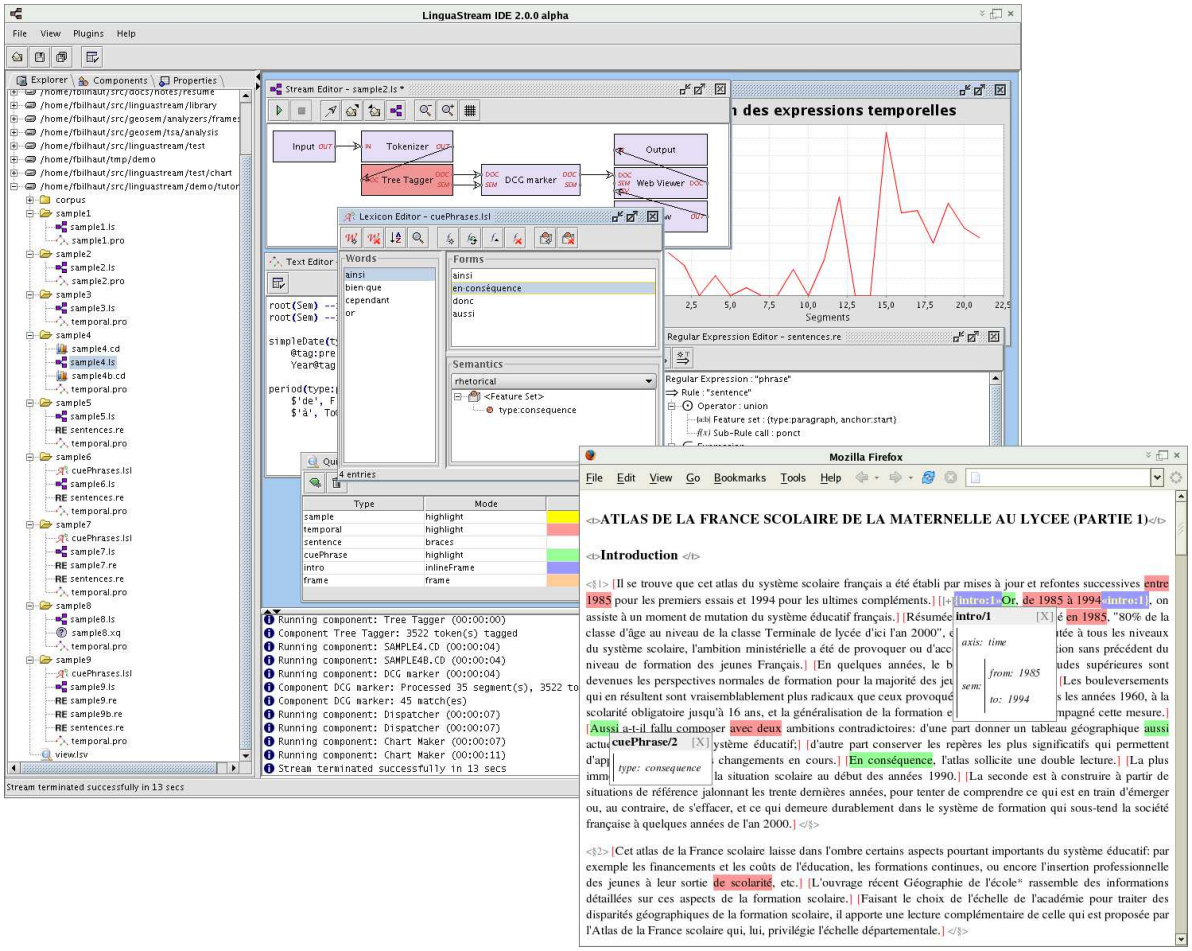


FIG. 9.1 – Environnement intégré de la plate-forme LinguaStream

tains sont spécifiquement dédiés au traitement automatique des langues, et d'autres permettent de résoudre différents problèmes liés à la gestion des documents électroniques (traitements XML en particulier). D'autres composants peuvent être utilisés pour effectuer des calculs sur les annotations produites par les analyseurs, ou générer des diagrammes.

La plate-forme encourage le recours à des représentations déclaratives pour spécifier les traitements et leur enchaînement. Une chaîne incorporera des outils exécutant des tâches génériques (telles que le découpage et l'étiquetage de mots), mais aussi et surtout des composants réalisant des analyses intégralement spécifiées par l'utilisateur. Les chaînes de traitement elles-mêmes sont représentées visuellement, l'appareil procédural qui en découle étant masqué.

Elle est par ailleurs conçue pour exploiter la complémentarité des modèles d'analyse, plutôt que privilégier un hypothétique modèle « omnipotent » capable d'exprimer efficacement tout type de contrainte. Nous faisons l'hypothèse qu'un analyseur complexe doit adopter successivement plusieurs regards sur le même matériau linguistique, auxquels répondront des formalismes distincts. On pourra combiner, au sein d'un même traitement, des expressions régulières au niveau morphologique, une grammaire locale au niveau syntagmatique, un transducteur au niveau phrastique et une grammaire de texte au niveau discursif.

Tous les modèles d'analyse proposés par la plate-forme s'appuient sur une représentation unifiée des marquages et des annotations. Ces dernières sont uniformément représentées sous forme de structure de traits, modèle communément utilisé en TAL et en linguistique, et permettant de représenter

des annotations riches et structurées. Tout composant d'analyse pourra produire son propre marquage en s'appuyant sur les analyses précédentes : les formalismes proposés permettent de spécifier des contraintes sur les annotations existantes par unification. Les marquages résultants sont organisés en couches indépendantes et hiérarchisées, supportant enchâssements, superpositions et chevauchements.

La plate-forme favorise l'abstraction progressive des formes de surface. En effet, chaque palier d'analyse pouvant accéder simultanément aux annotations produites par tous les paliers antérieurs, les analyseurs de plus haut niveau sont généralement conduits à s'abstraire progressivement du matériau textuel pour ne plus reposer que sur des représentations symboliques antérieurement calculées.

Elle autorise également la variabilité du grain d'analyse au cours du traitement. De nombreux modèles d'analyse imposent la définition d'un grain d'analyse minimal, dit « jeton » ou token. C'est par exemple le cas de toute grammaire ou transducteur : ces formalismes supposent l'existence d'une unité textuelle (comme le caractère ou le mot) à laquelle s'appliquent les patrons. Quand la définition de ce grain minimal est nécessaire au fonctionnement d'un composant, la plate-forme permet de spécifier le ou les types d'unité à considérer comme jetons. Toute unité préalablement délimitée peut jouer ce rôle : il pourra s'agir du découpage habituel en mots, mais aussi de toute autre unité pourvu qu'elle ait été préalablement marquée : syntagmes, phrases, cadres du discours, etc. Le grain minimal à prendre en compte peut donc être différent pour chaque palier de l'analyse. Ce système augmente considérablement la portée des différents modèles d'analyses utilisables dans la plate-forme.

Filtrage des analyses : LinguaStream prévoit que chaque module d'analyse spécifie les marquages antérieurs auxquels il souhaite faire référence. Chaque analyseur peut ne tenir compte que des marquages qu'il estime pertinents et les marquages devenus inutiles peuvent être définitivement supprimés en cours de traitement pour des raisons de performance ou de commodité. Combinée au libre choix du grain et du modèle d'analyse, cette fonction permet d'adopter un point de vue spécifique sur le document à chaque étape d'une chaîne de traitement.

Le recours systématique à l'élaboration de chaînes de traitement, constituées de divers composants élémentaires et de sous-chaînes, conduit naturellement à mener une analyse descendante de la tâche à réaliser, qui se trouve scindée en un ensemble de sous-tâches qui peuvent être traitées indépendamment. Ce processus de séparation des problèmes conduit à une meilleure compréhension de l'ensemble du système. En outre, la qualité descriptive de la chaîne elle-même contribue à la pérennité du procédé développé.

La modularité des chaînes de traitements favorise la réutilisabilité des composants dans des contextes différents : un module d'analyse développé au sein d'une première chaîne de traitement pourra par la suite être réutilisé dans d'autres chaînes. De façon similaire, toute chaîne de traitement pourra être réutilisée en tant que constituant d'une chaîne de plus haut niveau, sous forme de « macro-composant ». La possibilité de réutiliser des modules existants réduit significativement le coût de développement induit par la conduite de nouvelles expériences.

Pour une chaîne donnée, on pourra substituer à un composant tout autre composant fonctionnellement équivalent. Cela permet d'utiliser, pour une sous-tâche donnée, un prototype rudimentaire pouvant être remplacé in fine par un équivalent pleinement opérationnel. On pourra également évaluer les variations du résultat selon que l'on utilise tel ou tel composant (par exemple deux étiqueteurs morpho-syntaxiques différents).

LinguaStream se base de façon systématique sur les standards et outils XML disponibles. Les annotations insérées par la plate-forme appartiennent à un espace de nom spécifique, et sont donc transparentes pour les autres outils. Réciproquement, le balisage initial du document est transparent pour les chaînes de traitement, qui pourront être appliquées sans distinction à des documents de formats variés. Si la structure logique du document (par exemple le découpage en paragraphes) doit être prise en compte, des composants spécialisés en déduisent les annotations nécessaires. Le respect des recommandations XML permet en outre de bénéficier de la panoplie grandissante d'outils associés, no-

tamment pour la visualisation et l'annotation manuelle. Une collection de composants est fournie avec la plate-forme pour tirer parti des vocabulaires XML les plus courants, comme XHTML ou DocBook.

LinguaStream incorpore nativement plus d'une cinquantaine de composants différents (ensemble facilement extensible grâce une API Java, un système de macro-composants, et des templates). Certains sont spécifiquement dédiés au traitement automatique des langues, et d'autres permettent de résoudre différents problèmes liés à la gestion des documents électroniques (traitements XML en particulier). D'autres composants peuvent être utilisés pour effectuer des calculs sur les annotations produites par les analyseurs, ou générer des diagrammes.

Chaque formalisme permet d'exprimer des contraintes tant sur les formes de surface que sur les annotations insérées par les analyseurs précédents. Toutes les annotations sont représentées sous forme de structures de traits, et les contraintes sont systématiquement spécifiées par unification sur ces structures. Si la syntaxe varie d'un formalisme à l'autre, ce mécanisme est uniforme pour tous les modèles d'analyse proposés. En revanche, chaque modèle d'analyse s'appuie sur un mode particulier de représentation des contraintes. Parmi les différents formalismes disponibles, on trouvera notamment :

- un système noté EDCG (pour « Extended Definite Clause Grammar »), permettant de décrire des grammaires locales d'unification ;
- un système nommé MRE (pour « Macro-Regular Expression »), permet de décrire des patrons sous la forme d'expressions rationnelles ;
- un système d'annotation à partir de lexiques sémantiques LSL (pour « LinguaStream Semantic Lexicon ») ;
- le formalisme CDML (pour « Constraint-based Discourse Modelling Language »), qui permet d'exprimer des contraintes au niveau discursif (conçu et implémenté par Antoine Widlöcher).

S'il est désormais nécessaire d'envisager la mise en place des traitements sur corpus de manière cyclique, en alternant les étapes d'évaluation/validation et les ajustements des modèles et traitements, la nature des procédures d'observation à mettre en oeuvre n'est pas sans poser problème. Nous proposons les principes et méthodes suivants.

Un premier aspect essentiel consiste tout simplement dans la visualisation du résultat des traitements mis en place, c'est-à-dire dans les outils d'observation du corpus marqué et annoté. Sur ce premier point, nous proposons de multiplier les vues possibles, en nous appuyant essentiellement sur différents standards XML (XHTML, XSL-FO, SVG...). Le mode le plus courant consiste à présenter un corpus surligné, chaque marquage donnant accès à la représentation symbolique, la structure de traits, associée à l'analyse dont il résulte. L'utilisation d'espaces de noms spécifiques pour le marquage, ainsi que l'indifférence de la plate-forme à l'égard des formats de documents fournis en entrée (pourvu que ceux-ci respectent le standard XML) permet la conservation des spécificités structurelles et visuelles des documents : un corpus DocBook conservera, après analyse, sa validité DocBook et pourra donc, sans perte, être rendu par les moyens habituels. Une vue en concordancier sera prochainement disponible et permettra également d'accéder simultanément aux segments marqués et à leurs annotations.

Un autre élément fondamental pour l'évaluation découle directement du principe de modularité des chaînes de traitement. La possibilité de remplacer à tout instant l'un quelconque des composants par un élément fonctionnellement équivalent rend possible la mise en comparaison des traitements, en soumettant ces derniers à des contextes rigoureusement identiques, condition *sine qua non* d'une comparaison pertinente.

Enfin, la plate-forme propose des outils permettant de produire des diagrammes issus de calculs numériques et statistiques sur les différents marquages présents dans un document, et fournit des composants permettant d'obtenir des mesures traditionnelles telles que précision/rappel, cette dernière autorisant en particulier la comparaison d'annotations automatiques et manuelles.

Il est difficile à ce stade de présenter en détail ce qui différencie LinguaStream des systèmes comparables, qu'ils aspirent au même degré de généricité, comme la plate-forme GATE (Cunningham *et al.*, 2002), ou qu'ils soient adossés à un modèle théorique plus spécifique, comme c'est le cas pour ContextO (Déclès et Minel, 2000) ou Intex (Muller *et al.*, 2004). Nous ferons plutôt mention de ses caractéristiques spécifiques au fur et à mesure des chapitres qui suivent.

Chapitre 10

Principes méthodologiques

10.1 Approche par composants

L'un des fondements méthodologiques de la plate-forme est la systématisation de l'approche par composants. Cette approche, issue du génie logiciel, peut être vue comme un prolongement de la conception orientée objet. Sans rentrer dans la définition exhaustive de ses principes, nous nous contenterons ici de mentionner ceux qui nous paraissent les plus essentiels.

Elle promeut tout d'abord la *fragmentation récursive* de problèmes complexes en sous-problèmes plus simples et plus faciles à appréhender. Il s'agit bien sûr d'un principe commun à beaucoup de méthodes d'analyse, mais l'approche par composants demande expressément que le résultat de cette fragmentation se matérialise par des unités concrètes autonomes et indépendantes. En pratique, chaque sous-tâche sera devra être réalisée par un module distinct, indépendant du système global auquel il prendra part *in fine* et donc facilement réutilisable dans d'autres contextes.

Elle s'appuie par ailleurs sur la notion de *service* : suite à l'identification de chaque sous-problème, celui-ci se traduit avant-tout par une *interface* définissant les modalités d'interaction entre le module correspondant et le reste du système. On définit ainsi formellement les sollicitations auxquelles ce module est capable de répondre, la nature des données qui doivent lui être fournies, le résultat attendu, etc. De ce fait, les modalités pratiques permettant de réaliser ce service sont totalement masquées, et tous les composants rendant un même service sont considérés, du point de vue du système global, comme équivalents (ce qui ne signifient pas nécessairement qu'ils rendent la même « qualité » de service).

Un corollaire, plus technique, de l'approche par composant est l'indépendance *physique* des différents modules prenant part à un système. On parle alors de système distribué, dont les différents modules peuvent appartenir à des processus logiciels différents et même s'exécuter sur des machines distantes. Les environnements logiciels aujourd'hui disponibles dans ce domaine sont aujourd'hui légion[1], de plus en plus intégrés aux les langages de programmation[2], et de mieux en mieux standardisés[3].

Pour résumer, on pourrait dire que l'approche par composants confère à une architecture les qualités suivantes :

Ré-utilisabilité : chaque composant se limite à un sous-problème particulier, et pourra prendre part à l'assemblage de différents systèmes complexes.

[1] On citera notamment des systèmes comme Corba, DCOM, RMI, ROAP ou encore REST.

[2] Le développement d'un composant distribué sous Java ou .NET est aujourd'hui à peine plus complexe que de créer une simple classe...

[3] Notamment avec les protocoles de type « service Web ».

Transparence : un composant ne se définit que par le service qu'il rend, et peut être remplacé dans système par tout autre composant rendant ce même service.

Distribution : les différents composants d'un système peuvent être physiquement éloignés.

L'application de ces principes au traitement automatique des langues présente de nombreux avantages. En premier lieu, la décomposition d'un problème complexe en sous-problèmes plus simples est, dans ce domaine comme ailleurs, souvent salutaire. Tout d'abord parce que ce processus de fragmentation est en soi une méthode d'analyse du problème global. Et ensuite parce qu'elle permet de se concentrer tour à tour sur chaque sous-problème, sans nécessairement avoir une vue sur la globalité du système.

D'autre part, le fait de chercher à construire des composants indépendants permet de sélectionner pour chacun d'entre eux les outils conceptuels et techniques les plus adaptés à la tâche considérée. Comme nous le verrons plus loin, la plate-forme permet notamment d'utiliser un formalisme différent pour écrire les règles de chaque composant (même si ce n'est bien sûr pas obligatoire). On pourra ainsi faire explicitement apparaître dans une chaîne de traitement les différents niveaux linguistiques traités. Par exemple, bien que la plate-forme elle-même ne fixe aucune limite entre ces niveaux, on pourra utiliser des outils différents pour les niveaux morphologique, syntagmatique, propositionnel, phrastique ou encore discursif, et les voir apparaître sous forme de composants (ou même de sous-chaînes) distincts dans une même chaîne de traitement.

Ainsi, au sein d'une chaîne de traitement, chaque composant procède à sa propre annotation en se basant sur les annotations réalisées par les composants précédents, les annotations étant toujours représentées sous la forme de structures de traits (cf. chapitre 11). Ainsi, une fois qu'une unité d'un certain type a été marquée dans le document, les composants suivants pourront s'appuyer sur la présence de cette unité ou sur les annotations associées pour générer à leur tour un nouveau marquage. De cette façon, la notion de chaîne de traitement aboutit généralement à une abstraction progressive des formes de surface.

L'exemple des cadres de discours que nous avons décrit dans le chapitre 7 est un bon exemple d'application de ce principe de fragmentation par couches linguistiques. Comme nous l'avons décrit alors, et comme le montre la figure 10.1, on pourra observer sur la chaîne de traitement élaborée à cette fin une premier ensemble de composants dédié aux traitements d'ordre morphologique et lexical (comme le découpage en mots ou l'étiquetage morpho-syntaxique), un autre lié aux traitements d'ordre syntagmatique (comme le repérage des expressions temporelles ou des introducteurs), un troisième lié au niveau discursif proprement dit (délimitation des cadres), etc.[4] On voit également comment, à chaque étape, une nouvelle analyse peut exploiter les annotations antérieurement calculées, par exemple quand l'analyse des expressions temporelles se base sur le découpage en mot et leur étiquetage, quand la détection des introducteurs s'appuie sur celle des expressions temporelles et des connecteurs de discours, ou encore quand l'analyse de la portée de ces introducteurs exploite la quasi-totalité des marquages antérieurs.

L'indépendance des composants permet également de bénéficier directement des avantages plus généraux que nous avons décrits plus haut. Cela inclut tout d'abord la réutilisation d'un composant développé pour une tâche donnée dans un autres contexte et/ou par une tierce personne. Cela garantit également la possibilité de substituer à un composant un autre composant rendant le même service. Par exemple, on pourra tester dans une même chaîne différents étiqueteurs morpho-syntaxiques afin d'observer l'incidence des différences entre leurs annotations sur le reste de cette chaîne. Enfin, l'indépendance physique des composants permet d'exploiter à distance des composants tiers mis à disposition sur d'autres machines. À l'extrême, une chaîne de traitement pourrait ne faire appel qu'à des

[4] On remarquera à ce sujet que l'ordre naturel des niveaux linguistiques peut ne pas être strictement respecté, par exemple lorsqu'on exploite les limites des phrases pour détecter les introducteurs

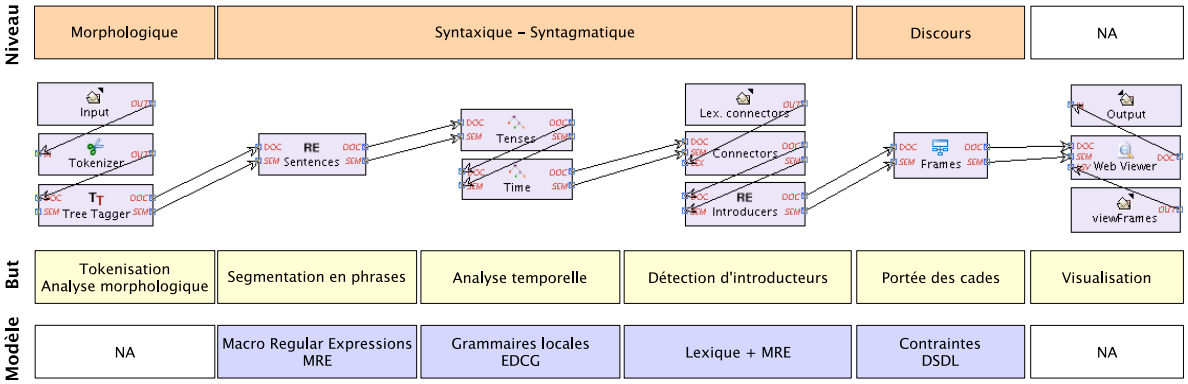


FIG. 10.1 – Approche modulaire du TAL : l'exemple des cadres de discours.

composants distants (via un réseau local ou Internet) pour les réunir au sein d'un même processus[5].

L'idée d'appliquer une approche par composants au traitement des langues n'est pas nouvelle en elle-même : elle est notamment décrite dans (Cunningham, 2000), et concrétisée par la plate-forme GATE (Cunningham *et al.*, 2002). Nous prétendons toutefois aller plus loin dans ce sens, en appliquant ces principes de façon systématique à tous les niveaux de notre plate-forme. En effet, au delà des fonctionnalités offertes par différents systèmes déjà proposés, LinguaStream offre les possibilités suivantes :

- Elle permet d'assembler des dispositifs expérimentaux sous la forme de chaînes de traitement arbitrairement complexes (graphes acycliques quelconques). Ces chaînes sont elles-mêmes décrites déclarativement, et peuvent être construites visuellement dans un environnement graphique.
- Ces chaînes de traitements permettent de faire collaborer des composants de natures très différentes. En s'appuyant sur l'hypothèse de la complémentarité des modèles d'analyse, la plate-forme permet notamment d'assembler des modules construits à l'aide de formalismes foncièrement différents. Comme nous le verrons plus loin, cela permet de choisir le formalisme le plus adapté au niveau linguistique traité par chaque composant. D'autre part, des composants spécifiquement liés au traitement des langues peuvent collaborer avec des composants de natures totalement différentes, tels que des outils de transformation XML, des moteurs de requêtes RDF ou même des systèmes de fouille de données. La plate-forme permet également de constituer des composants à l'aide de composants logiciels tiers, pour peu que l'utilisateur spécifie les modalités de transformations de données nécessaires.
- Les modules crées sous la forme de règles déclaratives, de scripts ou même de composants *ad-hoc* sont très facilement isolables et partageables. Dans la plupart des cas, la cession d'un simple fichier suffit à transférer un module à une tierce personne. Une chaîne de traitement complète peut également être « exportée » sous la forme d'un paquet autonome et facilement réutilisable. Dans le cas de composants logiciels spécifiques, un système de *plugins* permet de les mettre à disposition de façon indépendante et de les charger dynamiquement dans la plate-forme.

La notion de chaîne de traitement concrétisant l'approche par composants constitue donc le coeur de la plate-forme. C'est tout d'abord le cas au niveau de son architecture interne, principalement articulée autour du moteur qui se charge de représenter et d'exécuter ces chaînes. Ce dernier est basé sur un modèle relativement abstrait de la notion de composant, et est donc en lui-même indépendant de toute application au traitement des langues. Il est donc capable d'animer une chaîne de traitement

[5]On retrouve alors l'idée de construction de processus par assemblage de services Web qui fonde les formalismes tels que BPML (Business Process Modeling Language).

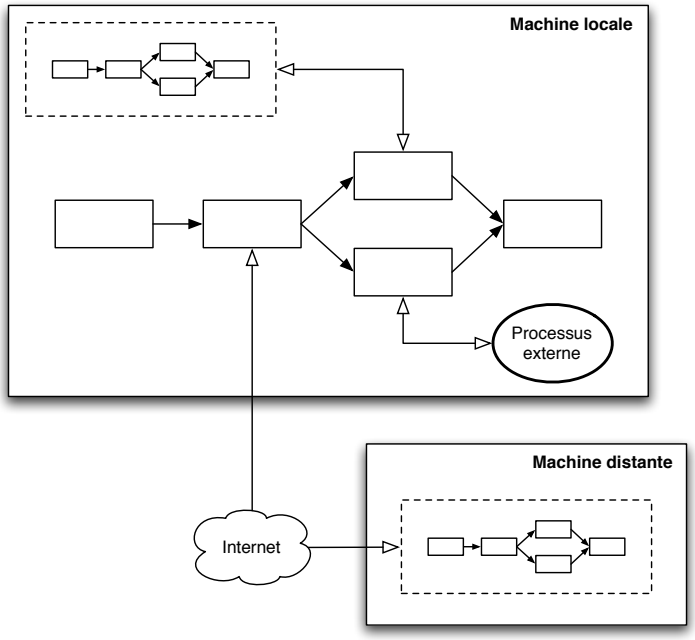


FIG. 10.2 – Modularité des chaînes de traitements. Une chaîne peut faire appel, outre les composants de la plate-forme elle-même, à des processus externes ainsi qu'à des sous-chaînes locales ou distantes.

comprenant des composants quelconques, pourvu que ces derniers répondent à une interface bien définie. Bien-sûr, la majorité des composants que nous avons développés sont bien dédiés au traitement des langues, et s'appuient pour cela sur le modèle documentaire et les API correspondantes qui seront décrits au chapitre 11. D'autres feront appel à des processus externes à la plate-forme elle-même, par exemple pour faire appel à l'étiqueteur TreeTagger (Schmidt, 1994) ou à l'analyseur syntaxique Syntex (Bourigault et Fabre, 2000). D'autres encore seront plus généraux, par exemple dédiés aux traitements XML (comme les composants XSLT ou XQuery). Enfin, comme nous le mentionnions plus haut, ce moteur permet de constituer un macro-composant à partir d'une chaîne de traitement locale ou distante, composant qui pourra à son tour prendre par à une nouvelle chaîne de traitement. Finalement, une même chaîne de traitement peut effectivement constituer un processus à partir de constituants très divers, comme le montre la figure 10.2.

Précisons que ce moteur a été conçu pour être largement ouvert et accessible. Tout d'abord, une chaîne de traitement est elle-même représentée en XML, ce qui facilite la création dynamique de ces chaînes, pourquoi pas *via* XSLT au sein d'une autre chaîne de traitement. Une chaîne est également paramétrable via des variables qui peuvent être utilisées pour le règlage de n'importe quel composant : une même chaîne peut donc être rendue relativement adaptable. D'autre part, le moteur lui-même est accessible via une API, ce qui permet d'intégrer très facilement une chaîne de traitement au sein d'une application tierce (par exemple une application Web). Enfin, une chaîne de traitement peut également être exécutée comme n'importe quel outil de ligne de commande (indépendamment de l'environnement graphique) grâce à un *runtime* fourni avec la plate-forme.

Bien-sûr, la notion de chaîne de traitement constitue également un point central de l'environnement d'expérimentation offert par la plate-forme, où les composants peuvent être choisis dans une « palette » et assemblés visuellement. Nous reviendrons plus longuement sur ce point dans le chapitre 13.

10.2 Formalismes déclaratifs et complémentarité des modèles d'analyse

Outre l'approche par composants qui s'attache plutôt aux modalités pratiques d'interaction entre les modules d'une chaîne de traitement, la plate-forme promeut la complémentarité des modèles d'analyse, qui s'appuie d'une part sur l'utilisation de formalismes purement déclaratifs pour produire des modèles opératoires, et d'autre part sur l'hypothèse que les règles dédiées aux différents niveaux linguistiques d'un traitement ne peuvent toutes s'exprimer à l'aide d'un seul et même formalisme.

Précisons tout d'abord ce que nous entendons par « modèle d'analyse ». Nous désignons par ce terme la combinaison des deux facettes indissociables d'un outil permettant de procéder à l'annotation automatique d'un document à partir d'un modèle formel. Il s'agit donc de la combinaison d'un outil conceptuel d'une part et d'un outil technique d'autre part :

- un formalisme permettant d'exprimer déclarativement des règles constituant un modèle opératoire des objets linguistiques à annoter ;
- un algorithme permettant de projeter ces règles sur corpus.

On peut citer comme exemples de formalismes les grammaires d'unification, les expressions régulières, les langages de contraintes, ou encore les règles de type système expert. Comme exemples d'algorithmes on peut citer les transducteurs déterministes ou non, les moteurs d'inférences sur base de faits, les moteurs de résolution de contraintes, etc. Ces deux aspects ne sont bien sûr pas indépendants : à chaque formalisme correspondront des algorithmes plus ou moins efficaces, dont l'un sera effectivement choisi pour implémenter le composant effectivement utilisé. Selon les compétences de chaque utilisateur, la connaissance des propriétés de ces deux facettes pourra entrer en considération dans le choix du modèle d'analyse le plus adapté à une tâche donnée, bien que la seconde concerne essentiellement les performances des analyseurs (en temps de calcul notamment, mais aussi en mémoire dans certains cas) et puisse être ignorée lorsque les performances apparaissent comme secondaires.

Nous nous concentrerons ici sur la facette liée à la combinaison de formalismes directement opératoires, qui nous paraît avoir le plus d'impact d'un point de vue méthodologique, et constitue, avec la notion de perspective d'analyse que nous verrons plus loin, une des hypothèses fortes soutenues par LinguaStream.

10.2.1 Formalismes déclaratifs

Comme nous l'avons déjà évoqué dans le premier chapitre de cette partie, l'un des objectifs que nous visons consiste à réduire la distance qui sépare les modèles descriptifs des modèles opératoires. Cette distinction est généralement pratiquée (voir entretenue) dans le domaine du TAL dans la mesure où la recherche y relève souvent d'une collaboration entre linguistes et informaticiens. Une certaine conception de cette collaboration[6] consiste à considérer que les premiers détiennent la connaissance sur le matériau linguistique quand les seconds sont à même de les transposer avec un minimum de pertes sous la forme d'un logiciel. L'approche que nous souhaitons encourager par le biais de notre plate-forme est totalement inverse.

En effet, si l'on considère qu'une discipline se définit autant par la méthodologie qu'elle emploie que par son objet d'étude, le TAL peut être considéré comme une discipline à part entière, partageant certes le même objet d'étude que la linguistique, mais en employant une méthodologie sensiblement différente. Dans cette perspective, on ne cherchera pas nécessairement de concordance immédiate entre un modèle « purement » linguistique et un modèle « purement » opératoire, même s'ils visent la description d'un même objet, ni forcément à dériver l'un de l'autre. Chacun de ces modèles ne constitue par définition qu'une représentation simplifiée de la réalité observée, et la part de réalité dont chacun rendra compte est de notre point de vue directement dépendante de la méthodologie

[6] Qui tombe heureusement en désuétude, mais semblent encore loin d'avoir disparu.

employée, et donc de la discipline dans laquelle on s'inscrit. En d'autre termes, nous défendons l'idée qu'un modèle de TAL ne se limite en aucun cas à la simplification d'un autre modèle qui serait, lui, linguistique. Il constituerait au contraire un modèle autonome, différent car résultant de de choix de modélisation différents, dictés par la finalité opératoire. Mais cela n'exclut bien entendu en rien les fructueux rapports entre les disciplines, qui restent évidemment indispensables. Il s'agit simplement de garder à l'esprit qu'au sein de la discipline TAL, il est également question de modèles à part entière et non pas seulement transposition d'une connaissance extérieure à des programmes informatiques impénétrables.

Cependant, il reste à expliciter ces modèles propres au TAL. Car il ne suffit bien sûr pas pour cela de produire un logiciel produisant les résultats attendus. Lorsqu'on se contente de transposer directement sous la forme d'un programme un modèle conçu dans une perspective purement descriptive avec toutes les approximations que cela suppose, le modèle opératoire n'est à aucun moment rendu explicite : il est au mieux présent à l'esprit du programmeur au moment de la phase de codage, mais se perd rapidement et souvent définitivement dans les méandres du code source ainsi produit.

L'utilisation systématique de formalismes déclaratifs est une réponse à ce problème, puisqu'ils garantissent en effet une indépendance franche entre la description formelle de l'objet étudié et les outils techniques susceptibles d'appliquer ces descriptions pour l'annotation automatique. En d'autres termes, les règles (au sens large) qui sont ainsi décrites ont valeur de modèle, tout en étant projetées sur le corpus sans nécessiter aucune phase de d'adaptation qui risquerait de le dénaturer. Ou encore, pour reprendre les termes d'Antoine Widlöcher, elles constituent un modèle à la fois *descriptif* et *prescriptif* (Widlöcher, 2006).

10.2.2 Combinaison des modèles d'analyse

Malgré les avantages qu'il présente, l'usage systématique de représentations déclaratives pour réaliser des tâches non triviales en traitement des langues pose d'autres problèmes. En effet, quel que soit le formalisme choisi pour modéliser une réalité linguistique donnée, il va de soi que ce formalisme est nécessairement limité quant à ce qu'il permet d'exprimer. Et ce sont souvent ces limitations qui conduisent *in fine* à l'emploi de langages de programmation « classiques » pour procéder à la projection d'un modèle sur corpus.

Il est bien sûr possible de chercher à élaborer un modèle à la plus large couverture possible, le plus flexible possible, le plus indépendant possible d'une tâche particulière. On peut toutefois douter sérieusement de cette possibilité, ou en tout cas de la concision et de la qualité purement déclarative d'un tel formalisme. Nous pensons au contraire qu'il est préférable de multiplier les formalismes, en partant du principe que différentes modèles ou sous-modèles, généralement liés à différents niveaux linguistiques, ne peuvent s'exprimer à l'aide d'un seul et même formalisme. Cette hypothèse est directement concrétisée au sein de LinguaStream, et il s'agit là d'un point qui la différencie fortement d'autres outils qui privilégient bien souvent un modèle d'analyse unique. Comme le montre la figure 10.3, il est donc possible de combiner au sein d'une même chaîne de traitement différents composants nourris de règles données dans des formalismes distincts, ici un automate et une grammaire d'unification.

Le plus souvent, l'utilisation de formalismes différents sera liée à la prise en compte successive de différents niveaux linguistiques. Typiquement, si une grammaire d'unification pourra se révéler très efficace pour procéder à une analyse d'ordre syntagmatico-sémantique, ce formalisme s'avérera généralement inadapté au niveau discursif. L'exemple de l'analyse des cadres de discours que nous avons décrite dans le chapitre 7 constitue là encore un cas d'école, puisqu'elle met à contribution une grande variété de formalismes différents (cf. figure 10.1) :

– le découpage en mots s'appuie sur un formalisme de type expressions régulières au niveau des caractères (règles de type « RT » – Regexp Tokenizer) ;

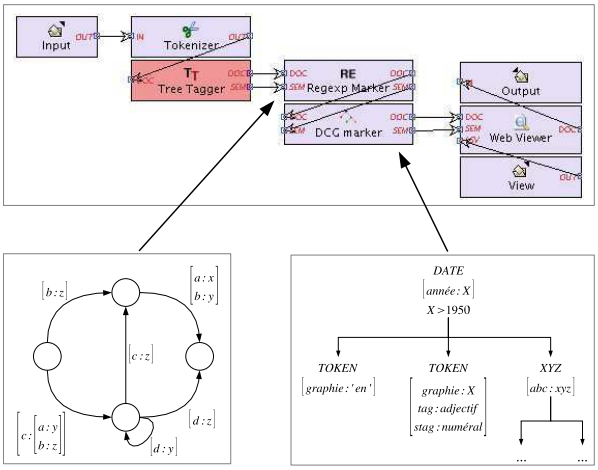


FIG. 10.3 – Complémentarité des modèles d'analyse au sein d'une même chaîne.

- l'analyse syntagmatico-sémantique des expressions temporelles fait appel à des grammaires d'unification (règles de type « EDCG » – Extended Definite Clause Grammar) ;
- la détection de connecteurs de discours fait appel à la projection d'un lexique sémantique (règles de type « LSL » – LinguaStream Lexicon) ;
- la détection des bornes des phrases et de certaines propositions exploite des macro-expressions régulières (règles de type « MRE » – Macro Regular Expression) ;
- les introducteurs de cadres sont également décrits à l'aide de ce dernier formalisme ;
- les propriétés des cadres de discours eux-mêmes sont décrites sous la forme d'une grammaire de contraintes (règles de type « CDML » – Constraint-based Discours Modelling Language).

On remarquera dans cet exemple qu'à chaque niveau linguistique correspond un formalisme. Chacun d'entre eux a de fait été utilisé pour traiter le niveau linguistique pour lequel il a été préférentiellement conçu. Il est toutefois important de mentionner que la notion de perspective d'analyse que nous décrirons dans la section suivante augmente largement le rayon d'action de chacun de ces formalismes, puisqu'il permet par exemple de jouer sur le niveau de grain pour appliquer des grammaires d'unification au niveau discursif.

Il convient également de mentionner un dernier « formalisme », même s'il s'éloigne de ce que recouvre généralement ce terme : il s'agit de la représentation du traitement global par une chaîne de traitement. En effet, on peut considérer que la chaîne de traitement comme une certaine forme de représentation du traitement à effectuer, qui rend explicite la collaboration au sein du modèle global d'un ensemble de sous-modèles, tout en prescrivant la machine la succession des tâches à réaliser pour aboutir au résultat final.

Techniquement, ce principe de collaboration entre modèles d'analyse est rendu possible par l'architecture modulaire que nous décrivions dans la section 10.1, puisqu'à chaque formalisme correspond un composant différent. Mais aussi et surtout, elle est rendue possible par la mise en oeuvre d'un modèle documentaire uniforme, qui sera décrit dans le chapitre 11. En particulier, tous les formalismes s'appuient sur une représentation unifiée des unités préalablement marquées et des annotations (sous forme de structures de traits).

10.3 La notion de perspective d'analyse

Nous avons vu dans la section précédente que la plate-forme permet d'exploiter un formalisme adapté à chaque unité prenant part à un modèle global, se matérialisant par différents modules d'une même chaîne de traitement. Dans cette partie nous allons évoquer la notion de « perspective d'analyse », qui permet à chacun de ces modules d'adopter un *point de vue spécifique* sur le texte. En d'autres termes, une perspective d'analyse définit comment le texte et surtout les annotations qu'il contient seront perçus par un module donné.

Les différentes propriétés d'une perspective d'analyse sont les suivantes :

- classes d'unités à considérer comme jetons ;
- classes d'unités définissant le domaine d'analyse ;
- filtres à appliquer sur les annotations antérieures.

La possibilité de définir *localement* la (ou les) classe(s) d'unités qui doivent être considérées comme jetons garantit la *variabilité du grain d'analyse* au cours du traitement. Elle concerne les nombreux modèles qui imposent la définition d'un grain minimal, dit jeton ou *token*. C'est par exemple le cas de toute grammaire ou transducteur : ces formalismes supposent l'existence d'une unité textuelle (comme le caractère ou le mot) à laquelle s'appliquent les patrons. Quand la définition de ce grain minimal est nécessaire au fonctionnement d'un composant, la plate-forme permet de spécifier localement le ou les types d'unités à considérer comme jetons. Toute unité préalablement délimitée peut jouer ce rôle : il pourra s'agir du découpage habituel en mots, ou de toute autre type d'unité : syntagmes, phrases, cadres du discours, etc. Le grain minimal peut donc être différent pour chaque palier d'une chaîne, ce qui augmente considérablement la portée des différents modèles d'analyse disponibles dans la plate-forme. Il sera par exemple possible d'écrire une grammaire de texte à l'aide de grammaires d'unification (formalisme EDCG).

Ce principe fonctionne de la manière suivante. Au niveau du modèle documentaire, il n'existe pas de jeton dans le sens où toutes les annotations sont équivalentes : comme nous le verrons dans la section 11, toute annotation produite par la plate-forme se matérialise par des jalons qui délimitent des segments textuels. Par défaut, un tel marquage sera considéré comme une séquence d'objets textuels composée du jalon de départ, puis des différentes unités contenues (texte ou autres marquages), et enfin du jalon de fin[7]. En revanche, lorsqu'une unité appartient à une classe « jeton », elle n'est plus considérée comme une séquence mais comme une unité atomique, dont le contenu est purement textuel.

La figure 10.4 représente deux « visions » possibles d'un même marquage, selon les unités qui seront considérées comme jetons. Il pourrait par exemple s'agir d'un marquage délimitant les mots et d'un autre autre délimitant les phrases. Dans ce cas, on serait généralement amené à considérer les unités de type « mot » comme jetons et les unités de type « phrase » comme succession de jalons. Dans ce cas, le document sera perçu comme une séquence d'objets qui pourrait être :

$$\ldots \quad debut_{phrase} \quad jeton_{mot} \quad jeton_{mot} \quad \ldots \quad jeton_{mot} \quad fin_{phrase} \quad \ldots$$

Mais il est également possible de spécifier le type « phrase » comme désignant des jetons. Dans ce cas, les phrases seront considérées comme des unités atomiques, et le marquage des mots (du moins ceux qui sont effectivement compris dans un marquage de type « phrase ») ne sera plus accessible :

$$\ldots \quad jeton_{phrase} \quad jeton_{phrase} \quad jeton_{phrase} \quad \ldots$$

Dans tous les cas, les annotations associées à un objet (jalon ou jeton) seront bien sûr disponibles

[7] Nous écartons ici le cas des marquages ambigus ou non-contigus, sur lesquels nous reviendront plus tard.

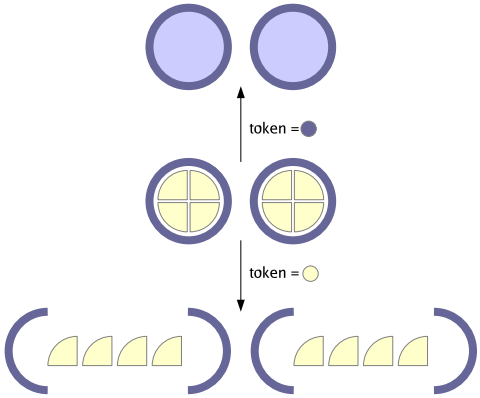


FIG. 10.4 – Illustration de la variabilité du grain « jeton ». Un même type d'unité pourra apparaître soit sous la forme d'entités indivisibles, soit en tant que jalons entourant d'autres unités plus petites.

et il sera possible de les exploiter pour spécifier des contraintes. Les jetons ont la particularité de posséder un contenu textuel (à l'inverse des jalons qui sont vides), qui sera également accessible à la spécification de contraintes. En pratique, ce contenu textuel sera surtout utilisé pour les unités de grain relativement faible (comme les mots), alors que pour les unités de grain plus important on n'exploitera généralement que les structures de traits associées. Mais bien sûr rien n'empêche de procéder différemment si nécessaire (on peut par exemple rechercher un syntagme ou même un paragraphe dont le dernier caractère serait un « s » ou un « x »).

La seconde composante d'une perspective d'analyse concerne le filtrage. Beaucoup plus simple que celle que nous venons de décrire, cette fonctionnalité est toutefois intéressante car elle permet de se détacher partiellement de la linéarité du texte, c'est-à-dire de la nécessité d'envisager systématiquement tous les éléments du flot textuel. Elle autorise en effet un composant à adopter une vue sélective sur les annotations déjà présentes dans le document, comme l'illustre la figure 10.5, de façon ne prendre en compte dans l'écriture des règles que les éléments voulus. Précisons que là encore, il n'est question que du point de vue adopté par un module sur le texte : les annotations filtrées ne sont en aucun cas supprimées du document, mais seulement masquées à un certain point de la chaîne de traitement.

La troisième et dernière composante d'une perspective d'analyse permet de restreindre le champ d'application des règles à certaines classes d'unités, champ que nous qualifions de « domaine d'analyse ». On pourra par exemple demander à ce que les unités reconnues par un module donné se positionnent obligatoirement à l'intérieur de paragraphes ou de phrases. Cela peut tout d'abord être utile pour éviter que les unités reconnues ne chevauchent d'autres unités déjà marquées (par exemple des syntagmes, qui ne sauraient chevaucher des phrases). Mais cela est également important du point de vue des performances. Ainsi, quand l'algorithme du modèle d'analyse employé n'est pas déterministe, on s'efforcera de limiter le domaine d'analyse au grain le plus fin possible de façon à éviter des calculs inutiles. Ou plus simplement, on pourra souhaiter éviter le marquages de certaines zones du document qui n'ont pas lieu d'être analysées (par exemple les zones de métadonnées).

10.4 Exploitation systématique des standards et outils XML

Un dernier parti-pris, plus technique, que nous revendiquons à travers LinguaStream est l'utilisation systématique des standards et outils appartenant à « la sphère XML ». Tout d'abord, et de façon aujourd'hui tout à fait naturelle, nous utilisons systématiquement ce méta-langage pour décrire les

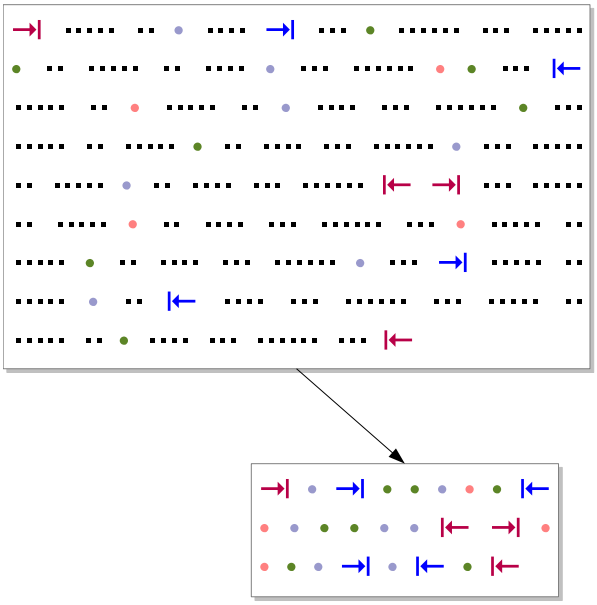


FIG. 10.5 – Illustration du filtrage des annotations. Les flèches représentent des jalons et les points des jetons. Chaque composant peut spécifier un filtre lui permettant de modifier sa vue sur le texte, en ne sélectionnant qu'une partie des analyses déjà présentes dans le texte.

fichiers manipulés par la plate-forme : chaînes elles-mêmes, lexiques, différents types de règles etc. L'intérêt de cette démarche est qu'il est alors très facile de générer automatiquement ces fichiers, par exemple pour fabriquer automatiquement un lexique, une règle MRE, ou même une chaîne de traitement. Il est également très simple de produire à partir d'eux de nouveaux fichiers, comme des triplets RDF à partir d'un lexique ou encore une documentation XHTML+SVG d'une chaîne de traitement.

D'autre part, les documents traités par la plate-forme seront eux-mêmes représentés en XML[8]. La plate-forme a pour particularité, comme nous le verrons dans le prochain chapitre, de pouvoir traiter tous les documents de ce type, indépendamment de leur propre schéma. Nous nous autorisons ainsi l'accès à tous les outils associés au méta-langage, et nous avons fait le choix de les exploiter dès que cela était possible. Cela se traduit tout d'abord par l'inclusion dans la plate-forme d'une série de composants permettant d'insérer dans une chaîne différents traitements XML aujourd'hui standards, comme XSL-T, XQuery ou SPARQL, ainsi que des applications plus spécifiques comme FOP, qui permet de mettre en page des documents XSL-FO.

D'autre part, nous exploitons les outils XML généralement disponibles en dehors de la plate-forme. C'est par exemple le cas concernant la visualisation des documents annotés : les composants de visualisation considéreront que le document fourni en entrée est capable de s'afficher par lui même dans un navigateur[9] acceptant les standards XML (soit qu'il s'agisse d'un document XHTML, soit qu'il soit muni d'une feuille de style), et ajoutera simplement des informations de style relatives à l'affichage des annotations elles-mêmes. Nous appliquons le même principe à l'annotation manuelle des documents, en nous reposant sur les outils adaptés (même s'il faut bien reconnaître que l'offre reste à ce jour limitée sur ce point précis). Nous nous soulageons ainsi de tâches de développement relativement lourdes, tout en bénéficiant d'une ouverture significative vers les applications tierces.

[8]On notera que cela n'est pas une obligation liée à la plate-forme en tant que telle, mais plutôt aux composants orientés TAL qu'elle propose : beaucoup d'autres composants de la plate-forme sont en fait capables de traiter d'autres types de fichiers.

[9]Comme FireFox de la Fondation Mozilla.

Chapitre 11

Modèle documentaire

L'un des aspects importants de LinguaStream en tant que plate-forme est le modèle documentaire sur lequel elle repose, qui est commun à tous les composants dédiés au traitement du langage naturel. À chaque étape d'une chaîne de traitement, il est utilisé pour représenter les marquages et annotations produits, assurant ainsi l'homogénéité des informations produites par les différentes analyses ainsi que le transfert de ces informations d'un composant à l'autre, et constitue à ce titre l'épine dorsale de la plate-forme. Il se matérialise d'une part par un jeu de schémas XML mis en oeuvre pour représenter les annotations elles-mêmes, et d'autre part par une API sur laquelle reposent les composants LinguaStream spécifiquement liés au TAL.

Nous allons tout d'abord décrire ce modèle de façon abstraite, tel qu'il est perçu par les modèles d'analyse. Dans un second temps, nous décrirons les modalités concrètes de sa réalisation dans la plate-forme, qui concernent les procédés de balisage XML et une API de gestion des documents ainsi marqués, que nous décrirons succinctement.

11.1 Modèle abstrait

Le modèle documentaire mis en oeuvre par la plate-forme est délibérément simple, de façon à garantir une grande flexibilité et une indépendance totale vis à vis des différents types de documents à annoter et des différents types d'analyse linguistique qui seront représentées.

11.1.1 Unités primitives et secondaires

Dans ce modèle comme dans beaucoup d'autres, un document est tout d'abord conçu comme une séquence d'objets indivisibles, correspondant aux unités définies par le système d'encodage utilisé pour représenter numériquement ce document sur son support. Nous qualifierons ces unités de *primitives*. En pratique, celles-ci correspondront aux caractères définis par Unicode ou tout autre procédé d'encodage de signes graphiques. Dans la grande majorité des cas, il s'agira des graphèmes d'une langue (lettres d'un alphabet, idéogrammes), mais il pourra également s'agit d'unités de grain inférieur, par exemple si les signes diacritiques sont encodés séparément, ou même d'unités d'ordre différent comme celles d'un alphabet phonétique.

Nous appellerons *unité secondaire* un ensemble d'au moins deux *jalons* insérés au sein de la séquence d'unités primitives qui forment le document, constituant un ou plusieurs groupes d'unités primitives. Bien entendu, ces nouvelles unités correspondront généralement à celles d'un certain modèle linguistique (syllabes, mots, syntagmes, propositions, paragraphes, etc.), mais notre modèle documentaire ne fait aucune hypothèse quant à leur nature effective : tout ensemble d'unités primitives peut

former une unité secondaire.

Les unités secondaires sont elles-mêmes regroupées en *classes* identifiées par un identifiant choisi arbitrairement que nous appellerons *type*. D'autre part, les différentes unités appartenant à une même classe seront distinguées par des identifiants uniques au sein de cette classe, constitués par des valeurs numériques. Ainsi, l'identification univoque d'une unité secondaire donnée nécessite la mention de sa classe (ou type) ainsi que de son identifiant au sein de cette classe. Nous utiliserons également le terme de *marquage* pour désigner une classe d'unités secondaires, et les désignerons généralement par un identifiant évoquant le type d'objet documentaire ou linguistique auquel ce marquage se rapporte (comme « mot », « word », « syntagme », « SN » ou « cadre »).

11.1.2 Organisation en couches

Un seul et même document pourra contenir un nombre quelconque d'unités de classes différentes. Plusieurs marquages pourront en effet cohabiter sans contrainte particulière, dans la mesure où les unités secondaires sont spécifiées par des ensembles de jalons (initiaux ou finaux) qui doivent être considérés comme « ponctuels ». Différentes unités secondaires pourront également se chevaucher, dans le sens où l'un des jalons délimitant une unité pourra s'intercaler entre les deux jalons délimitant une autre unité. De même, les unités secondaires sont indépendantes de la structure propre au document traité. Par exemple, si le balisage propre du document identifie les bornes de paragraphes, on pourra tout de même identifier des unités chevauchant ces bornes.

Les différents marquages d'un même document peuvent donc être considérés comme appartenant à des couches distinctes et indépendantes, reposant sur une couche d'arrière-plan constituée par la structure propre du document. Dans certains cas, des conflits pourront bien sûr apparaître lorsque l'on cherchera à considérer certaines unités secondaires comme des ensembles autonomes (constitués de l'ensemble de leurs jalons et des unités primaires qu'ils encadrent), mais comme nous le verrons dans la section suivante, cela ne relève que d'un point de vue local sur le marquage et ne remet aucunement en cause l'autonomie des couches en tant que telles.

Le modèle prévoit par ailleurs qu'à chaque classe d'unités soit associée une valeur numérique spécifiant la profondeur de la couche à laquelle elle appartient. On peut en effet souhaiter, lorsque plusieurs jalons marquent le même point d'un document, qu'un ordre précis soit systématiquement respecté. En d'autres termes, il est possible de spécifier un ordre dans lequel seront ordonnées les différentes couches de marquage. Cela permet notamment de garantir qu'un certain grain de marquage soit systématiquement inclus dans un autre. On pourra par exemple souhaiter qu'un marquage d'ordre syntagmatique soit systématiquement placé, en cas d'ambiguïté, à l'intérieur des marquages de grain supérieur (phrase, paragraphe, etc.). Il suffira pour cela d'associer une « profondeur » plus importante au marquage syntagmatique, signalée par une valeur numérique plus faible. Cette valeur numérique peut être négative ou positive, la couche par défaut possédant la valeur zéro.

11.1.3 Constitution de « jetons »

Nous avons décrit dans la section 10.3 la possibilité de définir pour chaque composant d'une chaîne de traitement le ou les types des unités qui doivent être considérées comme jetons, les autres étant considérées comme simples jalons. Cela signifie qu'une même unité secondaire pourra être considérée de deux façons :

- Soit comme une unité en tant que telle, c'est-à-dire un groupe d'unités primitives que l'on considère alors comme un tout autonome appelé *jeton*.
- Soit comme un simple couple de jalons se mêlant au flux des autres unités secondaires. Dans ce cas, ce sont les jalons eux-mêmes qui sont considérés comme des unités autonomes.

Il nous semble important d'insister ici sur le fait que les jetons sont constitués *dynamiquement*, selon la perspective d'analyse adoptée[1]. Ils n'ont donc aucune existence au sein du modèle documentaire que nous décrivons ici. Bien-sûr, si plusieurs unités considérées comme jetons se chevauchent, il conviendra d'appliquer des règles de résolution de conflits que nous évoquerons plus loin. Mais cela ne remet pas en cause l'indépendance de couches de classes d'unités secondaires puisque ces conflits ne résultent que d'un point vue ponctuellement porté sur le document et non modèle documentaire lui-même.

11.1.4 Jalons multiples

De toute évidence, une unité secondaire est au moins délimitée par deux jalons, l'un précédant la première unité primitive du groupe, et l'autre succédant à la dernière unité primitive. Toutefois, notre modèle prévoit l'utilisation de jalons supplémentaires, dans deux cas :

- Une unité secondaire peut être composée de plusieurs groupes disjoints d'unité primitives. Dans ce cas, un nombre pair de jalons sera utilisé pour identifier chacune des composantes disjointes de l'unité secondaire. Cette mode d'annotation sera typiquement utilisé pour regrouper virtuellement les différentes composantes non consécutives d'un même objet linguistique (verbe composé par exemple).
- Le modèle prévoit également qu'il y ait plus d'un jalon de début et/ou de fin, généralement pour rendre compte d'une ambiguïté. Par exemple, si l'on souhaite indiquer que l'on n'a pu statuer sur l'emplacement de la fin d'une phrase, il est possible d'attribuer plusieurs jalons de fin à une même unité secondaire, de façon à rendre compte de l'indécision, qui pourra éventuellement être résolue par la suite.

De façon à identifier les différents jalons appartenant à une même unité secondaire, il est indispensable que tous ses jalons partagent à la fois la même classe et le même identifiant. On notera également que dans le cas où une unité secondaire est spécifiée par plus de deux jalons, on sera généralement amené à considérer ces jalons comme unités autonomes, de façon à pouvoir appréhender correctement cette multiplicité selon l'application visée. Si toutefois on persistait à considérer de telles unités comme des jetons, les règles de résolution de conflits décrites plus loin s'appliqueront. On remarquera enfin que tous les modèles d'analyses ne sont pas nécessairement aptes à gérer les unités disjointes ou ambiguës.

11.1.5 Représentation des annotations

Outre la délimitation des unités elles-mêmes, notre modèle documentaire permet de leur associer des informations représentées sous forme de structures de traits. Il s'agit d'un modèle de données simple mais souple, très communément utilisé en linguistique formelle et en traitement automatique des langues, pour l'expression de grammaires fonctionnelles par exemple. Il a pour avantages principaux d'autoriser la représentation d'informations relativement complexes, et de se prêter facilement à l'expression de contraintes via des mécanismes tels que l'unification. Le modèle prévoit que chaque unité secondaire puisse être associée à au plus une structure de traits, sans faire d'hypothèse sur la nature des informations ainsi représentées : les structures de traits pourront aussi bien représenter un étiquetage morpho-syntaxiques que des représentations sémantiques, ou encore le résultat d'une analyse rhétorique.

Les structures de traits sont définies récursivement : une structure est déterminée par un ensemble de couples attribut/valeur (dits « traits ») ; une valeur peut être soit atomique, soit une nouvelle structure de traits. Chaque attribut est lui-même déterminé par un identifiant arbitraire, qui doit être unique au

[1] C'est en pratique l'API « EBMS » décrite plus loin qui se charge de constituer les jetons.

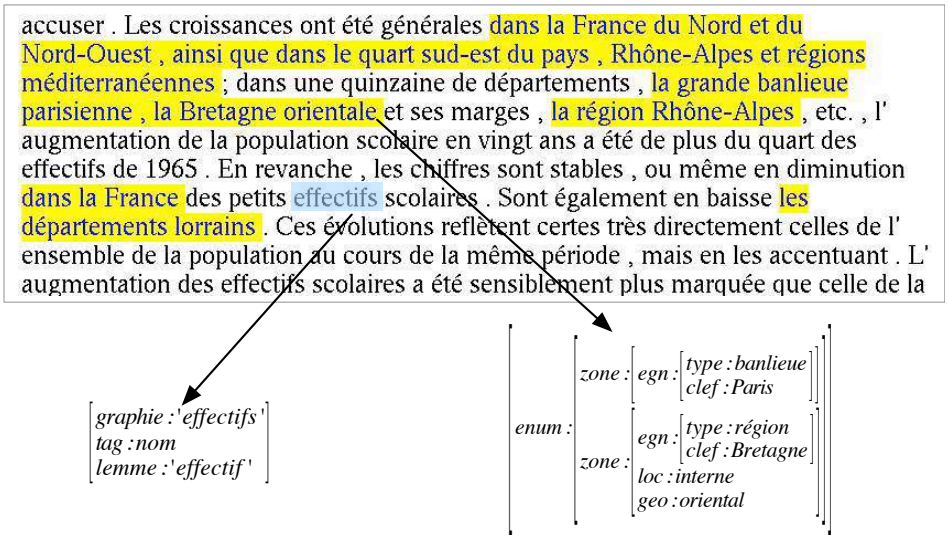


FIG. 11.1 – Exemples d'annotations LinguaStream.

sein de l'ensemble de traits auquel il appartient. Il s'agit donc de structures arborescentes où les noeuds sont constitués des traits, et les feuilles des traits dont la valeur est atomique[2]. On notera que nous ne faisons pour l'heure pas usage des diverses « extensions » possibles de ce modèle, comme le typage des traits ou la représentation de listes, même s'il n'est pas exclu que cela soit le cas dans une prochaine version de la plate-forme.

Voici un exemple de structure contenant trois traits « a », « b » et « c », les valeurs des deux premiers étant atomiques alors que celle du troisième est une nouvelle structure contenant les sous-traits « d » et « e », etc. :

$$\begin{bmatrix} a : & \text{x} \\ b : & 42 \\ c : & \begin{bmatrix} d : & \begin{bmatrix} f : & \text{y} \\ g : & \text{z} \end{bmatrix} \\ e : & \ldots \end{bmatrix} \end{bmatrix}$$

Toutes les annotations produites par les composants d'une chaîne de traitement seront donc représentés sous cette forme, quelle que soit la nature du traitement sous-jacent. Inversement, tous les formalismes proposés par la plate-forme permettent d'exploiter ces données, généralement par l'expression de règles portant sur les structures de traits des unités préalablement analysées.

La figure 11.1 reproduit un exemple de texte présentant à la fois l'annotation des expressions spatiales (cf. section 6) et un étiquetage morpho-syntaxique. On peut y observer à la fois les marquages eux-mêmes et les annotations associées sous forme de structures de traits :

11.1.6 Représentation des relations

En plus des marquages et des annotations que nous venons d'évoquer, le modèle documentaire de LinguaStream prévoit la représentation de relations entre les unités marquées. Là encore, aucune hypothèse n'est faite quant à la nature de ces relations (ni des unités mises en relation), qui pourront

[2]On remarquera qu'une structure correspond plus exactement à une forêt, puisqu'une structure de traits n'a pas de racine, mais est au contraire formée de plusieurs arbres

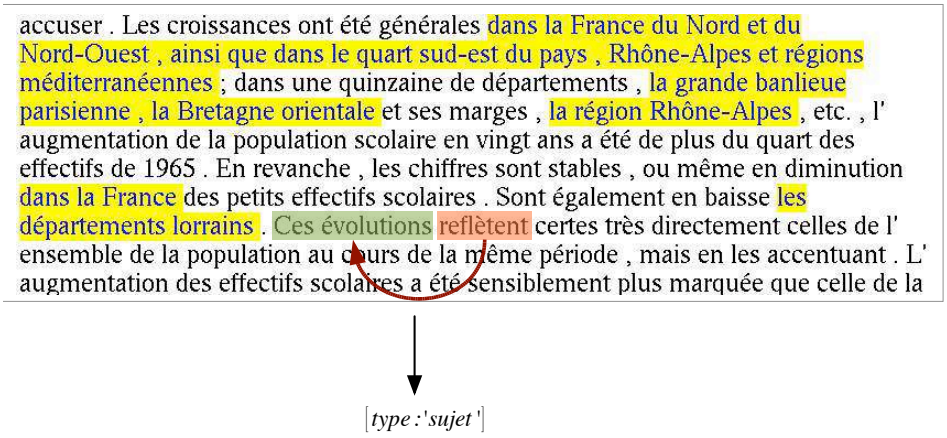


FIG. 11.2 – Exemple de relation LinguaStream.

aussi bien représenter des liens d'ordre syntaxique, rhétorique, thématique, ou même d'alignement (cf. section 7.3). Une relation est constituée d'un ou plusieurs arcs, dont chacun connecte deux unités secondaires. Même si dans le cas général une relation correspond à un seul arc, notre modèle prévoit donc qu'une seule et même relation soit constituée d'un nombre arbitraire d'arcs. En outre, chacun de ces arcs peut éventuellement être orienté.

Tout comme les marquages, les relations sont regroupées en classes, identifiées par un identifiant dit « type ». Là encore, ces identifiants sont arbitraires mais feront généralement référence à la réalité linguistique recouverte par lesdites relations. À chaque relation peut également être associée une structure de traits, sous la même forme que les annotations des marquages eux-mêmes. Ces structures sont totalement indépendantes de celles des unités connectées, et peuvent être utilisées pour représenter des informations spécifiquement liées à la relation elle-même.

La figure 11.2 présente un exemple de texte où deux unités ont été marquées (un groupe sujet et un groupe verbal) puis connectées par une relation syntaxique sujet-verbe.

11.2 Représentation concrète

Comme toutes les autres données manipulées par la plate-forme, les annotations des documents (marquages, structures de traits et relations) sont représentées en XML, de façon à bénéficier des apports bien connus de ce format standard, notamment en termes d'interopérabilité avec d'autres sources de données et de facilité de manipulation grâce aux nombreux outils qui lui sont associés. Toutefois, le codage XML d'annotations à la fois conformes aux objectifs fixés dans le chapitre précédent et au modèle abstrait que nous venons de spécifier est loin d'être immédiat, et implique un certain nombre de choix de conception qui font l'objet de cette partie.

Le premier problème concerne la possibilité que nous souhaitions donner à la plate-forme d'annoter tout type de document XML, indépendamment de son schéma, et en préservant totalement son intégrité. Cela signifie qu'au fil des modifications induites par les différentes étapes d'une chaîne de traitement, aucune information du document original ne devra être perdue, et que les informations ajoutées par la plate-forme devront être transparentes pour les applications qui ne sont pas expressément conçues pour les exploiter. Bien-sûr, cela implique également que lors de l'insertion de nouvelles annotations ou même de nouveaux éléments dans le document ne perturbe en rien les annotations préalablement insérées. Et enfin, nous avons mentionné plus haut le fait que notre modèle documentaire supporte les chevauchements entre unités secondaires, ce qui n'est pas, on s'en doute, sans poser de

difficultés liées au modèle purement arborescent imposé par XML.

Au sein des différentes approches envisageables pour procéder à l'annotation de documents, on distingue généralement deux grandes catégories dites « débarquées » (ou *stand-off*) et « embarquées » (ou *inline*). La première correspond à une annotation externe, n'impliquant absolument aucune modification du document lui-même. Les annotations sont ainsi stockées dans un format quelconque, tout en étant liées aux segments textuels qu'elles décrivent grâce un mécanisme permettant de « pointer » vers une portion précise d'un document XML, par exemple XPointer (Grosso *et al.*, 2003). La possibilité de conserver intact le document annoté est bien sûr très intéressante, et cette approche a de plus pour intérêt de gérer *de facto* le problème du chevauchement des annotations entre elles. Cette approche a toutefois pour inconvénient majeur d'être très sensible aux altérations du document après annotation. En effet, les mécanismes de pointage vers un fragment textuel reposent dans ce cas sur l'identification de sous-chaînes par les index de leurs caractères de début et de fin, du moins pour les unités qui ne sont pas déjà balisées dans le document annoté, comme c'est souvent le cas pour les unités de grain inférieur au paragraphe. Et même pour des unités déjà balisées, le pointage devra reposer sur les positions respectives de ces unités, à moins que des identifiants uniques ne leur soient systématiquement associées, ce qui est loin d'être toujours le cas. Or il est bien évident que toute modification ou fragmentation du document invalide les pointeurs de ce type. Un autre inconvénient réside dans le fait que le post-traitement de documents ainsi annotés (par exemple par un composant aval de la chaîne de traitement) est très peu efficace puisque la détection des unités marquées peut avoir un coût non négligeable. On notera enfin que le rendu visuel de ces annotations nécessite des outils spécialisés, encore rares aujourd'hui.

L'annotation *inline* est au contraire interne au document, puisque toutes les unités annotées sont directement balisées dans le document lui-même, et sont accompagnées des informations qui leurs sont associées. Les annotations de ce type sont, à l'inverse des annotations *stand-off*, très peu sensibles aux altérations du document. En revanche, le stockage des données au sein même du document peut avoir d'autres inconvénients. D'une part, la présence même de ces informations au sein du texte peuvent, si l'on n'y prend pas garde, rendre ces documents illisibles pour des outils qui ne sont pas conçus pour exploiter (ou simplement ignorer) ces données. Le volume des documents annotés peut aussi augmenter de façon rédhibitoire si les annotations sont nombreuses, rendant inopérants les outils XML procédant à la représentation exhaustive des documents en mémoire (type DOM). D'autre part, la gestion du chevauchement des annotations entre elles ou avec le balisage propre du document peut poser problème si des précautions adéquates ne sont pas prises. Enfin, l'accès aux annotations nécessite une recherche au sein du document lui-même, alors que dans bien des cas on souhaitera pouvoir y accéder directement, par exemple en les stockant dans un entrepôt de données adéquat afin de pouvoir formuler des requêtes efficacement[3].

Pour toutes ces raisons, aucune des approches *inline* et *stand-off* ne nous a paru pleinement satisfaisante, et l'ensemble des objectifs que nous avons décrits mentionnés plus avant nous a finalement conduit à adopter une approche *hybride*, dans le sens où elle conduit à la modification du document annoté par insertion de balises autour des unités annotées, tout en stockant les annotations elles-mêmes (structures de traits et relations) dans des documents distincts. Cette solution convient à nos objectifs dans la mesure où elle résiste aux modifications du document tout en insérant le minimum d'informations dans le document lui-même, et en séparant clairement les marquages des informations associées.

Il nous faut toutefois préciser les modalités de marquages que nous mettons en oeuvre pour résoudre certains problèmes qui ne sont pas immédiatement résolus par l'approche hybride, notamment la préservation de l'intégrité du document original et la gestion du chevauchement potentiel des annotations (entre elles ou avec le balisage propre du document).

[3]Cela sera typiquement le cas dans des applications de type recherche d'information, où les annotations constitueront *in fine* des index.

Outre la mécanique logicielle qui reproduit fidèlement tous les éléments du document lu, la préservation de l'intégrité du document original est garantie par la forme des balises insérées autour des unités à marquer. Celles-ci ont en effet les propriétés suivantes :

– Elles appartiennent à un espace de nom spécifique à la plate-forme, au sens défini dans (Bray *et al.*, 1999). Cela garantit d'une part qu'il n'y ait pas de collision avec un balisage appartenant au document original, précaution indispensable dans la mesure où nous ne faisons aucune hypothèse sur le schéma de ce document. D'autre part, cela permet aux applications tierces d'ignorer purement et simplement le marquage généré par la plate-forme si elles ne le reconnaissent pas. Cela suppose évidemment que les dites applications se comportent correctement face à des espaces de noms inconnus, ce que l'on peut raisonnablement attendre de leur part dans la mesure où il s'agit là d'une « bonne pratique » largement répandue.
– Elles constituent d'autre part des couples de balises vides, que nous appellerons *ancres*[4]. Ces balises ont pour particularité de n'avoir aucun contenu propre, et constituent donc des objets purement « ponctuels » qui ne peuvent en aucun cas interférer avec les autres éléments d'un document. En d'autres termes, l'ablation ou l'insertion d'une telle balise ne peut en aucun cas mettre en cause la bonne forme du document.

Précisons brièvement les modalités d'utilisation de ces ancres pour procéder au balisage des unités secondaires. Dans le cas d'une unité continue et non-ambiguë, seulement deux ancres seront insérées, l'une marquant le début et l'autre la fin, celles-ci étant distingués par la valeur d'un attribut prévu à cet effet. Le lien entre les deux ancres s'opère grâce à la classe (ou type) d'annotation auquel s'ajoute un identifiant unique au sein de cette classe. Ce couple (type ; id) forme un identifiant univoque au sein du document, et se trouve mentionné sur chaque ancre sous la forme d'attributs. Précisons que cet identifiant est bien propre à l'unité secondaire ainsi marquée, et non aux différents noeuds XML qui la matérialisent. Pour identifier ces derniers, des identifiants XML au sens de (Marsh *et al.*, 2005) peuvent également être ajoutés, tout en demeurant extérieurs à notre modèle documentaire. Un dernier attribut est également positionné sur chaque ancre de façon à mentionner, le cas échéant, la couche à laquelle appartient l'unité marquée.

Dans le cas d'une unité discontinue, c'est-à-dire formée de plusieurs segments non consécutifs, plusieurs couples d'ancres seront utilisés. Dans le cas d'une unité ambiguë, plusieurs ancres de début et/ou de fin seront insérés. Comme pour les autres unités, l'ensemble de ces ancres seront liées entre elles par le couple (type ; id) qui identifie l'unité secondaire qu'ils marquent dans le document.

Précisons enfin que le schéma XML utilisé par LinguaStream prévoit un mode alternatif de marquage, qui peut être utilisé lorsque les contraintes mentionnées plus haut, notamment en termes de chevauchement, peuvent être relâchées au profit d'un marquage plus conventionnel. Le marquage par ancres présente en effet l'inconvénient d'être difficilement accessible aux outils XML habituels tels que XSLT ou CSS. Dans certains cas particulier, un utilisateur souhaitant disposer d'un balisage XML plus « conventionnel » dans le cadre d'une application particulière pourra également l'obtenir via un réglage adéquat des composants ou de l'API. Le balisage obtenu dans ce cas est dit de type « bloc », et se matérialise par des balises XML habituelles, comportant une balise ouvrante et une balise fermante, le contenu de l'unité étant alors inclus, au sens DOM, entre ces balises.

Dans ce cas, une balise spécifique est utilisée, comportant seulement les attributs spécifiant la classe, l'identifiant et la couche. Bien-sûr, ce mode de marquage est réservé aux cas où l'on peut s'assurer de l'absence de risque de chevauchement (dans le cas contraire une erreur sera signalée par la plate-forme), et ne peut s'appliquer aux unités ambiguës. Notons que le passage des balises de types « ancre » en balises de type « bloc » est réalisé au cas par cas (pour une classe d'unités ou même seulement certaines unités), et qu'il peut également intervenir *a posteriori* à l'aide d'un composant fourni

[4]Les balises de ce type sont souvent appelées *jalons* (ou *milestones*), mais nous réserverons ici ce terme aux unités du modèle abstrait.

avec la plate-forme, ou d'une simple feuille XSLT. Il est important de mentionner que l'API que nous décrirons plus loin garantit l'abstraction totale des modalités pratiques de marquage. En particulier, le choix d'un marquage de type « ancre » ou « bloc » n'aura aucun impact sur la vision qu'en donnera cette API, et donc aucune incidence sur les traitements susceptibles d'exploiter ce marquage.

Le schéma spécifiant ces modalités de marquage se limite à deux types d'éléments correspondant aux balisages de type « ancre » (élément « a ») et de type « bloc » (élément « b »). Les attributs utilisés pour spécifier le type de l'unité, son identifiant, sa couche et le cas échéant le type d'ancre sont respectivement « type », « id », « layer » et « anchor ». Le schéma complet au format XSD est donné en annexe E.5.

Les structures de traits associées aux marquages et les relations sont comme on l'a dit stockées dans des fichiers séparés, en faisant appel à des schémas spécifiques qui ne présentent pas de particularité notable. On notera simplement que pour une plus grande facilité d'utilisation par le biais des outils XML habituels, les structures de traits sont représentées par des balises qui sont le reflet immédiat des traits utilisés : nous n'utilisons donc pas de meta-modèle sur ce point, ce qui signifie que chaque trait est immédiatement transposé sous la forme d'une balise portant son nom. On notera toutefois la disponibilité de mécanismes permettant quand cela est nécessaire d'utiliser des espaces de noms propres à chaque type d'analyse. On notera par ailleurs l'existence au sein de la plate-forme d'outils permettant d'exporter ces fichiers sous une forme RDF.

11.3 L'interface de programmation « EBMS »

Pour faciliter le développement de nouveaux modules de traitement conformes au modèle que nous venons de spécifier, la plate-forme propose une interface de programmation (API) qui masque toute la complexité inhérente aux modalités de marquage XML (cf. section 11.1) tout en reprenant les concepts du modèle abstrait (cf. section 11.2). Cette API est appelée « Event-Based Markup System » (EBMS), pour des raisons que nous mentionnerons plus loin. Elle constitue le fondement de la très grande majorité des composants fournis avec la plate-forme, et est également utilisée par les utilisateurs développant des modules tiers (cf. section 14.2). Elle permet également de développer facilement des modules de traitement *ad-hoc* à l'aide des différents langages de script interprétés qui sont directement accessibles depuis la plate-forme.

La figure 11.3 représente une vue simplifiée de cette API. Son coeur est le moteur de traitement des documents XML (classe « Engine »), qui prend en charge tous les principes de marquage décrits précédemment. Il permet de lire un document XML quelconque, de rendre compte à l'utilisateur de l'API des annotations qu'il contient, puis si nécessaire de reproduire fidèlement le document original en ajoutant éventuellement de nouvelles annotations. Le moteur tiendra compte de la perspective d'analyse (cf. 10.3) spécifiée par l'utilisateur, qui concerne notamment ici les types de jetons et le filtrage des jalons, et éventuellement le domaine d'analyse.

Les autres composantes de l'API sont directement liées aux principes du modèle abstrait. On trouvera notamment une classe abstraite « DiscourseObject » qui désigne un objet textuel quelconque. Une première spécialisation de cette classe est appelée « Milestone », et correspond immédiatement à la notion de jalon. La classe « Token » correspond quant à elle à un jeton, les classes d'unités à considérer comme jetons étant spécifiées auprès du moteur (instance de la classe « Engine »). C'est donc cette dernière classe qui se charge de distinguer les jetons des jalons au sein des annotations du document analysé, conformément aux indications de l'utilisateur de l'API. Elle se charge également de prendre en compte les options de filtrage des jalons, qui permettent de ne prendre en compte qu'un sous-ensemble des annotations présentes dans le document.

Pour rendre compte des annotations présentes dans le document, le moteur demande que lui soit

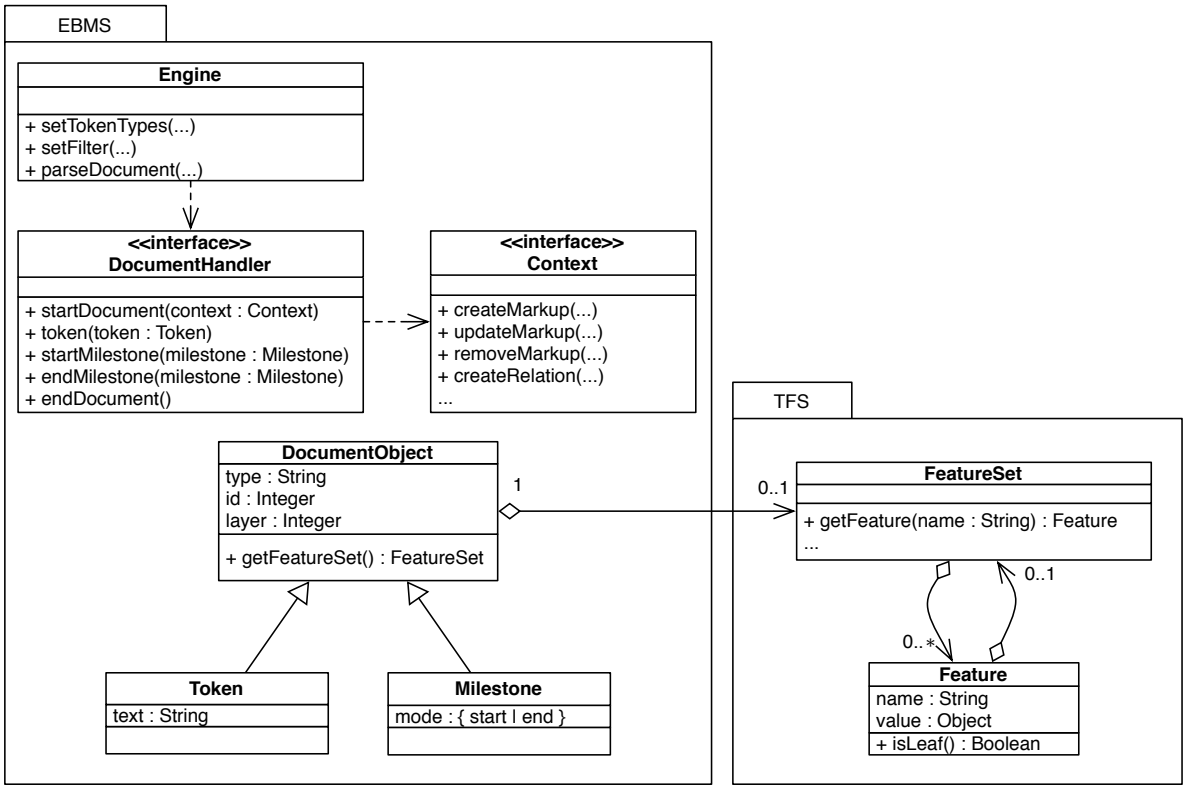


FIG. 11.3 – Extrait de l'API « EBMS ».

fourni un objet implémentant l'interface « DocumentHandler », dont il appellera les méthodes adéquates au fur et à mesure de la lecture du document. Nous mettons ici en oeuvre le *design pattern* qui fonde la méthode d'analyse de documents XML dite SAX[5], et qui repose sur le parcours linéaire du document et l'émission d'événements *via* au fur et à mesure que les éléments recherchés sont reconnus. L'application de ce principe à notre modèle documentaire aboutit au fonctionnement suivant (le terme client désigne les composants logiciels « utilisateurs » de l'API) :

- Le client crée une instance de la classe « Engine », et lui fournit une perspective d'analyse : type(s) d'unités formant les jetons, filtre(s), domaine(s) d'analyse, etc.
- Le client implémente l'interface « DocumentHandler » selon ses besoins (cf. ci-dessous), et fournit une instance au moteur. Cette implémentation peut si besoin utiliser l'instance de l'interface « Context » qui lui est fournie pour créer de nouveaux marquages, annotations et/ou relations.
- Le client lance le moteur en spécifiant les documents d'entrée (document lui-même, et éventuellement annotations et relations) ainsi que les documents de sortie (idem).

Les événements émis par le moteur *via* l'interface « DocumentHandler » sont les suivants :

startDocument() : émis lorsque débute la lecture du document.

token(t : Token) : émis lorsqu'un jeton est rencontré (selon la perspective d'analyse). Dans ce cas les jalons correspondants ne sont pas perçus, mais le texte contenu dans le jeton est accessible, ainsi bien sûr que les annotations et relations associées. Bien-sûr, le jeton n'est émis que s'il appartient au domaine d'analyse et qu'il n'est pas filtré.

startMilestone(m : Milestone) émis lorsqu'un jalon de début est rencontré (si le type correspondant

[5]Simple API for XML, qui est d'ailleurs elle-même utilisée dans l'implémentation d'EBMS.

ne constitue pas un jeton d'après la perspective d'analyse). Seules les annotations et relations associées sont accessibles, puisqu'un jalon ne contient pas de texte. Là encore, le jalon n'est émis que s'il appartient au domaine d'analyse et qu'il n'est pas filtré.

endMilestone(m : Milestone) idem pour un jalon de fin.

endDocument() : émis lorsque la lecture du document s'achève.

On le voit, les éléments (y compris textuels) du document qui ne constituent pas des annotations LinguaStream ne sont pas perçus par le client, qui ne considère que les objets du modèle abstrait. Cela ne convient bien sûr pas dans tous les cas, notamment quand aucune unité secondaire n'a encore été marquée dans le document. C'est par exemple le cas lorsque l'on procède au marquage des unités minimales, typiquement les mots. Dans ce cas, on ne peut plus se reposer sur le modèle documentaire que nous avons défini dans ce chapitre, et il est nécessaire de prendre en compte les éléments « bruts » du document traité. Cela reste rare, puisque la plate-forme repose d'une part sur le principe d'éloignement progressif des formes de surface, et fournit d'autre part des outils permettant de prendre en compte les formes des unités secondaires déjà analysées dès que cela est nécessaire. Une autre API a toutefois été développée pour traiter ce cas de figure, et qui est notamment utilisée par les composants de découpage en mots. Cette API baptisée « Tokenizer-Based Markup system » (TBMS) ne sera pas décrite ici, mais repose sur des principes analogues à ceux d'EBMS, appliqués au niveau des unités primitives.

Un aspect important de ce principe de traitement des documents est qu'à aucun moment une quelconque représentation globale du document n'est calculée. Au contraire, le moteur n'a jamais qu'une vue locale sur le document, et « oublie » à chaque nouvel événement tous ceux qui ont précédé. Ce principe de lecture linéaire et événementielle a pour avantage de n'imposer aucune contrainte en termes de mémoire nécessaire au stockage des informations. En d'autres termes, elle permet de traiter des documents de taille quelconque (voir même infinie dans le cas d'un flux), contrairement aux approches qui procèdent par représentation complète du document en mémoire.

Le « revers de la médaille » est qu'il est de la responsabilité du client de mémoriser, au fur et à mesure de la réception des événements, les informations dont il a besoin pour faire ses propres calculs. Dans le meilleur cas, par exemple s'il se contente d'animer un transducteur déterministe, aucune information ne devra être mémorisée. Mais dans d'autres cas, il sera nécessaire de stocker les portions du document qui doivent subir un traitement global, en prenant si possible les précautions adéquates pour éviter que le volume de ces informations soit fonction de la taille du document, ce qui reviendrait à rendre le traitement limité quant à la taille des données qu'il peut gérer. Pour cela, on procédera généralement à la définition d'un grain dont chaque instance sera mémorisée, pour être traitée globalement avant d'être supprimée de la mémoire. Par exemple, on pourra choisir de mémoriser la totalité du contenu de chaque paragraphe pour lui appliquer un traitement global, mais de ne garder en mémoire qu'un seul paragraphe à la fois. On notera que la plate-forme fournit une extension de EBMS facilitant grandement cette tâche, baptisée « DOM-Based Markup System » (DOM-BMS) et développée par Antoine Widlöcher (Widlöcher, 2006). Cette extension permet, étant donné un domaine d'analyse, de construire une représentation objet partielle du document, segment par segment.

Le lecteur désireux d'obtenir plus d'informations pratiques sur ces APIs et les modalités de leur utilisé pour la création de nouveaux modules d'analyse pourra se reporter au manuel de LinguaStream (Bilhaut *et al.*, 2006) ainsi qu'à la documentation de son API.

Chapitre 12

Modèles d'analyse

Nous avons évoqué à plusieurs reprises dans les chapitres précédents un principe fondateur de LinguaStream visant la combinaison de modèles d'analyse (voir en particulier la section 10.2). Nous allons évoquer dans ce chapitre les différents modèles disponibles en standard dans la plate-forme, ce qui permettra notamment d'apprécier plus concrètement les apports de ce principe. La description de chacun d'entre eux sera très succincte : nous ne nous étendrons ni sur les algorithmes sous-jacents ni sur la syntaxe des formalismes associés. Concernant ce dernier point, le lecteur intéressé pourra toutefois se rapporter au manuel de la plate-forme (Bilhaut *et al.*, 2006). Nous ne donnerons pas non plus d'exemples de règles, le lecteur pouvant se rapporter à la partie II et à l'annexe E pour examiner certaines des règles développées dans le cadre de nos propres travaux.

12.1 Modèles génériques

Nous allons tout d'abord nous attacher aux modèles d'analyse que l'on pourrait qualifier de « génériques » dans le sens où ils sont réellement applicables à tout type de grain (selon la perspective d'analyse adoptée), bien que certains soient *préférentiellement* adaptés à certains niveaux linguistiques.

12.1.1 Grammaires locales d'unification (EDCG)

Ce formalisme est basé sur le langage Prolog et plus particulièrement sur la syntaxe dite DCG (Definite Clause Grammar) qui permet de construire facilement des grammaires d'unification. Cette syntaxe a toutefois été largement agrémentée, en s'appuyant notamment sur la méthode GULP (Covington, 1994) pour représenter et manipuler les structures de traits, de façon à obtenir une syntaxe purement déclarative et masquer les particularités du langage Prolog. Le langage obtenu est appelé EDCG pour « Extended Definite Clause Grammar ». Insistons sur le fait que les grammaires ainsi définies sont uniquement *locales*, l'algorithme associé se chargeant de rechercher au sein du texte les segments qui correspondent à des productions de cette grammaire.

Ce formalisme permet de décrire des grammaires complexes tout en bénéficiant du mécanisme d'unification pour exprimer des contraintes sur les terminaux (jetons ou jalons) et surtout sur les structures de traits qui leur sont associées, tout en construisant de nouvelles structures de traits qui seront associées aux unités détectées. L'écriture de ces grammaires est grandement facilitée par le fait que leur mise en oeuvre s'appuie sur le système de *backtracking* de Prolog. On notera toutefois que le non-déterminisme introduit par cette dernière propriété peut en théorie s'avérer limitatif en termes de performances, même si en pratique une définition adéquate du domaine d'analyse permet de lever facilement cette limitation.

Un autre intérêt de ce formalisme est qu'il donne accès aux utilisateurs qui le souhaitent à l'outillage logique de Prolog, qui pourra intervenir sous forme de contraintes au sein de la grammaire. En pratique, cela s'avère notamment utile pour procéder à des calculs complexes sur les structures sémantiques. Précisons que ces outils sont disponibles aux utilisateurs expérimentés, tout en restant totalement transparents pour ceux qui ne souhaitent pas en faire usage.

Étant basé sur la notion de grammaire, ce formalisme est préférentiellement adapté aux analyses d'ordre syntagmatique ou phrastique. Il est toutefois important de répéter que comme tous les autres modèles décrits dans cette section, son application à tous les autres niveaux de grain est rendue possible par la notion de perspective d'analyse. Ainsi, si le jeton que l'on définira généralement pour une grammaire EDCG est de type « mot », il reste tout à fait possible de décrire des règles portant sur des jetons de type « phrase », « paragraphe » ou autre.

Ce formalisme a notamment été utilisé pour décrire les analyseurs d'expressions spatiales et temporelles dans différentes langues (Beaudet, 2002; Bilhaut *et al.*, 2003a; Tantzen, 2004; Loustau, 2005). Nous l'avons également utilisé pour extraire des syntagmes nominaux, ainsi qu'au niveau discursif pour décrire des combinaisons de cadres de discours.

12.1.2 Macro-expressions régulières (MRE)

Ce formalisme repose sur le principe des expressions rationnelles (séquences de terminaux agrémentés des opérateurs de Kleene (1956)), en appliquant à des unités de grain variable une syntaxe proche de celles des expressions dites « régulières ». Même s'il est beaucoup moins expressif qu'EDCG, ce formalisme a pour avantages d'être plus immédiatement accessible (sa syntaxe est simple et très répandue) et surtout de reposer sur un algorithme déterministe dont l'efficacité ne dépend ni de la taille du document ni de la complexité des règles. Il peut donc être utilisé pour reconnaître des patrons de taille arbitraire.

Plusieurs particularités sont à remarquer en comparaison avec les outils reposant habituellement sur des expressions régulières. En premier lieu, il s'appuie comme tous les autres modèles d'analyse sur le modèle documentaire précédemment décrit. De ce fait, il n'opère pas directement sur des formes de surface mais sur des unités textuelles annotées (jetons ou jalons). Il opère donc sur les structures de traits associées à ces unités[1], et/ou à leur forme de surface quand il s'agit de jetons. Bien-sûr, il bénéficie d'autre part de la notion de perspective d'analyse, ce qui permet de faire varier le grain « jeton » sur lequel porteront les règles du caractère à la section titrée (par exemple).

Une autre particularité de ce système est qu'il permet de spécifier librement la ou les portions de l'expression régulière que l'on souhaite effectivement marquer. Ainsi, si une règle reconnaît un patron « a b c d e », il sera possible de demander à ce que seulement la séquence « b c d » soit effectivement marquée. Il est également possible de demander plusieurs marquages pour une même règle. Bien-sûr, on pourra associer une structure de traits à chacun de ces marquages, qui pourra si nécessaire reprendre des valeurs présentes dans les structures de traits de unités constituant l'unité reconnue. Mentionnons également que plusieurs règles peuvent être appliquées en parallèle, et qu'une règle peut également faire appel à une autre règle (en lui passant des arguments si nécessaire).

En interne, les règles sont transformées en automates déterministes reconnaissant des langages dont l'alphabet est constitué de l'ensemble des structures de traits. On remarquera que la déterminisation de tels automates ne va pas sans poser de problèmes spécifiques, dus au caractère non fini de l'alphabet des structures de traits et surtout au fait que les éléments de cet alphabet ne sont pas disjoints. Ce problème a fait l'objet d'un travail spécifique réalisé en collaboration avec Arnault Soulet, dont on pourra trouver les détails en annexe C.

[1] On notera qu'il s'agit ici de filtrage et non d'unification comme en EDCG.

Le formalisme MRE a été utilisé à des fins très diverses, le plus souvent pour repérer des unités structurellement peu complexes et nécessitant peu de calculs sémantiques, mais fonctionnant à des niveaux de grains variés. Nous l'avons ainsi utilisé tant pour repérer des introducteurs de cadres de discours (cf. section 7) que pour mener une analyse rhétorique très simple dans le contexte du résumé automatique (cf. section 14.2). Il a également été utilisé de façon significative dans les travaux d'Arnault Soulet sur la segmentation discursive (Soulet, 2002) et de Marion Laignelet sur les relations entre titres et cadres temporels (Laignelet, 2003).

12.1.3 Langage de modélisation du discours par contraintes (CDML)

Le formalisme CDML (Constraint-based Discourse Modelling Language), conçu et développé par Antoine Widlöcher (Widlöcher, 2006), permet d'exprimer des contraintes au niveau discursif (même si là encore il ne s'agit que d'un grain de prédilection). Il se veut à la fois descriptif (il permet l'explicitation formelle d'organisations discursives) et prescriptif (l'expression commande un analyseur) et permet l'exploration des organisations discursives par l'expression et la satisfaction de contraintes pouvant être non séquentielles (sans présupposé sur l'ordre des éléments) et non linéaires (la linéarité du texte peut être ignorée). Exprimées à l'aide d'un ensemble de fonctions discursives primitives (présence/absence, cohérence sémantique...), les contraintes peuvent porter en particulier sur les annotations produites en amont de la chaîne de traitement et sur des relations entre ces dernières. Nous ne rentrerons pas ici dans plus de détails, et le lecteur pourra se référer aux publications sus-mentionnées pour obtenir plus d'informations.

Parmi les travaux mentionnés dans ce mémoire, ce formalisme a notamment été utilisé pour l'analyse de la portée de introducteurs de cadres temporels (cf. section 7), à la suite d'une première version non déclarative qui avait été développée avant la naissance de CDML.

12.1.4 Règles de type « système expert » (CLIPS)

Ce modèle d'analyse repose sur le langage CLIPS, qui permet de développer des systèmes à base moteur d'inférence avec une syntaxe Lisp, communément utilisé pour construire des systèmes experts, et ayant la particularité d'être orienté à la fois « objets » et « règles ». Ce système fonctionne à partir d'une base de faits (parfois dite « tableau noir » ou *« blackboard »*) où sont représentées les connaissances du système à un instant donné, faits qui seront unifiés avec les prémisses des certaines règles, ces dernières générant à leur tour de nouveaux faits, etc.

Pour appliquer ce principe au traitement automatique des langues, on peut inscrire sur le tableau noir un ensemble de faits représentant d'une part l'ensemble des unités textuelles pertinentes qui sont présentes dans chaque segment analysé, et d'autre part leurs propriétés (ordre d'apparition, annotations associées, etc.). On peut alors écrire des règles qui seront déclenchées par la présence des faits linguistiques recherchés au sein du segment courant, et générer de nouveaux faits qui définiront *in fine* une nouvelle annotation. Dans LinguaStream, le contenu de la base de faits est le reflet direct de la perspective d'analyse adoptée. Cette dernière définit tout d'abord quels sont les jalons et jetons qui y apparaîtront en compagnie de leurs structures de traits (grain d'analyse et filtrage). Elle définit par ailleurs quels seront les segments qui seront vus comme des groupes d'objets indépendants (domaine d'analyse), les faits linguistiques étant effacés du tableau noir avant chaque nouveau segment.

L'idée d'appliquer ce type de méthodes au traitement des langues est due à Cédric Person, qui l'a exploitée dans ses travaux sur le traitement automatique de la temporalité fondé sur le modèle de Laurent Gosselin (Person, 2004). Son implémentation n'a pas été initialement développée sous LinguaStream, mais a fait l'objet d'une intégration *a posteriori* de façon à la rendre intégrable à d'autres travaux. Elle se matérialise donc maintenant sous la forme d'une chaîne de traitement incluant un

module CLIPS (reposant pour l'heure sur l'implémentation Java de ce langage, nommée Jess).

12.1.5 Règles scriptées

Dans le cas où aucun formalisme proposé par la plate-forme ne s'applique, il reste possible d'utiliser différents langages de script, en s'appuyant sur l'API EBMS (cf. section 11.3. Bien que des compétences minimales en programmation soient alors requises, on notera que l'utilisation de cette API rend la tâche extrêmement aisée, en masquant toute la complexité inhérente à la représentation physique des documents, en donnant « gratuitement » accès à la notion de perspective d'analyse, et en permettant donc de se concentrer sur l'essentiel du travail à réaliser. Les langages aujourd'hui disponibles pour cela dans la plate-forme sont BeanShell[2], Python[3] et Groovy[4].

12.2 Modèles spécifiques

Nous allons dans cette section décrire des modèles d'analyse qui sont plus particulièrement adaptés à un niveau d'analyse particulier, même si certains bénéficient partiellement des possibilités offerte par la notion de perspective d'analyse. Nous ne mentionnerons pas ici les usages qui en ont été faits, dans la mesure où il s'agit d'outils très généraux et utilisés dans des chaînes de traitement très diverses.

12.2.1 Délimitation d'unités minimales par expressions régulières (RT)

Toute chaîne de traitement doit débuter par la définition des unités secondaires minimales, sur la base desquelles les analyses subséquentes seront menées. Il s'agit généralement d'un découpage en mots, mais cela n'est bien sûr qu'une possibilité parmi d'autres, et il est par ailleurs possible de définir un nombre quelconque de classes d'unités minimales.

Outre un module de découpage en mots « standard », la plate-forme fournit un modèle d'analyse dit « Regex Tokenizer » permettant d'écrire des règles spécifiques de découpage du texte à partir d'expressions régulières, opérant cette fois au niveau caractère. Il s'agit d'un formalisme extrêmement simple, où chaque règle est constituée de l'association d'un patron, d'un type identifiant la classe à laquelle appartiendront les unités reconnues, et d'une structure de traits. On pourra ainsi construire un ensemble de règles permettant de reconnaître les différentes chaînes d'unités primitives constituant les unités minimales. On notera qu'il est possible avec ce formalisme d'utiliser des portions des formes de surfaces reconnues au sein de la structure de traits associée aux marquages produits. Il s'agit donc d'un système comparable à l'outil classique « Lex », à la différence qu'aux unités reconnues ne sont pas associées de simples étiquettes mais des structures de traits, et bien sûr que nous traitons ici des fichiers XML et pas seulement du texte brut.

12.2.2 Règles d'étiquetage de jetons basées sur les formes de surface (TM)

Bien que la plate-forme soit orientée vers l'éloignement progressif des formes de surface, il est parfois nécessaire de « redescendre » au niveau des formes au cours d'une chaîne de traitement. La plate-forme offre pour cela le modèle d'analyse dit « Token Marker » qui permet d'écrire des règles portant sur le contenu textuel des jetons afin de leur ajouter de nouveaux traits.

[2] `http://www.beanshell.org`
[3] `http://www.python.org`
[4] `http://groovy.codehaus.org`

Ce système est très proche du précédent dans la mesure où il s'appuie également sur des expressions régulières au niveau caractère. Toutefois, il s'agit ici d'agir sur des unités préalablement délimitées, et constituant des jetons selon la perspective d'analyse adoptée. Les nouveaux traits peuvent, au choix, se substituer aux traits existants ou s'y ajouter.

12.2.3 Projection de lexiques sémantiques (LSL)

Ce modèle d'analyse permet de projeter un lexique sémantique sur un document, c'est-à-dire de procéder au marquage de toutes les unités présentes dans un tel lexique. Le formalisme associé permet de spécifier une liste d'entrées, chacune étant tout d'abord caractérisée par un ensemble de formes. Chaque forme est donnée par un ou plusieurs jetons correspondant généralement à des mots (mais encore une fois il ne s'agit là que d'une possibilité parmi d'autres). Il sera nécessaire pour procéder au marquage proprement dit que le grain « jeton » spécifié dans le modèle d'analyse soit en adéquation avec le grain utilisé dans le lexique. À la suite de la projection, de nouvelles unités secondaires seront créées autour des groupes de jetons reconnus. Chaque entrée du lexique peut par ailleurs être associée à une ou plusieurs structures de traits (distinguées par des symboles appelés « clefs ») qui seront associées aux marquages produits.

Ces lexiques peuvent être directement projetés sur un document à l'aide d'un composant développé par Antoine Widlöcher, ou être directement appelés depuis les règles écrites à l'aide d'autres formalismes comme EDCG ou CDML. Précisons que les lexiques peuvent être stockés soit dans des fichiers XML, soit dans une base de données relationnelle.

12.2.4 Règles d'identification d'unités linguistiques basées sur le balisage XML (XM)

Nous avons déjà mentionné dans le chapitre 11 le fait que les annotations LinguaStream sont indépendantes du balisage XML originel du document, et que seules les premières sont accessibles aux modèles d'analyses. Il est toutefois souvent nécessaire de prendre en compte ce balisage, quand il délimite des objets linguistiques pertinents pour l'analyse que l'on souhaite mener.

La plate-forme fournit pour cela un modèle d'analyse particulier qui permet d'établir une correspondance entre le balisage XML et un marquage LinguaStream. On peut ainsi écrire des règles associant un type particulier de balise XML (identifié par un nom qualifié et éventuellement des contraintes sur ses attributs) à une classe d'unité secondaire. On pourra par exemple choisir de considérer, si le document est au format XHTML, que les balises « p » et « li » délimitent des paragraphes et les balises de type « hX » délimitent des titres.

Les règles ainsi décrites sont bien sûr spécifique à chaque schéma XML particulier. Toutefois, tous les modules d'analyse subséquents seront bien indépendants de ces schémas, puisqu'ils ne se reposeront plus que sur les annotations LinguaStream générées à partir des balises XML. En d'autres termes, ce système permet de dissocier très clairement ce qui relève de la modélisation XML de ce qui relève des objets linguistiques.

12.3 Intégration de systèmes d'analyse externes

Outre les systèmes précédents qui constituent des modèles d'analyse en tant que tels, la plate-forme permet d'exploiter les résultats produits par des logiciels tiers au sein d'une chaîne de traitement. Précisons que ces logiciels ne sont pas fournis avec la plate-forme, et doivent être obtenus séparément, en acquérant le cas échéant une licence d'utilisation.

12.3.1 L'étiqueteur morpho-syntaxique « TreeTagger »

Le logiciel « TreeTagger » (Schmidt, 1994) est un étiqueteur morpho-syntaxique et lemmatiseur multilingue très répandu, fonctionnant par apprentissage sous forme d'arbre de décision. S'il est installé sur la même machine, LinguaStream permet de l'utiliser sous la forme d'un composant habituel au sein d'une chaîne de traitement. Dans ce cas, il agit en ajoutant des traits sur un marquage des mots qui devra être réalisée en amont.

Les traits ajoutés correspondent à l'étiquette syntaxique ainsi qu'au lemme. Le composant permet soit de conserver les traits tels qu'ils sont produits par le TreeTagger, soit de les transformer en étiquettes conformes à un catalogue spécifique à la plate-forme (uniquement pour le français). Dans ce dernier cas, on s'assure que la suite de la chaîne de traitement est indépendante des étiquettes particulières du TreeTagger, afin de pouvoir lui substituer un autre étiqueteur si nécessaire.

Pour l'utilisateur, il s'agit d'un composant quelconque, et le fait que soit fait appel à un logiciel externe est totalement transparent. Les propriétés de ce dernier permettent d'ajuster le comportement de l'étiqueteur, notamment en sélectionnant les ressources adaptées à la langue traitée. Le TreeTagger étant capable de prendre en compte un pré-étiquetage des mots, la plate-forme assure le transfert d'un éventuel étiquetage déjà présent sur les mots (que l'on pourra par exemple effectuer lors du découpage en mots (règles RT) ou *a posteriori* (règles TM).

12.3.2 L'analyseur syntaxique « Syntex »

L'analyseur syntaxique de corpus Syntex (Bourigault et Fabre, 2000) permet notamment d'extraire automatiquement d'un corpus une liste de noms et syntagmes nominaux, structurée par des relations de dépendance syntaxique. Le résultat de l'analyse se présente sous la forme d'un réseau de dépendance, dans lequel chaque syntagme extrait est relié à sa tête et à son expansion syntaxiques.

Il peut être utilisé dans LinguaStream en tant qu'analyseur syntaxique, à l'aide d'un module développé par Jérôme Fleury (Fleury, 2006), qui permet de restituer l'analyse d'un texte produite par Syntex sous la forme d'un marquage des mots et des syntagmes, et surtout de structures de traits reflétant les étiquettes syntaxiques et de relations entre ces unités. Précisons que l'utilisation de ce module requiert l'acquisition d'une licence d'utilisation du logiciel Syntex auprès de ses propriétaires.

Chapitre 13

L'environnement d'expérimentation intégré

L'environnement d'expérimentation constitue la partie visible de la plate-forme, sous la forme d'une interface graphique où peuvent être réalisées la totalité des tâches liées à la constitution d'un dispositif expérimental complet. Nous nous basons ainsi sur le même principe que les environnements de développement intégrés (IDE), afin d'offrir une interface cohérente et uniforme.

Au risque de donner à ce chapitre un aspect « catalogue », l'environnement sera ici présenté à partir d'un ensemble de copies d'écran, qui nous semblent les plus à même de rendre compte des diverses fonctionnalités disponibles. Là encore, le lecteur intéressé pourra bien sûr se reporter au manuel d'utilisation pour obtenir plus de détails pratiques.

13.1 Vue d'ensemble

L'environnement est scindé en trois zones principales : une zone d'édition, où peuvent être manipulés les différents objets prenant part à une chaîne de traitement ; une zone utilitaire où se trouve la palette de composants, l'explorateur de fichiers et l'éditeur de propriétés ; une zone où la plate-forme affiche des messages relatifs à son fonctionnement. Dans la zone d'édition prendront place une certaine variété d'éditeurs plus ou moins spécifiques, comme le montre la figure 13.1. On pourra y manipuler aussi bien des chaînes de traitement que des règles sous différents formalismes, des paramètres de visualisation ou encore des lexiques sémantiques.

L'éditeur principal est bien sûr celui qui permet de manipuler les chaînes de traitement elles-mêmes. Comme on le voit sur la figure 13.2, ces chaînes sont représentées graphiquement sous formes de boîtes reliées par des connecteurs qui figurent les flux de données. Les composants à ajouter dans une chaîne sont choisis dans une « palette », où ils sont organisés sous une forme arborescente (visible dans la partie gauche de la figure précédente).

Tous les fichiers manipulés dans la plate-forme sont stockés au format XML, mais sont également associés à des interfaces spécifiques au sein de l'environnement. Par exemple, l'édition de lexiques sémantiques peut se faire à l'aide d'un éditeur spécifique illustré en figure 13.3.

13.2 Outils de visualisation

La plate-forme offre de multiples outils permettant de procéder à la visualisation des résultats produits par une chaîne de traitement. Le mode de visualisation le plus communément utilisé consiste à

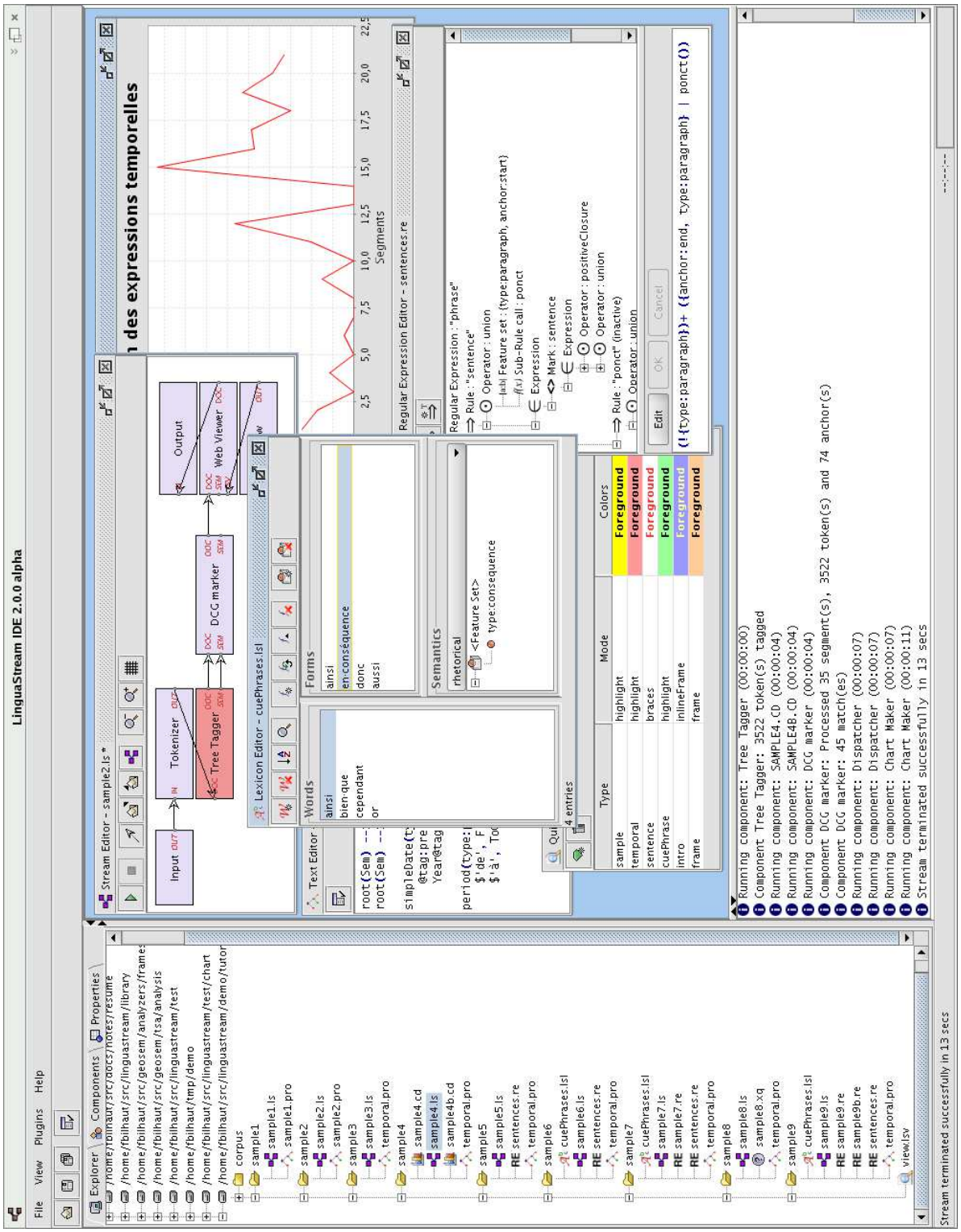


FIG. 13.1 – Différents éditeurs dans l'environnement intégré.

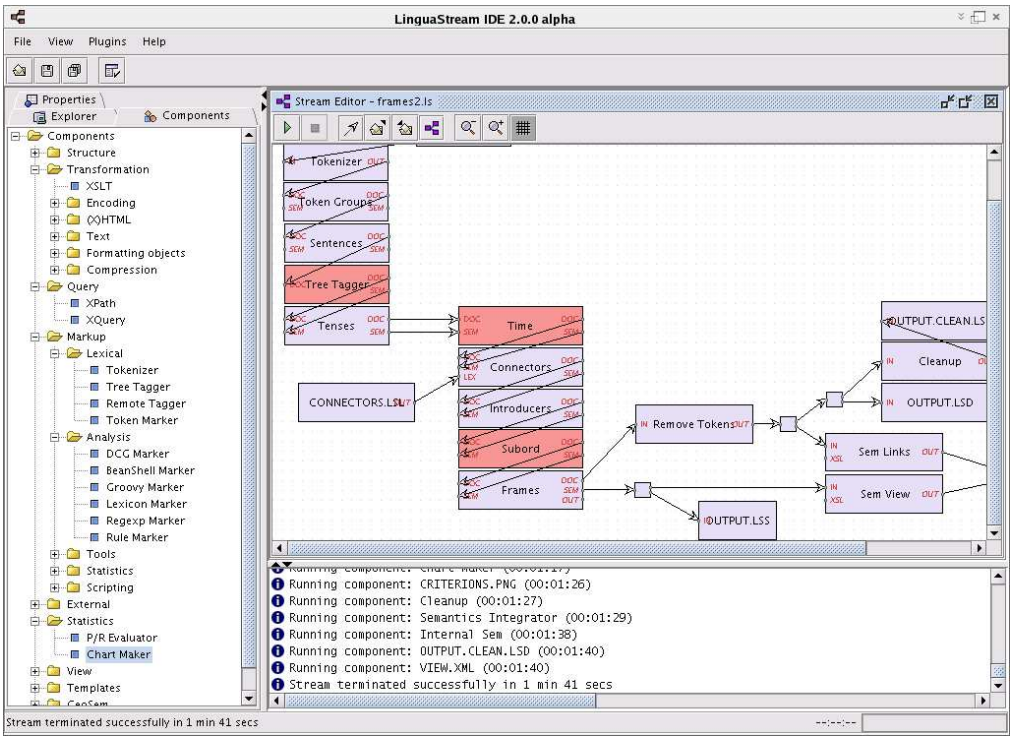


FIG. 13.2 – L'éditeur de chaîne de traitement.

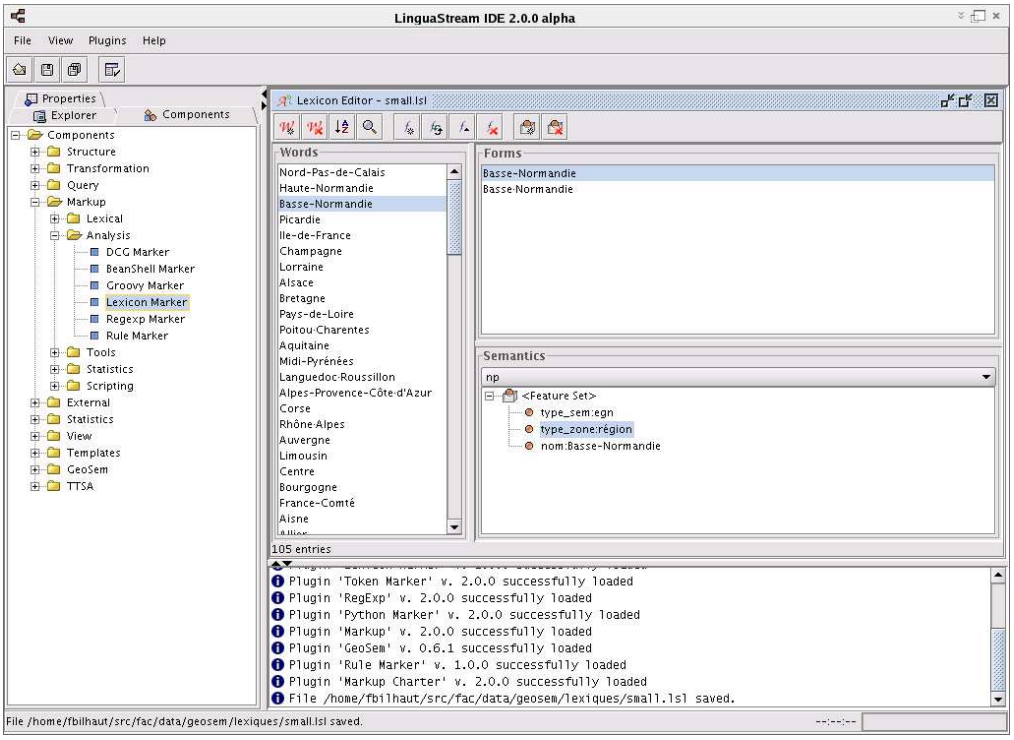


FIG. 13.3 – L'éditeur de lexiques sémantiques.

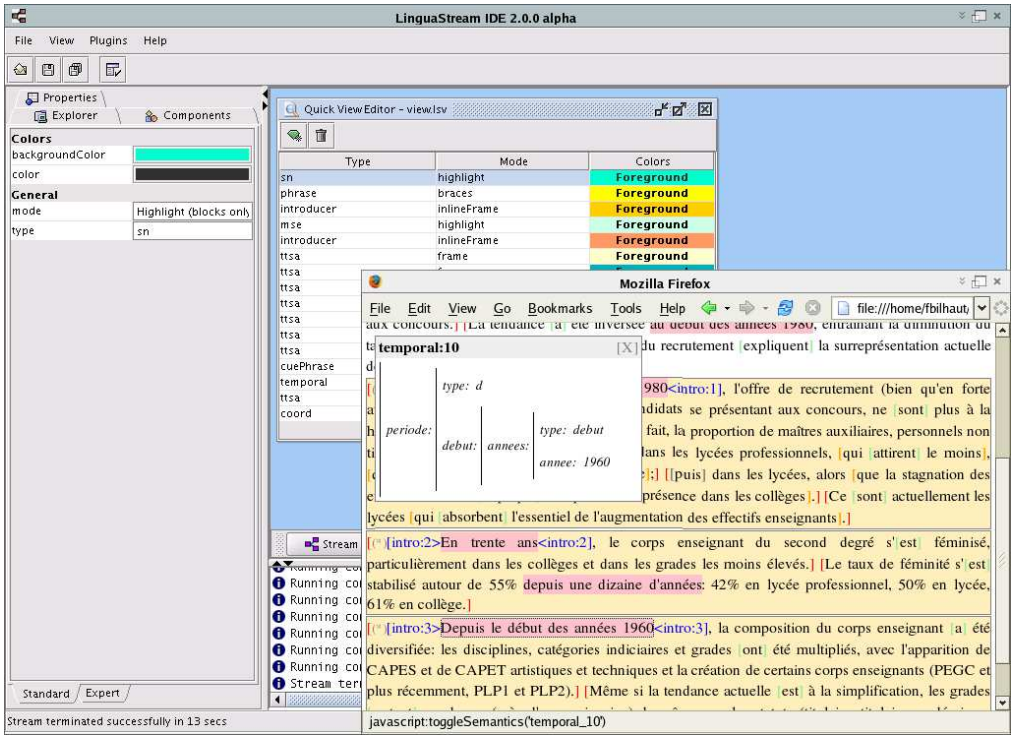


FIG. 13.4 – Visualisation des annotations au sein du document analysé.

rendre visibles les marquages et annotations au sein du document lui-même. La plate-forme étant totalement indépendante de tout format XML particulier, il est toutefois nécessaire que le fichier traité soit visible en tant que tel dans un navigateur ou autre (fichier XHTML, XML+CSS, etc.). La plate-forme se chargera d'ajouter les modalités de visualisation des marquages et annotations, qui apparaîtront donc au sein du document lui-même.

La figure 13.4 illustre ce procédé. On voit dans l'interface de plate-forme l'éditeur appelé « QuickView » qui permet de déterminer les modalités d'affichage de chaque marquage, sous la forme d'un « descripteur de vue ». On voit au premier plan de cette même figure un exemple décoré tel qu'il apparaît dans un navigateur, au sein duquel un clic sur un segment annoté permet de faire apparaître la structure de traits associée. Il est également possible de masquer l'ensemble des annotations LinguaStream de façon à visualiser le document dans son état initial.

Insistons sur le fait que ce procédé permet de préserver à l'identique la mise en forme initiale du document. On peut par exemple observer sur la figure 13.5 un document XHTML après annotation.

Outre la visualisation au sein même du document, la plate-forme propose d'autres modes qui permettent, cette fois sans préserver la forme initiale du document, de disposer d'une vue spécifique. Il s'agit en fait de modules intermédiaires qui procèdent à une transformation du document pour en produire une vue particulière, celle-ci pouvant finalement être décorée par le système QuickView précédemment mentionné (le même paramétrage peut donc être utilisé quel que soit le mode de visualisation adopté).

Un premier mode de visualisation reprend la vue classique de type « concordancier », comme le montre la figure 13.6. Cette vue permet d'observer les contextes gauches et droits d'un ou plusieurs types d'expressions analysées. On remarquera qu'il ne s'agit que d'une vue : toutes les contraintes portant sur le type d'unité visualisées relèvent ici de la chaîne de traitement amont. On remarquera toutefois qu'un concordancier plus élaboré est en cours de développement à partir du travail réalisé

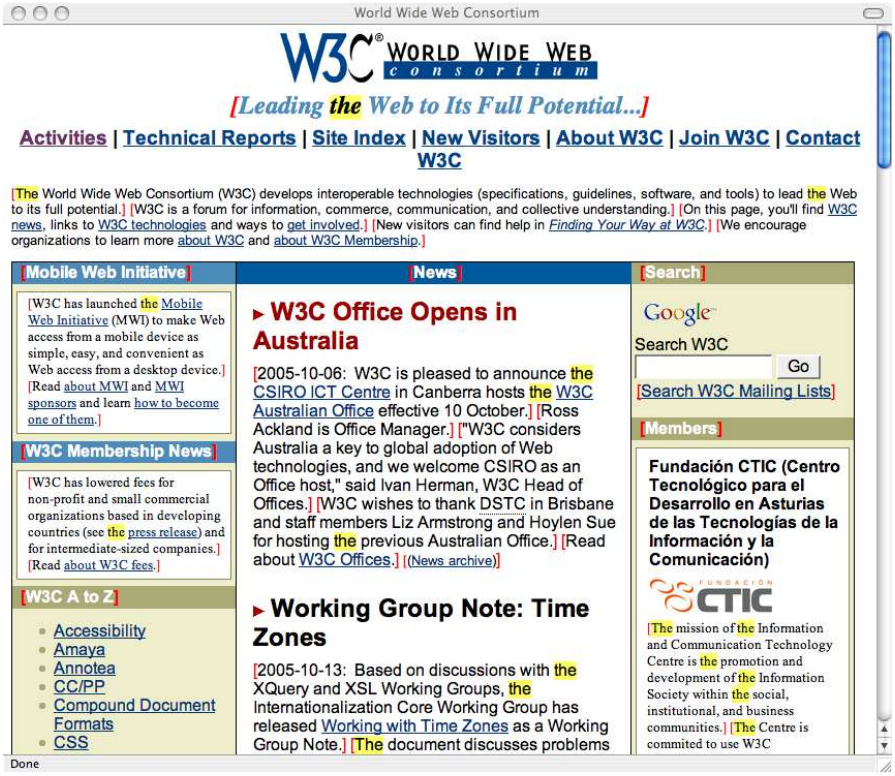


FIG. 13.5 – Préservation de la mise en forme du document original (ici un document XHTML). Le marquage des analyses produites par LinguaStream (ici un exemple « jouet » délimitant les phrases et reconnaissant le mot « the ») sont visibles sous la forme de crochets et de surlignages.

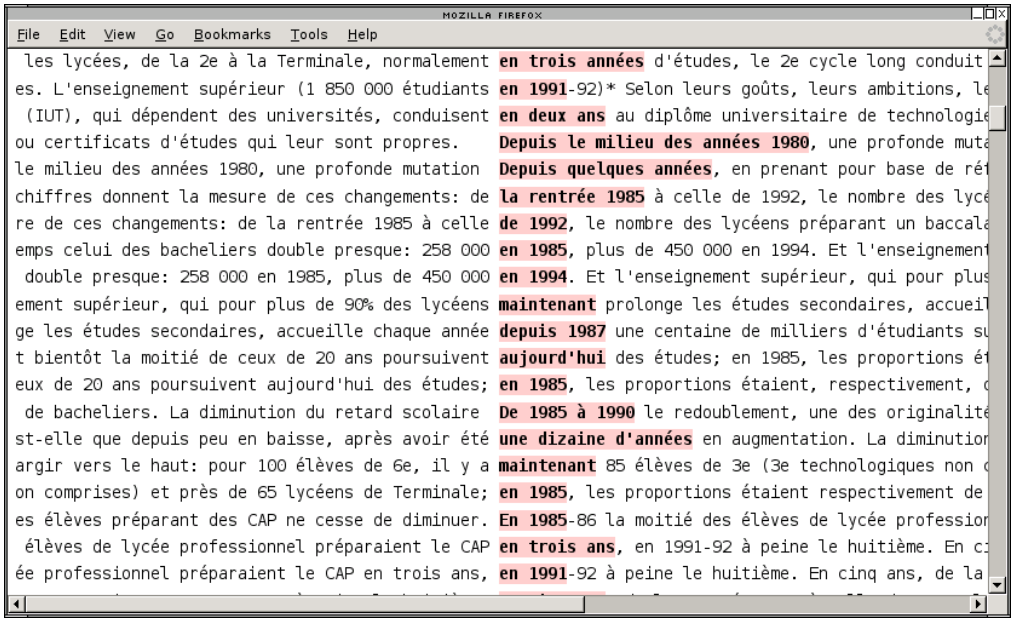


FIG. 13.6 – Visualisation des annotations type « concordancier ».

par Thomas Courdille, qui permettra de procéder dynamiquement à des tris et au filtrage sur la forme et/ou les structures de traits.

Un autre mode de visualisation produit une vue dite « macro-concordancier », qui offre une vue sélective adaptée à des objets de grain plus important (objets discursifs notamment). La figure 13.7 montre un exemple de vue générée par cet outil à partir d'un document où des cadres temporels ont été annotés à la fois manuellement et automatiquement (en plus d'un ensemble d'annotations intermédiaires dans ce dernier cas, cf. section 7). Il est évident que les objets de type « cadre » ne peuvent être affichés dans une vue concordancier standard, mais il est toutefois intéressant de pouvoir les observer isolément, tout en bénéficiant d'une vue synthétique sur leur contenu et/ou leur contexte. Par exemple, on pourra s'intéresser spécifiquement à certaines unités qui les composent, en l'occurrence les expressions temporelles, les connecteurs de discours ou les verbes, qui ont une importance particulière dans le calcul de la portée. La vue « macro-concordancier » permet cela, à partir d'un paramétrage qui comprend le ou les types d'unités que l'on souhaite observer, ainsi que le ou les types des unités internes que l'on souhaite voir apparaître, le reste du texte étant seulement figuré par des points de suspension. On notera que comme avec un concordancier habituel, le même texte peut apparaître plusieurs fois s'il prend part au contexte de plusieurs unités observées, ce qui se produit dans notre exemple quand le même cadre a été annoté plusieurs fois (manuellement et automatiquement).

Un autre mode de visualisation permet d'observer les relations et les structures de traits associées, comme le montre la figure 13.8, à l'aide d'un composant développé par Antoine Widlöcher.

On notera enfin que l'ensemble de la plate-forme s'appuyant sur Unicode, elle permet de traiter et visualiser des documents dans la plupart des langues. La figure 13.9 montre un exemple de document en Chinois simplifié où ont été repérées certaines expressions temporelles à partir d'une grammaire EDCG (cette grammaire contient donc des idéogrammes chinois en guise de terminaux).

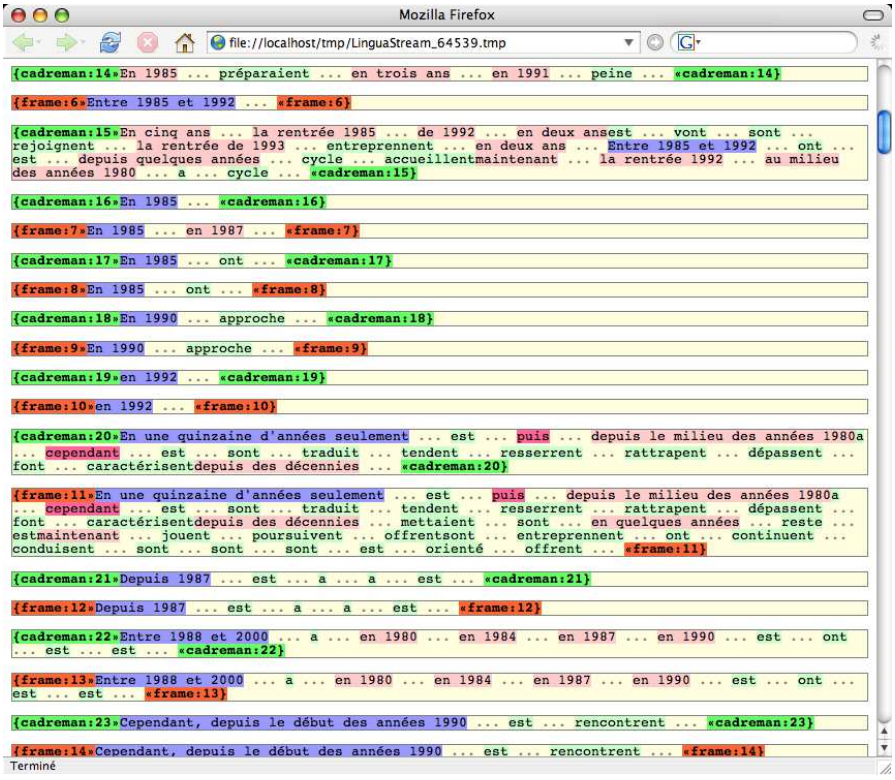


FIG. 13.7 – Visualisation des annotations type « macro-concordancier ».

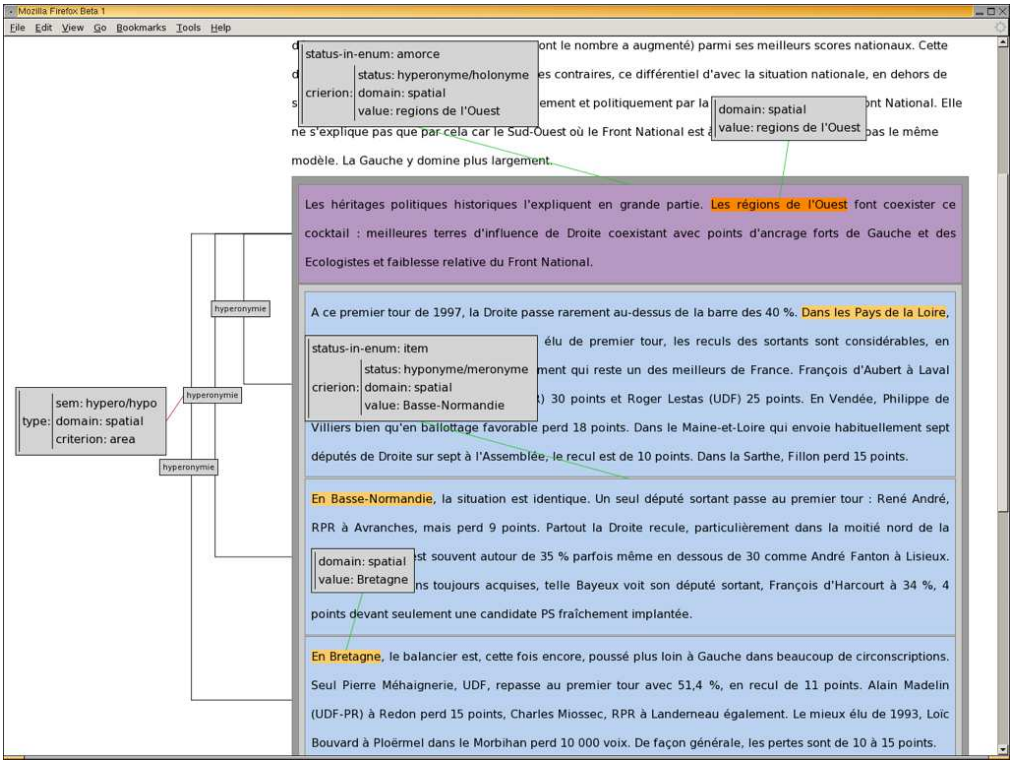


FIG. 13.8 – Visualisation des relations (module développé par Antoine Widlöcher).

FIG. 13.9 – Visualisation des annotations dans un document en Chinois.

13.3 Autres outils d'expérimentation

Outre les vues « textuelles » mentionnées dans la section précédente, la plate-forme permet de générer différents types de graphiques relatifs aux annotations présentes dans un document. Une interface spécifique permet de créer des « descripteurs de graphiques » qui seront fournis au générateur de graphiques, ce dernier étant là encore accessible sous la forme d'un composant. Une fois créé, un même descripteur de graphique pourra bien sûr s'appliquer à différents documents, au même titre qu'un descripteur de vue. Ce système permet de différents types de graphiques : proportions relatives de marquages possédant des propriétés données (par filtrage sur les structures de traits), quantités de marquages par segment (le grain étant paramétrable), graphe représentant une valeur numérique présente sur un trait donné, etc.

La figure 13.10 montre une session de l'environnement où est édité un descripteur correspondant au graphique circulaire qui représente les proportions relatives des différents critères de ruptures qui sont intervenus dans un analyse de cadres temporels. On voit également en arrière-plan un graphique représentant les quantités d'expressions temporelles et de cadres par segment au sein du même document.

Parmi les outils susceptibles d'intervenir dans la création de systèmes expérimentaux, il convient également de mentionner ceux qui permettent de manipuler les chaînes de traitement elles-mêmes. Il s'agit tout d'abord d'un module permettant d'appliquer une chaîne de traitement à un jeu de documents, et de collecter les résultats et éventuelles erreurs. Ce système est accessible via une interface représentée dans la figure 13.11.

Un autre système permet quant à lui d'automatiser différentes tâches réalisables par la plate-forme sous la forme de scripts. Un tel script pourra par exemple lancer une requête sur Google en fonction de mots-clefs demandés à l'utilisateur, puis appliquer une chaîne de traitement à chacun des documents retournés (un script de ce type est reproduit en annexe E.3).

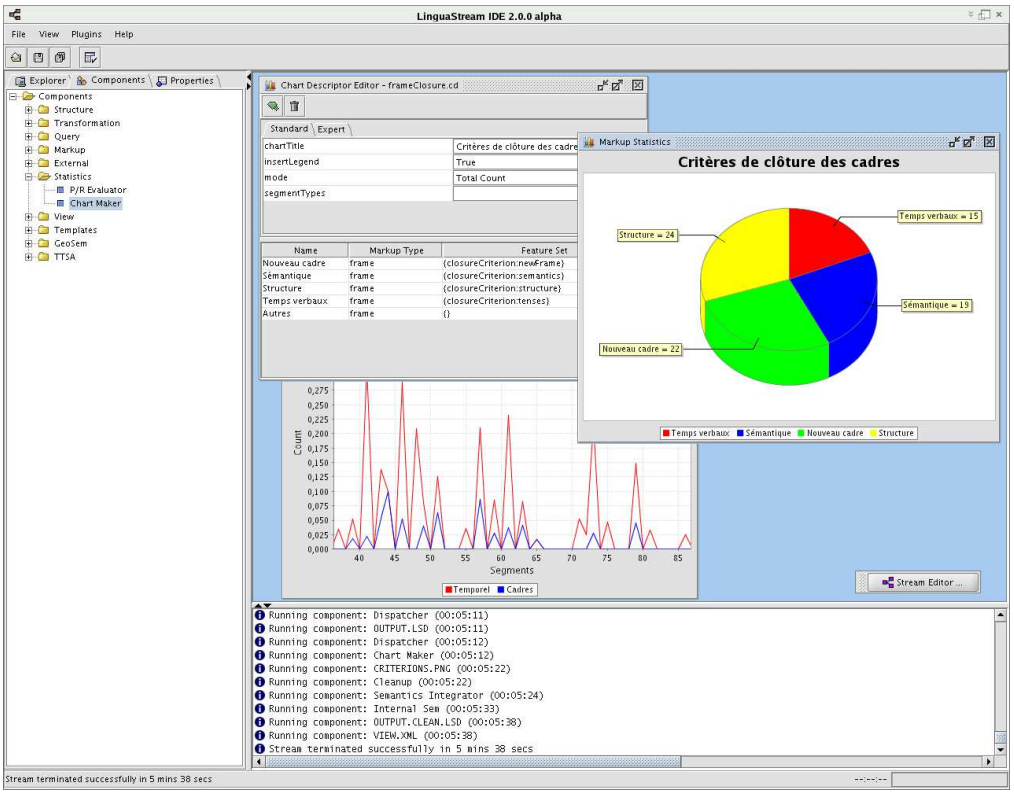


FIG. 13.10 – Visualisation des résultats sous forme de graphes.

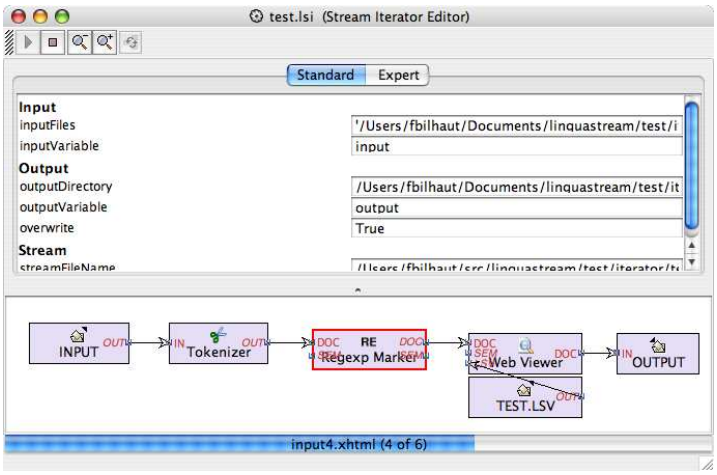


FIG. 13.11 – Interface de traitement par lots.

Chapitre 14

Conclusion

14.1 De l'outil aux instruments

Comme on aura pu s'en rendre compte à la lecture de l'ensemble de ce mémoire, LinguaStream peut être envisagée sous différents angles. Il s'agit en premier lieu d'un *laboratoire virtuel* où peuvent être mises en place rapidement des expériences qui demanderaient, sans outil adapté, un temps de développement considérable. Elle favorise également la reproductibilité de ces expériences, qui sont intégralement décrites sous forme déclarative, ainsi que la réutilisabilité de tout ou partie des procédés ainsi élaborés. Elle ne prétend évidemment pas apporter de solution immédiate à toute la diversité des problèmes que l'on peut rencontrer en TAL, mais on peut toutefois considérer qu'elle offre un cadre général ou pourraient venir se greffer bon nombre des formalismes et des composants qui pourraient se révéler nécessaires dans les situations ou les modules existants de suffisent pas.

Sous un autre angle, LinguaStream peut être vue comme une *plate-forme logicielle* permettant la mise en oeuvre effective de procédés de traitement des langues. En effet, elle offre plusieurs moyens de rendre disponibles, en tant que composants logiciels, les chaînes de traitement qu'elle permis d'élaborer. Les chaînes peuvent en effet être rendues paramétrables, et former des macro-composants susceptibles de prendre part à de nouvelles chaînes. Elles peuvent également être exécutées en ligne de commande, à partir d'une API Java ou même sous forme de services Web. D'autre part, la plate-forme est capable de traiter tous les types de documents XML, génère des annotations facilement exploitables par des outils tiers, et fournit des outils permettant de les exporter sous des formats standards tels que RDF. On notera à ce sujet que la plate-forme n'a pas pour autant été conçue pour traiter rapidement des volumes de données conséquents, et que son application à très grande échelle nécessiterait des développement spécifiques. Elle reste toutefois parfaitement utilisable pour mettre en place des prototypes fonctionnels, dont nous avons vu un exemple dans le chapitre 6 portant sur la réalisation d'un moteur de recherche complet.

Considérant ces deux facettes, et selon la terminologie proposée par Benoît Habert (2005), on remarquera donc que LinguaStream peut être vue comme un *outil* facilitant la mise en place de *dispositifs expérimentaux*, ceux-ci pouvant éventuellement devenir des *instruments* disponibles soit pour répondre à de nouveaux besoins expérimentaux, soit pour une utilisation applicative.

Enfin, si de très nombreux développements futurs sont bien sûr à envisager, on peut considérer que dans le cadre de nos propres travaux, LinguaStream a atteint son principal objectif, qui consistait paradoxalement à se révéler *aussi peu indispensable* que possible. Nous souhaitions en effet nous donner les moyens d'élaborer des procédés d'analyse des langues en leur garantissant une certaine pérennité vis à vis de leur réalisation logicielle, ce qui est bien le cas ici puisque même si LinguaStream venait à disparaître « du jour au lendemain », cela n'enlèverait rien à la transparence des chaînes de

traitement qu'elle aura servi à élaborer ni à l'accessibilité du savoir « computo-linguistique » qu'elles recèlent.

14.2 Autres cas d'utilisation

Comme on peut s'en douter, LinguaStream a avant-tout été développée pour répondre à nos propres besoins et à ceux de l'équipe. Mais elle a également été amenée à rendre d'autres services, à des fins de recherche ou d'enseignement, au sein même du laboratoire mais aussi à l'extérieur. Voici une liste présentant très succinctement les différentes utilisations qui en ont été faites au delà des travaux mentionnés dans ce mémoire :

- Antoine Widlöcher (GREYC), outre sa participation au développement de la plate-forme elle-même et l'élaboration du formalisme CDML (cf. section 12.1.3), utilise LinguaStream dans le cadre de ses travaux de thèse sur l'analyse de la structure rhétorique du discours (Widlöcher, 2004).
- Marion Laignelet (ERSS / société Initiales) exploite également la plate-forme dans le cadre de ses travaux sur l'étude lingustique des relations titres / cadres (Laignelet, 2003) et l'analyse automatique de segments d'information évolutive (Laignelet, 2006).
- La plate-forme est également utilisée au LIUPPA autour de la question de l'analyse de documents territorialisés (Etcheverry *et al.*, 2005), dans le cadre du projet GeoSem. Ces travaux mènent par ailleurs à de nouveaux développements au sein de la plate-forme elle-même, notamment autour des services Web (Lesbegueries, 2005).
- Un travail a également été entamé par Cédric Person (GREYC), visant l'intégration à LinguaStream de ses travaux sur le traitement automatique du temps et de l'aspect (Person, 2004), de façon à permettre l'exploitation aisée de ses résultats au sein de l'équipe.
- Elle est aussi mise en oeuvre au GREYC dans le cadre d'un projet TCAN intitulé « Intervalles temporels et applications à la linguistique textuelle » (action spécifique du département STIC du CNRS).
- Une collaboration a été entamée avec des membres de l'équipe DoDoLa du GREYC travaillant sur la fouille de données (et notamment avec François Rioult), où la plate-forme est exploitée pour expérimenter la collaboration de méthodes de type « fouille de texte » avec des méthodes « linguistiques ». Différents composants ont déjà été développés pour intégrer dans une chaîne de traitement des procédés principalement basés sur l'étude des motifs fréquents, et d'ores et déjà utilisés à des fins pédagogiques. Une expérimentation concrète portant sur des méthodes hybrides a été initiée dans le cadre du défi fouille de textes (DEFT) 2006.
- Sur le plan pédagogique, la plate-forme est utilisée auprès des étudiants des universités de Caen (informatique) et de Toulouse 2 (sciences du langage), que ce soit pour l'observation de corpus (principalement pour les linguistes) ou le traitement automatique des langues (principalement pour les informaticiens).
- Différents projets de Master ont également été conduits au GREYC autour du développement de LinguaStream, portant notamment sur les développements initiaux du modèle d'analyse MRE (Soulet, 2002), l'intégration de l'analyseur Syntex (Fleury, 2006), la réalisation d'un module « concordancier » (Courdille, 2005), ou encore la réalisation du moteur de recherche conçu dans le cadre de GeoSem (Benallel, 2005). D'autres projets ont également fait usage de la plate-forme en tant qu'outil, portant notamment sur l'analyse des expressions spatiales et temporelles (Beaudet, 2002; Tantzen, 2004).

14.3 Processus de développement

Rassurons immédiatement le lecteur qui aura eu la patience de nous lire jusque là : nous ne nous lancerons pas ici dans la description des détails d'implémentation de LinguaStream. Il s'agit seulement de donner quelques détails au sujet de l'effort de développement, non négligeable, impliqué par cette entreprise.

L'essentiel de ce développement s'est fait en s'appuyant sur la plate-forme Java, qui garantit son fonctionnement sur tous les systèmes d'exploitation disposant le machine virtuelle adéquate, c'est-à-dire la grande majorité d'entre eux (y compris Windows, MacOS et Linux). Hormis les développement relatifs à ses applications (GeoSem, nos propres travaux, etc.), l'implémentation de LinguaStream implique près de 2 000 classes[1], dont l'organisation ne peut bien sûr être décrite en détail ici. Outre Java, on notera que la plate-forme met en oeuvre des ponts avec différents autres langages, notamment Prolog, mais aussi Python, CLIPS ou encore BeanShell. On remarquera par ailleurs que nous avons fait appel à de nombreuses bibliothèques *open-source*[2], que nous n'énumérerons pas ici, mais qui nous ont fait bénéficier d'un gain de temps considérable.

Le plus grand soin a été apporté à la conception de l'architecture de LinguaStream, qui met en oeuvre bon nombre de *design patterns* et adopte partout où cela est possible une approche modulaire. On notera en particulier que l'architecture JavaBeans a été employée comme ossature du système de composants de la plate-forme, et qu'un système de *plugins* garantit une isolation logicielle complète entre le noyau de la plate-forme et ses fonctionnalités diverses, pour la plupart réalisées par des modules externes. Outre le fait de faciliter notre propre processus de développement, ce dernier point permet à des utilisateurs tiers de développer et de distribuer leur propres modules sans avoir à disposer des sources de la plate-forme elle-même. Globalement, l'architecture logicielle ainsi développée peut aujourd'hui être considérée comme robuste, puisqu'elle est restée stable durant près de cinq années de développement relativement soutenu.

Hormis les nombreux contributeurs déjà mentionnés dans la section précédente, le développement de la plate-forme implique aujourd'hui une équipe formée d'une part par Antoine Widlöcher, nouvellement Nicolas Hernandez, et nous-même au GREYC, et d'autre part par Marion Laignelet et Christophe Pimm à l'ERSS. Sur le plan pratique, l'équipe s'organise autour d'un système de gestion de configuration pour ce qui est des sources, ainsi que (depuis peu) d'un Wiki[3] dédié notamment au support utilisateur et au partage de ressources.

La plate-forme est disponible sur le site Web qui lui est consacré[4], et utilisable gratuitement à des fins personnelles ou de recherche. Nous espérons avoir à moyen terme l'opportunité d'en faire un logiciel *open-source*.

[1] Réparties sur environ 100 000 lignes de code.

[2] Publiées sous la licence LGPL.

[3] Site web dynamique dont tout visiteur peut modifier les pages à volonté. Il permet non seulement de communiquer et diffuser des informations rapidement, mais aussi de structurer cette information pour permettre d'y naviguer commodément.

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

Conclusion

Conclusion

Nous avons décrit au fil de ce mémoire une approche de l'analyse thématique visant à définir *a priori* les propriétés de l'objet « thème » auquel elle s'attache. Nous nous sommes par suite fondé sur la modélisation d'un certain nombre de phénomènes linguistiques qui y sont corrélés pour définir des procédés d'analyse automatique. Nous nous sommes également donné les moyens, en développant LinguaStream, de préserver la transparence des modèles opératoires qui en découlent, tout en proposant des outils méthodologiques applicables beaucoup plus généralement en TAL.

Nous avons ainsi été amené à nous concentrer sur quelques phénomènes thématiques bien particuliers, dont l'analyse automatique ne saurait bien sûr rivaliser, en termes de généralité ou de légèreté des ressources, avec d'autres méthodes déjà proposées dans ce domaine. La démarche nous semble toutefois salutaire dans un contexte où ces dernières semblent souvent se soustraire à la définition précise des phénomènes linguistiques auxquels elles s'attachent. Elle a en outre permis de faire apparaître différentes perspectives de recherche que nous allons maintenant évoquer, dont certaines vont dans le sens d'une généralisation qui pourrait permettre, à long terme, d'aboutir à une couverture plus large de la structure thématique des textes.

***

Un premier volet concerne l'analyse des cadres de discours, et tout d'abord celle des cadres temporels (chapitre 7). Comme nous l'avons vu, nous nous sommes dotés sur ce point d'outils d'évaluation qui devraient nous permettre de faire évoluer significativement notre méthode de détermination de la portée des introducteurs temporels. Plusieurs hypothèses assez précises pourront être explorées à cette fin.

i) Nous avons observé, au cours d'une première évaluation, une tendance forte de l'analyseur à calculer des portées trop grandes. Si elle est confirmée par les prochaines évaluations, elle pourra nous conduire à accroître de la rigidité des critères existants, ainsi qu'à rechercher de nouveaux critères. Ces derniers pourront notamment résulter d'une étude sur corpus doublement annotés (manuellement et automatiquement), à l'aide des outils offerts par LinguaStream. Nous espérons ainsi observer un certain nombre de régularités susceptibles de fonder de nouveaux indices de clôture.

ii) Afin de perfectionner les indices existants, nous souhaitons également considérer les bénéfices que pourrait apporter une analyse aspecto-temporelle des verbes, en lieu et place de l'étiquetage morpho-syntaxique, sur ce point très limité, dont nous disposons actuellement. Nous pourrons pour cela exploiter les résultats de Cédric Person (2004), déjà mentionnés à plusieurs reprises, et qui seront bientôt disponibles sous la forme de composants LinguaStream. Dans le même ordre d'idées, l'analyse des chaînes de référence pourra également être intéressante dans la contexte de l'analyse de la portée, par exemple en nous fondant sur les travaux de Mai Ho-Dac (2006) portant sur la relation entre encadrement et continuité référentielle.

iii) Un autre point concerne, au contraire du point (i), l'assouplissement des critères dans certaines positions caractéristiques. On remarquera notamment que la plupart des indices que nous exploitons

aujourd'hui devraient être localement inactivés dans des situations relevant d'un « échappement » temporaire de l'univers courant (propositions relatives, incises, comparatifs, parenthèses, etc.). Compte tenu du défaut sus-mentionné de l'analyseur, cette piste n'est certes pas prioritaire, mais la réalité du phénomène qu'il s'agirait de prendre en compte paraissant avérée, elle nous permettrait vraisemblablement d'améliorer notre modèle opératoire, au moins sur le plan théorique.

iv) Nous souhaitons également nous intéresser à l'étude sur corpus de l'incidence de la forme d'un introducteur sur son propre pouvoir cadratif. Les analyses en amont dont nous disposons, associées aux outils offerts par LinguaStream, nous permettront de mener assez facilement un certain nombre d'études visant à établir (ou infirmer) des corrélations entre l'ampleur de la portée d'un introducteur d'une part, et d'autre part des propriétés telles que la nature de la plage temporelle qu'il définit, le fait qu'il soit associé ou non à un connecteur de discours, ou encore sa position dans le paragraphe. Nous pourrions ainsi disposer de critères portant sur l'introducteur lui-même quant à la dimension de sa portée (complétant bien sûr les autres indices rencontrés au fil du texte).

v) Nous pourrons, à plus long terme, poursuivre ces travaux sur l'analyse automatique des cadres en nous attachant plus attentivement aux cadres spatiaux (chapitre 6), pour lesquels notre implémentation n'est que très partielle. Nous pourrons pour cela nous appuyer sur les résultats de Yann Mathet et Thierry Charnois concernant le *matching* d'expressions spatiales, qui fait appel à un système d'information géographique en corrélation avec un processus de « raisonnement » spécifiquement adapté à la sémantique spatiale telle qu'elle est traitée au sein de l'équipe. Nous pourrons d'autre part évaluer l'applicabilité aux cadres spatiaux des critères dont nous disposons déjà sur le plan temporel.

***

Un second ensemble de perspectives est lié à la notion de thème composite développée dans le chapitre 8. Une première étape consistera à poursuivre notre étude, pour l'heure très exploratoire, de la notion d'axe sémantique. Il s'agit d'une part de s'attacher au phénomène sémantique en tant que tel, mais aussi de persévérer dans la voie consistant à prendre en considération les interactions entre discours et structure des connaissances d'un domaine, et par suite les problèmes touchant à la représentation de ces connaissances ainsi qu'à leur constitution. Concernant l'analyse thématique proprement dite, il s'agira de poursuivre notre étude des différentes configurations discursives liées au phénomène de thématisation composite, et de considérer le plus généralement possible les possibilités d'analyser automatiquement les phénomènes de portée qu'elles impliquent. Là encore, plusieurs voies pourront être considérées.

i) Tenant une place de choix parmi les configurations que nous venons d'évoquer, nous retrouvons bien sûr le modèle de l'encadrement du discours, et en particulier un troisième type de cadre que nous avons pu observer à de multiples reprises dans ce mémoire, et que l'on pourrait qualifier très allusivement de « notionnels ». Il s'agit des cadres introduits par des expressions telles que « Dans le privé », pour lesquels il n'existe pas encore à notre connaissance de méthode d'analyse automatique. On peut certes douter de la possibilité de concevoir un seul et même procédé qui serait à même de traiter toute la diversité recouverte par cette catégorie, mais il s'agit selon nous d'une tâche dont le rôle ne peut être négligé en termes d'analyse thématique, et dont l'étude pourra conduire, si nécessaire, à une subdivision en plusieurs sous-catégories.

ii) Nous souhaitons également nous pencher sur d'autres configurations qui, comme nous avons pu l'observer sur différents exemples, semblent faire intervenir des phénomènes de portée sans pour autant impliquer d'introducteur au sens strict, ni même, dans certains cas, de constructions syntaxiquement détachées. Nous avons notamment évoqué à ce sujet la possibilité de rencontrer des phénomènes de détachement qui seraient d'une autre nature (notamment « référentielle » ou « sémantique »), que nous souhaitons explorer plus avant. Nous envisageons par ailleurs de considérer d'éventuels points de jonction entre la notion de portée en général et des structures d'ordre rhétorique. Nous avons par

exemple évoqué la possibilité, dans la section 8.5.4, d'exploiter la détection de segments conclusifs pour établir la clôture d'un cadre, et nous espérons pouvoir aborder des questions de cet ordre en collaboration avec Antoine Wiclöcher et Nicolas Hernandez.

iii) Une autre voie consisterait à explorer les mécanismes analogues à ceux abordés dans ce mémoire, mais qui concernent des niveaux de grain différents. Leur analyse pourrait venir compléter celle des structures discursives qui ont retenu notre attention jusqu'ici, et il s'agirait donc d'envisager en parallèle différentes manifestations d'un même phénomène thématique, de façon à obtenir, toujours sous l'angle particulier des thèmes composites, une couverture textuelle plus importante. Cela concerne d'une part des objets de grain plus fin, touchant notamment au plan terminologique. On voit bien effet qu'un thème composite tel que LE RETARD SCOLAIRE •—○ (DANS LE SECONDAIRE) pourra aussi bien être instauré par un syntagme tel que « le retard scolaire dans le secondaire » que par une structure cadrative du type « Dans le secondaire, le retard scolaire... », même si les mécanismes assurant leur persistance en discours seront différents (nous décrivons en annexe A les prémisses d'une méthode d'analyse allant dans ce sens). Cela concerne d'autre part des structures de grain élevé, et notamment celles de l'architecture textuelle. Il est en effet fréquent, dans les textes expositifs en particulier, que des thèmes composites soient instaurés par des successions de titres. En témoigne par exemple cet extrait de (HER) :

> X. Les instituteurs
>
> X.1. Les instituteurs publics
>
> X.2. Les instituteurs du privé
>
> Source : HER

vi) Une dernière piste, transversale à celles que nous venons d'évoquer, concerne l'identification des noyaux thématiques au sein des structures en thèmes composites. Nous avons vu dans la section 8.3 que notre méthode d'analyse automatique est fortement centrée sur les satellites thématiques, et que nous nous sommes rangés à des méthodes relativement classiques pour ce qui est des noyaux. Même si cela a été l'occasion de mettre en pratique la collaboration entre méthodes « linguistiques » et « quantitatives », nous souhaitons également envisager la possibilité d'exploiter des concepts gravitant autour des notions de structure informationnelle, de chaîne de référence, et de progression thématique. Nous rejoindrions en cela la problématique de l'interaction, déjà évoquée plus haut, entre portée et continuité référentielle.

***

La plate-forme LinguaStream, décrite dans la partie III, sera bien sûr mise à contribution pour tous les travaux que nous venons de mentionner. Ce qui ne manquera probablement pas, comme cela a été le cas lors de la mise en oeuvre de tous les procédés que nous avons décrits dans ce mémoire, de faire apparaître de nouveaux besoins et de susciter de nouvelles évolutions. Toutefois, nos perspectives à court terme concernant la plate-forme relèvent essentiellement de la consolidation de ce qui a déjà été acquis.

Un premier volet, très pratique mais néanmoins important pour l'élargissement de son champ d'application, concerne l'utilisabilité de LinguaStream en tant qu'environnement d'expérimentation. L'outil a bien sûr déjà été pris en main avec un certain succès par des utilisateurs tiers, mais il est souvent apparu que cela nécessitait un investissement relativement important de leur part, et un support assidu de la part de l'équipe. En effet, même si un effort conséquent a été consacré à la conception de l'environnement d'expérimentation, le foisonnement de ses fonctionnalités semble se traduire à la fois par une impression de trop grande complexité et, paradoxalement, par un manque de visibilité d'une partie des possibilités offertes.

Pour faire face à ce problème au demeurant très ordinaire, plusieurs axes sont à envisager. Le premier consiste bien sûr à produire une documentation exhaustive, ce qui représente un travail considérable mais actuellement en passe d'être finalisé avec le concours de toute l'équipe. Cela se traduit d'une part par la rédaction d'un manuel de référence (et sa maintenance), mais aussi par l'élaboration de tutoriels adaptés aux différents publics visés[5] et la mise en place d'un Wiki permettant à la fois de fournir un support collaboratif et de partager aisément des ressources utilisables avec la plate-forme.

Le second axe concerne l'ergonomie de la plate-forme proprement dite. Il s'agira notamment d'améliorer l'environnement de façon à mieux hiérarchiser la présentation des fonctionnalités disponibles : les plus communément utilisées devront être plus immédiatement accessibles, tout en aiguillant le cas échéant l'utilisateur expérimenté vers les fonctionnalités avancées dont il pourrait avoir besoin. Nous envisageons également la mise en place, au sein même de l'environnement, d'éléments de documentation immédiatement visibles auprès des différents composants offerts à l'utilisateur. Enfin, il s'agira de multiplier les interfaces de type « assistant », qui guident l'utilisateur dans la réalisation d'un certain nombre de tâches plus ou moins habituelles.

Au delà de ces aspects touchant à l'utilisabilité de l'environnement d'expérimentation, nous souhaitons également exploiter plus avant le cadre général fourni par LinguaStream en tant que plate-forme pour intégrer de nouveaux systèmes d'analyse. Nous rejoignons ainsi l'une des principales ambitions de départ, visant à capitaliser les travaux réalisés au sein de l'équipe et à faciliter leur réutilisation. Hormis les différents travaux qui ont été réalisés sous LinguaStream et qui sont donc, de fait, cessibles et réutilisables, il serait utile de rendre disponibles pour les utilisateurs de la plate-forme d'autres systèmes comme l'analyseur syntaxique de Jacques Vergne ou l'analyseur aspecto-temporel développé par Cédric Person[6]. L'intégration de systèmes tiers, comme nous l'avons fait pour TreeTagger ou Syntex, pourra également se révéler très utile ; on pensera sur ce point à des logiciels comme Cordial ou Unitex.

Nous souhaitons également poursuivre les applications au départ moins attendues de la plate-forme, tout d'abord sur le plan pédagogique. Dans ce contexte, la dissimulation de l'appareil procédural au profit des formalismes déclaratifs permet en effet de se concentrer sur l'analyse d'un phénomène linguistique sans être parasité par des difficultés d'ordre technique. Le principe de modularité permet pour sa part d'isoler un problème singulier sans perdre le bénéfice des analyses préalables éventuellement nécessaires, en considérant simplement leur apport comme une donnée du problème. Enfin, les différents modes de visualisation proposés permettent de rendre les phénomènes étudiés observables très concrètement, en contexte. On notera à ce propos que des outils spécifiques ont également été développés, en se fondant sur la plate-forme, à des fins pédagogiques, et que nous considérons à l'heure actuelle leur transformation sous forme d'applications Web, plus faciles à déployer. Enfin, parmi les applications de LinguaStream qui ne faisaient pas partie des plans initiaux, on notera son rôle en tant que point d'appui, déjà mentionné dans le chapitre précédent, de notre collaboration avec l'équipe « fouille de texte » du laboratoire, que nous souhaitons bien évidemment poursuivre avec le plus grand intérêt.

L'ensemble de ces tâches constituant un programme relativement lourd, et relevant pour partie de l'ingénierie logicielle pure, nous espérons à terme avoir l'opportunité de faire de LinguaStream un logiciel *open-source*, afin d'élargir aussi bien l'équipe de développement que la base de ses utilisateurs.

[5] De profils plutôt « informaticien » ou plutôt « linguiste ».

[6] Comme nous l'avons déjà dit, un travail sur ce point précis a déjà été entamé par l'intéressé.

Annexes

Annexe A

Analyse thématique et structures terminologiques

Comme nous l'avons vu dans le chapitre 4, une approche couramment adoptée en analyse thématique automatique consiste à étudier la distribution des formes de surface dans les documents, et plus particulièrement leur fréquence absolue ou relative. Cette méthode peut être utilisée en indexation automatique pour circonscrire un ensemble de mots ou de termes fréquents dans un document et donc, d'un certain point de vue, caractéristiques de son contenu informationnel. Elle peut également être appliquée au problème de la segmentation thématique, où les formes caractéristiques d'un fragment textuel sont comparées aux formes caractéristiques d'autres fragments pour évaluer leur cohésion thématique. Nous avons évoqué différents problèmes propres à cette approche, globalement liés au fait que la distribution des formes de surface n'est que très approximativement représentative des phénomènes sémantiques qui interviennent dans l'organisation thématique d'un document.

L'expérience dont nous faisons état dans cette annexe repose sur une étude purement quantitative de la distribution des mots dans les documents, à laquelle nous adjoindrons une étude de la structure des termes composés à partir de ces mots. Plus précisément, nous observerons comment les mots fréquents s'agencent sous la forme de syntagmes nominaux complexes, agencement que nous représenterons sous la forme de graphes. Nous verrons d'une part que l'on peut reconnaître un intérêt à ces graphes en tant que représentation structurée mais calculable à peu de frais de l'à propos d'un texte. D'autre part, nous observerons la structure de ces graphes pour faire apparaître d'une manière différente les phénomènes sémantiques qui nous ont intéressé dans ce mémoire.

La première phase de la constitution de nos « graphes terminologiques » repose sur une analyse des plus classiques de la fréquence relative des formes dans un texte. Rappelons que dans sa version la plus simple, ce procédé consiste à déterminer, pour un fragment textuel donné, l'ensemble des mots pleins[1] qui apparaissent fréquemment dans ce fragment relativement aux fragments qui l'entourent. Le coefficient $tf \cdot idf$ est classiquement utilisé à cette fin, et peut aussi bien s'appliquer au niveau documentaire (dans ce cas un mot est considéré comme thématiquement significatif s'il est fréquent dans un document relativement aux autres documents de la base) qu'à des fragments plus fins de l'ordre du paragraphe, comme cela est par exemple pratiqué par les méthodes de type *text-tiling* (dans ce cas un mot sera considéré comme thématiquement significatif s'il est fréquent dans un segment relativement à l'ensemble du document).

A priori, rien ne s'oppose à ce que cette méthode soit appliquée à des syntagmes complexes plutôt qu'à des mots simples : partant du principe communément admis en indexation que des syntagmes fourniront une représentation plus précise du contenu d'un document que des mots isolés, il est tout à

[1] Par opposition aux mots dits « vides » ou grammaticaux, qui sont généralements éliminés à l'aide d'un antidictionnaire.

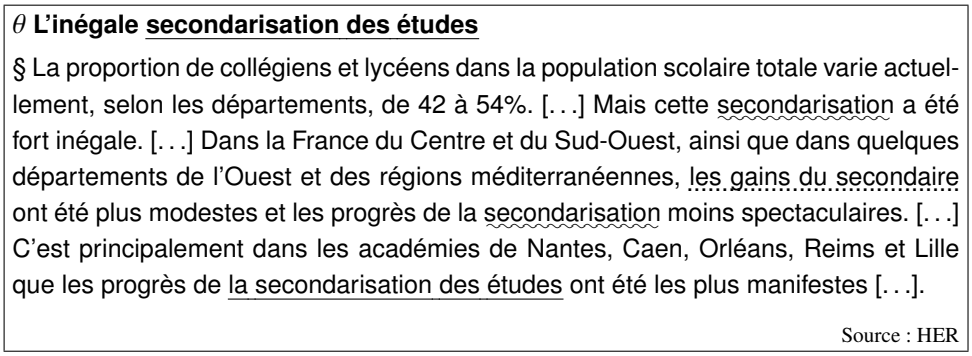


FIG. A.1 – Phénomènes de re-lexicalisation et de réduction terminologique.

fait possible d'appliquer la technique ci-dessus à ce niveau de grain. En pratique, cela pose toutefois le problème de la récurrence des syntagmes complexes, ou plutôt de leur non-récurrence. En effet, un syntagme nominal particulièrement représentatif de l'à propos d'un fragment peut ne pas être fréquent dans ce même fragment, en raison de différents phénomènes de re-lexicalisation, de paraphrase ou encore de réduction terminologique (Jacques, 2001). Et bien évidemment, ce problème se posera de façon d'autant plus importante que les fragments considérés seront courts.

Considérons par exemple le paragraphe de la figure A.1. Le syntagme « la secondarisation des études » mentionné dans le titre semble assez représentatif de l'à propos de ce passage, mais les différents phénomènes que nous venons d'évoquer font qu'il n'apparait qu'une seule fois dans son intégralité, alors que le mot « secondarisation » est relativement fréquent. On observera également différentes re-lexicalisations telles que « les gains du secondaire ».

Il est également fréquent qu'un syntagme représentatif du thème d'un segment n'y apparaisse jamais dans son intégralité. Par exemple, un paragraphe traitant de « l'augmentation des effectifs scolaires dans le secondaire » pourra ne contenir que les termes « les effectifs du secondaire » d'une part, et « l'augmentation des effectifs scolaires » d'autre part. Dans ce cas, on pourra au mieux espérer que les mots « augmentation », « effectifs », et « secondaire » comptent effectivement parmi les plus fréquents dans un tel segment, mais même dans ce cas, cet ensemble de mots isolés ne constituera pas une représentation réellement satisfaisante du contenu du passage.

La prise en compte de ces problèmes rejoint la problématique abordée par les systèmes d'analyse terminologique automatique tels que ceux qui ont été élaborés sur la base de Syntex (Bourigault et Fabre, 2000), et qui prennent en compte à la fois la distribution des mots et leur fonction syntaxique. Toutefois, ces méthodes sont plutôt orientées vers la constitution assistée de terminologie que vers le problème de l'analyse thématique en tant que tel, bien que ces deux problématiques se recouvrent pour une large part. L'expérience que nous allons détailler ici consiste à appliquer une méthode de ce type à l'analyse thématique, de façon à faire apparaître clairement la nécessité d'avoir recours à des *structures* dans la représentation de l'à propos d'un passage, en adoptant pour un temps une approche terminologique.

L'idée consiste à évaluer la représentativité distributionnelle des mots simples pour un grain donné (par exemple document ou le paragraphe), que nous utiliserons pour évaluer la représentativité des syntagmes nominaux complexes. Par suite, les syntagmes nominaux saillants seront regroupés sous forme de graphe pour obtenir une représentation de la structure terminologique représentative du fragment considéré. Voici le détail de cette méthode, mise en oeuvre à l'aide de la chaîne de traitement reproduite dans la figure A.2 :

– Après quelques étapes préliminaires visant à s'assurer de la validité XML du document à traiter, le système procède dans un premier temps aux étapes habituelles de découpage lexical, d'éti-

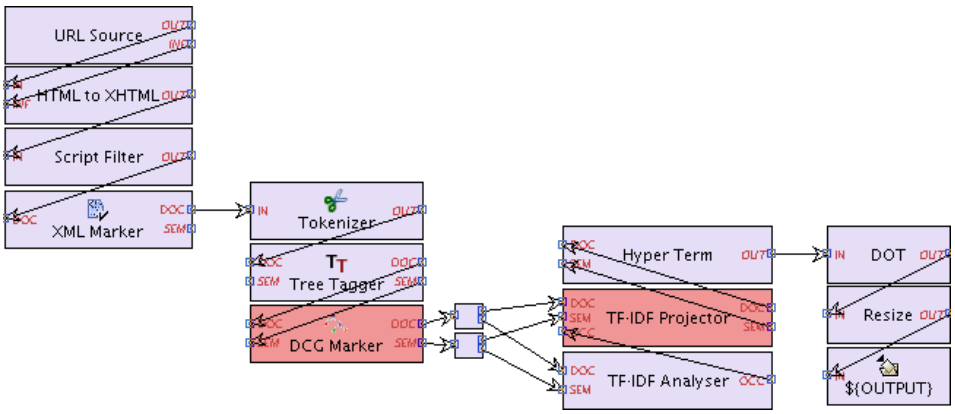


FIG. A.2 – Chaîne de traitement LinguaStream du processus d'extraction des hypertermes.

quetage morpho-syntaxique et de lemmatisation.

- Par la suite, une grammaire locale simple (reproduite en annexe E.1) permet de baliser les syntagmes nominaux complexes à partir des étiquettes préalablement attribuées aux mots simples. Finalement, chaque syntagme sera représenté par la liste des lemmes de ses mots pleins, par exemple « évolution / effectif / secondaire » pour « les évolutions des effectifs du secondaire ».
- L'étape suivante consiste à calculer le coefficient $tf \cdot idf$ pour chaque mot simple, étant donné un certain grain. Nos essais ont été menés en choisissant le grain paragraphe, pour les raisons que nous allons détailler par la suite.
- La dernière étape consiste à constituer des graphes à partir des syntagmes nominaux les plus représentatifs, graphes que nous appellerons « hypertermes ».
- Un outil de mise en forme de graphes est finalement utilisé pour représenter ces hypertermes sous forme graphique[2].

Nous ne nous étendrons pas ici sur les premiers niveaux de l'analyse, qui sont relativement habituels. Nous allons en premier lieu nous pencher sur le calcul de la représentativité distributionnelle des mots simples et des syntagmes nominaux complexes, et tout d'abord sur le grain choisi pour le calcul du $tf \cdot idf$, qui est de l'ordre du paragraphe. La première raison de ce choix est que nous souhaitons nous donner les moyens de représenter l'à propos de passages précis et non du document pris dans son intégralité. Une seconde raison est que nous ne chercherons pas ici à déterminer la représentativité des mots dans un document relativement à d'autres documents, c'est-à-dire leur caractère discriminant par rapport à l'ensemble de la base, ce que fournirait l'application du coefficient au grain documentaire. Au contraire, le calcul de ce coefficient nous fournira la représentativité de chaque mot relativement à chaque paragraphe. Pour un paragraphe donné, nous obtenons ainsi une représentation « brute » de son à propos sous forme d'une liste de mots localement saillants. Précisons que, de façon tout à fait classique, les coefficients sont en fait calculés pour les lemmes et non pour les graphies des mots.

À partir de cette représentation brute, nous allons chercher dans un premier temps à en déduire quels sont les syntagmes nominaux saillants pour chaque segment. Ce calcul est extrêmement simple, puisque nous nous contenterons de calculer le coefficient d'un syntagme en faisant la somme des coefficients des lemmes qui entrent dans sa composition. Il est clair que cette méthode pourrait être affinée à bien des égards, ce serait-ce qu'en considérant une moyenne ou en exploitant une pondération associée à la structure syntaxique des syntagmes (la tête pourrait par exemple se voir attribuer un poids plus important). En tout état de cause, nous obtiendrons ainsi pour chaque syntagme et pour chaque segment une valeur numérique que l'on considérera comme caractéristique de sa saillance au sein

[2] À l'aide de l'outil « DOT » inclus dans la suite « GraphViz ».

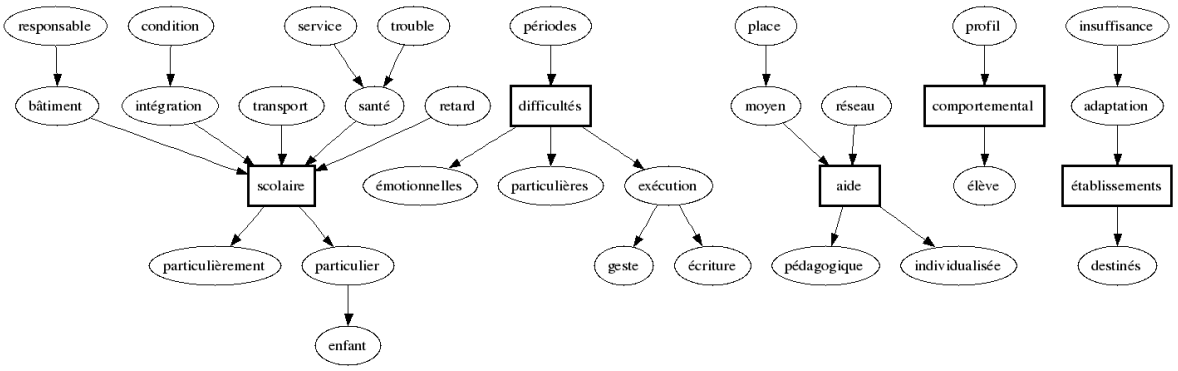


FIG. A.3 – Hypertermes caractéristiques extraits d'un document portant sur l'intégration scolaire des enfants atteints de myopathie.

d'un segment. Ainsi, un syntagme tel que « la secondarisation des effectifs » pourra se voir attribuer un coefficient élevé pour un segment donné même s'il est lui-même peu fréquent dans ce segment, pour peu que le soient les mots qui le composent.

L'étape suivante étape consiste à regrouper les syntagmes les plus saillants sous la forme de graphes que nous appelons « hypertermes ». Le principe de ce regroupement est extrêmement simple : si dans un passage ou un document donné apparaissent deux syntagmes composés des mots A-B et B-C (en ignorant les mots grammaticaux), on ajoutera la suite A-B-C dans le graphe. Si par la suite on rencontre un syntagme B-D, l'arc correspondant sera également ajouté au graphe. Par exemple, si apparaissent à la fois les syntagmes « l'augmentation des effectifs » et « les effectifs du secondaire », on ajoutera dans le graphe la séquence « l'augmentation des effectifs du secondaire », même si celui-ci n'apparaît jamais en tant que tel. Si le syntagme « les effectifs du primaire » apparaît à son tour, le noeud « primaire » sera ajouté et lié au noeud « effectifs ». En procédant ainsi pour un l'ensemble des syntagmes rencontrés, on obtient un ensemble de sous-graphes connexes, dont un exemple est donné dans la figure A.3.

Au fur et à mesure de la constitution de ces graphes, chaque sous-graphe connexe se voit attribuer un coefficient caractéristique de sa représentativité thématique. Là encore, le coefficient d'un sous-graphe correspond à la somme des coefficients des syntagmes qui le composent. Le coefficient d'un sous-graphe est donc d'autant plus important qu'il est constitué de syntagmes eux-mêmes représentatifs. Notons que la même remarque que précédemment s'applique également ici : différents modalités plus évoluées de calcul du « coefficient total » d'un sous-graphe pourraient se substituer à une simple addition.

Une fois les graphes constitués, une étape supplémentaire consiste à choisir un « noyau » pour chaque sous-graphe connexe, qui correspond au noeud dont le degré est le plus élevé, c'est-à-dire le noeud connecté au plus grand nombre d'arcs. En d'autres termes, il s'agit du mot qui se trouve au « centre » du graphe. Ce principe rejoint la notion de « productivité » qui est employée en extraction terminologique (cf. (Bourigault et Fabre, 2000)), et qui désigne la propension d'un mot à prendre part à la constitution de termes complexes. Ici, nous considérerons que ces noyaux sont particulièrement représentatifs d'un point de vue thématique, et sont représentés dans la figure A.3 sous forme de rectangles au trait gras. Notons que le degré du noyau d'un sous-graphe participe également au calcul du coefficient du dit sous-graphe, sous la forme d'un coefficient multiplicateur, considérant qu'un hyperterme doté d'un noyau très productif est d'autant plus représentatif du point de vue thématique.

Finalement, le calcul du coefficient W_{ig} d'un hyperterme g pour un segment i est le suivant, où w_j est un mot simple, t_k un syntagme nominal, tf_{ij} le poids du j^{eme} mot dans le segment i, df_j le nombre de segments dans lesquels apparaît ce mot, N le nombre total de segments, et D_g le degré du graphe g :

$$W_{ig} = D_g \cdot \sum_{t_k \in g} \sum_{w_j \in t_k} tf_{ij} \cdot log\left(\frac{N}{df_j}\right)$$

Une fois déterminé le noyau de chaque graphe connexe, nous procédons enfin à un élaguage des noeuds trop éloignés de celui-ci, étant donnée une distance maximale fixée arbitrairement. Dans cette même figure, la distance maximale du noyau a été fixée à 2. L'intérêt de cette manipulation est de produire des graphes de taille réduite que l'on peut présenter un utilisateur. L'autre intérêt est de scinder certains graphes en plusieurs sous-graphes s'ils sont liés par des noeuds trop éloignés de leurs noyaux, ce lien pouvant alors être considéré comme peu significatif.

Étant donnés l'ensemble des graphes ainsi obtenus et les « scores » qui leur sont attribués, on peut sélectionner les hypertermes les plus pertinents pour un segment ou un document, et les présenter à un utilisateur en tant de représentation de son contenu informationnel. Cette présentation peut paraître déroutante au premier abord, notamment parce que les graphes peuvent faire apparaître des termes qui n'existent pas physiquement dans le document, et parce qu'il est nécessaire pour l'utilisateur de reconstituer lui-même les termes possibles à partir de suites de lemmes privées des mots grammaticaux. Il nous semble toutefois que cette représentation peut fournir un aperçu intéressant du contenu d'un document : la mise en valeur des noyaux de chaque sous-graphe permet de rendre compte compte des mots qui sont non seulement fréquents dans le document mais qui sont également les plus « productifs » du point de vue terminologique, et surtout, les graphes eux-mêmes donnent un bon aperçu de la façon dont ces mots interviennent dans le réseau conceptuel propre au document.

Étant donnée la relative légèreté de la méthode, il est par exemple envisageable de l'utiliser en aval d'un moteur de recherche classique afin de proposer à l'utilisateur une vue condensée des documents retrouvés. Nous avons réalisé un prototype basé sur ce principe en encapsulant la chaîne de traitement de la figure A.2 dans un script LinguaStream (cf. annexe E.3) afin d'extraire les hypertermes représentatifs des documents retournés par une requête Google[3]. Les résultats obtenus sont présentés dans un navigateur sous une forme imitant la présentation du dit moteur de recherche, où sont intégrés sous forme graphique les hypertermes les plus représentatifs de chaque document (du point de vue du coefficient décrit plus haut). En cliquant sur un hyperterme, l'utilisateur peut alors accéder directement au(x) passage(s) dont il est issu. Un exemple d'affichage produit par ce prototype est donné dans la figure A.4.

Au delà de l'éventuel intérêt pratique des hypertermes, on peut s'intéresser aux indications qu'ils fournissent quant à l'organisation des concepts propres à un document, ou même à un domaine si ce document en est représentatif. Considérons par exemple les hypertermes de la figure A.5, extraits d'un ouvrage traitant du système scolaire français (HER). On peut voir apparaître deux parties au sein de chaque hyperterme, l'une étant constituée du noyau et des noeuds qui le précèdent dans le graphe, et l'autre de ceux qui suivent le noyau, par exemple « la baisse des effectifs » *vs.* « dans l'enseignement primaire ». Cette configuration apparaît comme un autre reflet des phénomènes évoqués dans la partie II.

Cette dualité constitue également une modalité courante de constitution de termes complexes, qui se base sur l'adjonction à une tête syntagmatique d'une ou plusieurs extensions prépositionnelles permettant de situer le référent de cette tête relativement à d'autres concepts « auxiliaires » ou « satellites », par rapport au temps, à l'espace, etc. Ce phénomène nous paraît important dans le cadre de l'analyse thématique, dans la mesure où elle constitue un moyen d'exprimer de façon concise une réalité conceptuelle relativement complexe. De fait, il peut être utilisé pour construire des titres faisant référence non seulement à une ou plusieurs notions qui seront abordées dans la section titrée, mais aussi au contexte auquel elles se rattachent. Par exemple :

[3] l'interaction avec ce dernier est réalisée à l'aide de son interface SOAP.

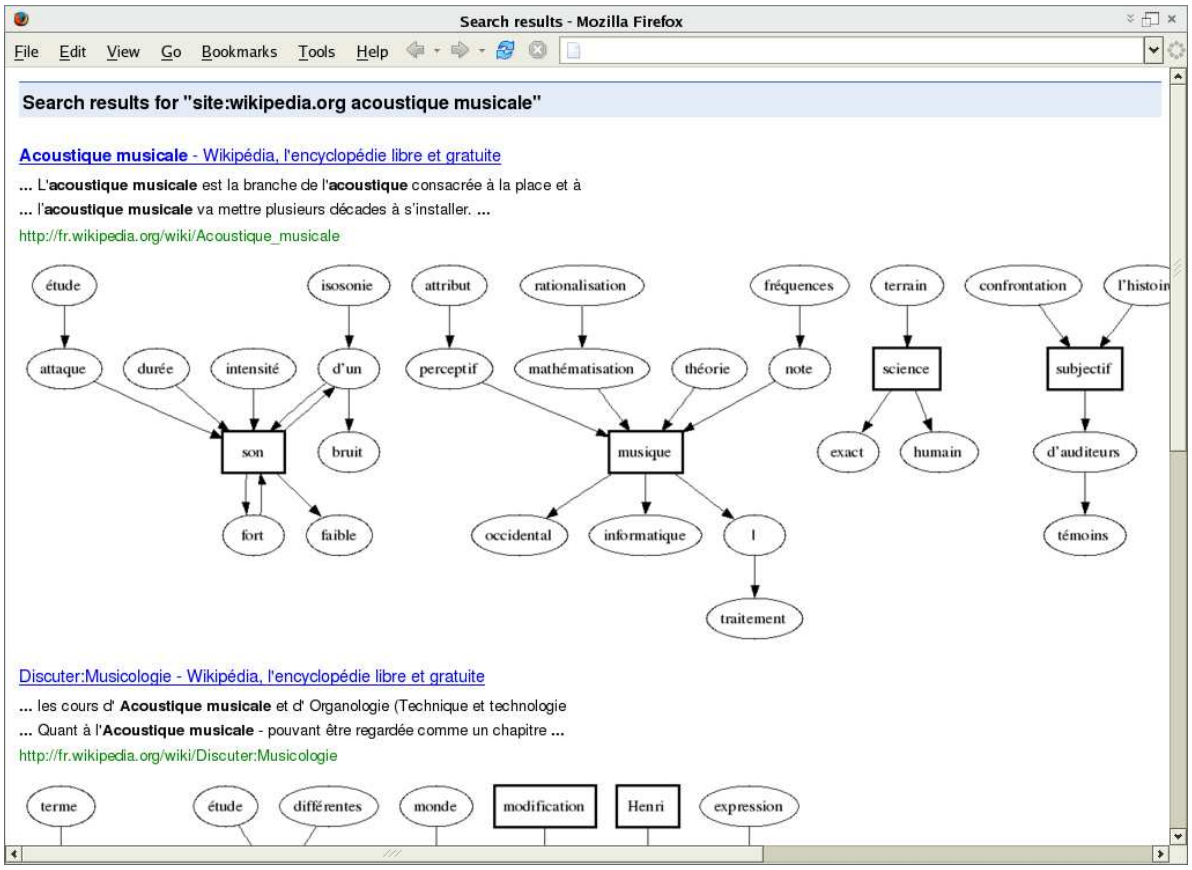


FIG. A.4 – Résultats de recherche issus du moteur Google, agrémentés des hypertermes caractéristiques des documents retrouvés.

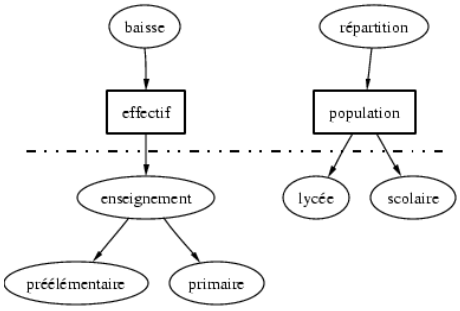


FIG. A.5 – Hypertermes extraits de (HER).

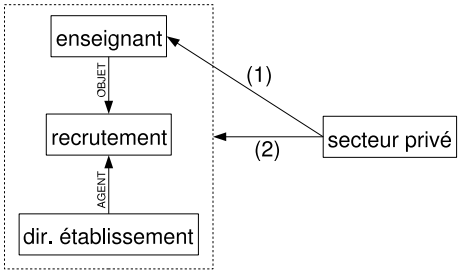


FIG. A.6 – Statut intra- ou extra-argumental d'un satellite.

Le tourisme dans les régions françaises de la zone Transmanche
Les liaisons maritimes en 1994
Professer au collège et au lycée
Les enseignants du public
Administration réseau sous Linux

Source : ATM, HER, ARL

Mais il est clair que le positionnement « terminologique » que nous avons adopté ici n'apporte pas de solution réellement satisfaisante. Car s'il est souvent possible d'exprimer sous la forme d'un syntagme complexe l'à propos d'un fragment textuel, il est bien évident que ce syntagme n'apparaîtra pas toujours littéralement dans le texte. Et même si la constitution des hypertermes permet de faire apparaître des structures syntagmatiques qui ne sont pas présentes en tant que telles dans le texte, il est bien évident que la portée de l'analyse décrite ici reste très limitée puisqu'elle ne s'écarte pas de ce niveau syntagmatique. Il lui est en effet impossible de rendre compte des diverses structures linguistiques responsables en discours de l'établissement des liens entre noyaux et satellites. Considérons par exemple ces trois réalisations d'une même relation sémantique entre « les enseignants » et « le privé » :

(1) Les enseignants du privé sont directement recrutés par le chef d'établissement.

(2) Les enseignants sont recrutés différemment dans le privé : c'est le chef d'établissement qui [...]

(3) Dans le privé, la situation est différente. La législation stipule que [...]. Les enseignants sont donc recrutés directement par le chef d'établissement.

Dans le cas (2), les deux composantes qui nous intéressent ici n'apparaissent plus au sein d'un même syntagme comme c'était le cas en (1), bien que le lien sémantique persiste. Les énoncés (1) et (2) ne sont certes pas équivalents en toute généralité : dans le second cas la contrainte véricconditionnelle imposée par le syntagme prépositionnel « dans le privé » s'applique à l'ensemble de la prédication, quand celle de son « homologue » du premier énoncé est en principe restreinte au syntagme auquel il appartient. Ces deux configurations réalisent ainsi deux structures sémantiques distinctes que nous avons représentées graphiquement dans la figure A.6. Toutefois, nous pouvons considérer que le lien entre les deux composantes qui nous intéressent ici est bien représenté dans les deux cas : dans le premier énoncé, la restriction au secteur public s'applique par extension à l'ensemble de prédication, alors que dans le second elle d'applique *a fortiori* aux composantes de cette prédication.

Or il est bien évident qu'une analyse terminologique ne peut rendre compte de ce lien dans le second cas : en toute rigueur, sa détection automatique devrait reposer sur une analyse syntaxique plus ou moins exhaustive de la phrase. Ou plus simplement, si l'on est capable de déterminer avec

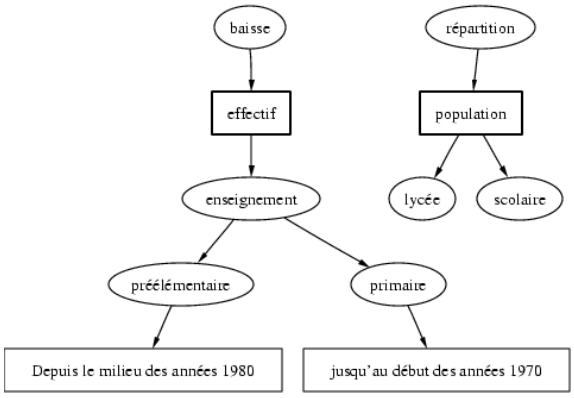


FIG. A.7 – Hypertermes « temporalisés » extraits de (HER).

une précisions satisfaisante les limites des phrases (ou mieux, des propositions) et de détecter les adverbiaux pertinents, une approximation acceptable peut consister à relier ensemble les composantes qui apparaissent au sein de la même phrase ou proposition. Cette approche peut d'ailleurs fournir un moyen peu coûteux d'agrémenter nos hypertermes en ajoutant comme successeurs d'un noeud certains adverbiaux qui apparaissent dans les mêmes phrases que lui. La figure A.7 montre un exemple de ce type, où nous avons utilisé l'analyseur d'expressions temporelles tel que présenté dans le chapitre 6 pour ajouter des satellites temporels aux hypertermes déjà présentés dans la figure A.5.

Envisageons maintenant le cas (3) : dans cette configuration, le satellite « le privé » n'apparaît plus dans la même phrase que le noyau « les enseignants ». Pourtant, en tant qu'introducteur de cadre de discours (cf. section 3.4.1), ce satellite apporte un critère susceptible de s'appliquer à plusieurs propositions, et sa portée s'étend en l'occurrence jusqu'à celle où apparaît notre noyau. Le lien noyau/satellite existe donc bien, mais l'analyse phrastique ne suffit plus à en rendre compte : il est nécessaire de prendre en considération une configuration discursive particulière, ici un cadre de discours.

À des niveaux de grain plus élevés, des phénomènes similaires peuvent également apparaître au sein de l'architecture textuelle. C'est par exemple le cas pour les titres reproduits plus haut, qui spécifient à chaque fois une composante du contexte valant, par défaut, pour l'ensemble de la section titrée. Plus généralement, du syntagme au texte, il existe manifestement une multitude de phénomènes linguistiques susceptibles de conduire à l'établissement de liens sémantiques entre un noyau et un satellite. Les phénomènes discursifs envisagés dans le chapitre 8 en font partie, mais on voit bien ici qu'il serait nécessaire, pour obtenir une couverture globale de la structure thématique d'un document, de prendre en compte d'autres phénomènes, de grains inférieur (niveau terminologique par exemple) et supérieur (architecture textuelle par exemple).

Annexe B

Application du modèle Anadia à la segmentation thématique

Anadia (Beust, 1998) est un modèle de la mémoire sémantique fondé sur la sémantique componentielle. En définissant le sens comme le produit du rapport entre les signifiés, ce modèle se matérialise par une organisation relationnelle des signes (ou plus précisément, des lexies) sur la base de leur valeur (au sens de Saussure), afin de rendre cette valeur calculable par une méthode de catégorisation, elle-même basée sur la typologie des sèmes. On se rapportera si nécessaire à la bibliographie pour des plus amples informations sur ce modèle qui ne sera pas décrit ici.

Nous montrons l'intérêt de l'utilisation d'un dispositif Anadia dans le contexte de l'analyse thématique. Plus précisément, on cherchera à évaluer l'apport d'un tel dispositif à une méthode de type *text tiling* (Hearst, 1994) (cf. section 4.1). Après avoir exposé succinctement une méthode d'analyse thématique distributionnelle simple, nous allons montrer comment le modèle Anadia peut être utilisé pour améliorer la segmentation produite par cette méthode, ainsi que pour mieux caractériser les relations existant entre les segments identifiés.

Méthode de segmentation de base

La méthode présentée ici est semblable à celle exposée dans la section 4.1. Il s'agit donc d'une méthode purement quantitative, fondée uniquement sur l'analyse de la distribution des occurrences des mots.

Le pré-traitement ici utilisé se limite à un étiquetage à l'aide du *TreeTagger* (Schmidt, 1994), qui permet à la fois d'effectuer une sélection sur la nature des mots (on ne conservera que les noms et les verbes), et d'obtenir une lemmatisation. À la suite de ces pré-traitements, on considère un premier niveau de segmentation *a priori*, définissant l'unité textuelle minimale. L'unité choisie ici est le paragraphe, mais ce choix pourrait éventuellement être adapté au volume du corpus traité.

On notera que nous effectuons ici un traitement particulier sur les descripteurs, qui sont regroupés sur un critère de racines communes, en fonction d'un ensemble de classes de suffixes *après* le calcul des vecteurs propres à chaque segment. Nous partons du principe que si deux termes sont saillants dans un même segment, qu'ils partagent une même racine, et que les suffixes qui les distinguent appartiennent à une même classe, alors ces deux termes peuvent être regroupés en un seul descripteur, le poids du descripteur obtenu étant égal à la somme de leurs poids respectifs. On regroupera ainsi par exemple « enseign-er », « enseign-ement » et « enseign-ant » en un seul descripteur si ces trois termes sont saillants au sein d'un même segment.

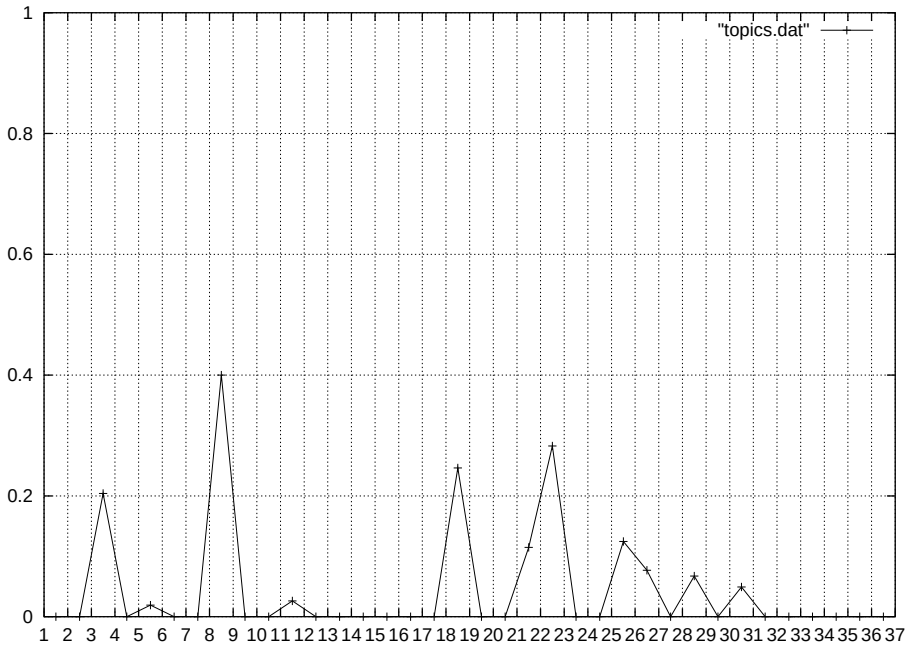


FIG. B.1 – Analyse thématique initiale

Les classes de suffixes ont été établies empiriquement, en se basant sur l'étude des descripteurs. Il faut bien noter qu'elles ne sont applicables que sur les descripteurs saillants d'un même segment, au sujet desquels on peut présumer d'une certaine cohérence thématique, et en risquant donc un taux d'erreur minimal. Bien que très simple, ce procédé a permis une amélioration sensible de la qualité des descripteurs (et donc de celle de la segmentation), sans pour autant nécessiter des connaissances lexicales supplémentaires.

La méthode a été implémentée et testée sur un texte technique dans le domaine de l'informatique. Le graphique de la figure B.1 présente les résultats obtenus. La courbe relie les points correspondant à la mesure de cohésion obtenue pour chaque couple de paragraphes :

Les pics témoignent d'une certaine cohérence thématique entre les paragraphes, par exemple 8-9, ou 25-26-27. Si ces regroupements sont pour la plupart justifiés, on constate à la lecture du texte que la qualité globale du résultat est assez faible. En effet, même s'il est difficile d'évaluer la qualité d'une segmentation thématique (très subjective par essence), il est flagrant que certains regroupements thématiques évidents ne ressortent absolument pas sur cette analyse.

Utilisation du modèle Anadia

Motivations

Nous allons à présent montrer comment le modèle Anadia peut être utilisé pour améliorer cette méthode de segmentation thématique. Au même titre que les différentes méthodes évoquées plus haut, le principe de cette extension à la méthode de base repose sur la notion de cohésion lexicale.

La méthode exposée plus haut est très restrictive quand aux relations existant entre les descripteurs. En effet, le calcul cohérence entre les segments est basée sur l'équivalence stricte des termes, les seules variations permises étant issues de la lemmatisation et du regroupement sur les préfixes communs. De

plus la relation entre deux descripteurs est purement binaire : la relation existe ou non, mais on ne peut pas quantifier le degré de cohésion. On va donc chercher à établir des relations plus évoluées entre les descripteurs afin d'afiner le calcul de la distance entre les segments.

L'intérêt d'un dispositif Anadia pour cette tâche réside dans le fait qu'elle offre une structuration des sèmes, des sémèmes et des taxèmes, auxquelles on pourra accéder par les lexies, qui correspondent dans notre cas aux descripteurs. Ces différentes relations seront exploitées pour évaluer finement la cohésion sémantique entre les descripteurs.

Principe général

Le principe proposé est le suivant : étant donnés un dispositif Anadia et deux descripteurs dont on souhaite mesurer la cohérence, on procédera comme suit :

- On recherche les lexies correspondantes au sein du dispositif. Si au moins l'un des deux descripteurs est absent, on considère que leur relation est nulle relativement à ce dispositif.
- Si les deux descripteurs appartiennent à une même table, on considère que la distance qui les sépare est proportionnelle au nombre de traits qui les distinguent au sein de cette table (si les deux lexies appartiennent à la même catégorie, la distance est donc considérée comme nulle).
- Si les deux descripteurs appartiennent à deux tables différentes, on considère que la distance qui les sépare est proportionnelle à celle qui sépare ces deux tables. La distance entre deux tables est définie comme le nombre de liens de spécialisation qui les séparent. Si deux tables sont totalement indépendantes (il n'existe aucun chemin entre elles dans le réseau des tables), la relation entre les deux descripteurs est considérée comme nulle.

Ce principe repose sur l'idée que le nombre de traits distinguant deux lexies au sein d'une table est révélateur de la distance sémantique existant entre ces lexies. De même, on suppose que les relations de spécialisations formant un chemin au sein du réseau entre deux tables distinctes permet également d'établir une notion de distance.

D'autre part, en plus de fournir une mesure de la cohésion entre deux descripteurs, ce principe permet en outre de caractériser cette relation :

Synonymie quand les deux descripteurs appartiennent à la même catégorie.

Isotopie quand les deux descripteurs appartiennent à la même table.

Généralisation/spécialisation quand les deux descripteurs appartiennent à des tables connexes.

Application à la méthode de segmentation

Nous allons montrer ici comment les principes évoqués plus haut peuvent être appliqués au calcul effectif de la distance entre deux descripteurs, et donc entre les segments eux-mêmes.

Etant donnés deux vecteurs $X = (x_1, ..., x_n)$ et $Y = (y_1, ..., y_n)$ décrivant les segments dont on souhaite mesurer la cohésion, l'idée est de modifier, pour chaque descripteur x_i de X, la pondération de chaque descripteur y_j de Y en fonction du rapport entre le poids $w(x_i)$ de x_i et de la mesure de cohésion donnée par un « coefficent Anadia » noté $A\left(x_i, y_j\right)$. Puis on procède symétriquement. C'est à dire :

$$\forall (i,j) \in \{1, ..., n\}^2 : \begin{cases} w(x_i) \leftarrow w(x_i) + w\left(y_j\right) \cdot A\left(x_i, y_j\right) \\ w(y_i) \leftarrow w(y_i) + w\left(x_j\right) \cdot A\left(x_j, y_i\right) \end{cases}$$

Ce calcul permet à chaque couple de descripteurs liés par le dispositif Anadia de se renforcer mutuellement, en tenant compte à la fois de la pondération de chacun et d'une mesure A de leur

cohésion sémantique. On définit ce « coefficent Anadia » comme suit (on dira que deux descripteurs sont indépendants si l'un et/ou l'autre est absent du dispositif, ou s'ils appartiennent à des tables non connexes) :

$$A(x,y) = \begin{cases} 0 & \text{si x et y ne sont pas liés} \\ \frac{1}{1+d_A(x,y)} & \text{sinon} \end{cases}$$

La valeur $d_A(x,y)$ correspond à la distance entre les descripteurs x et y au sein du dispositif. Cette distance est calculée comme explicité plus haut, c'est-à-dire soit par le nombre de traits distinguant ces descripteurs au sein de de la même table, soit par la longueur du chemin entre deux tables distinctes. On peut éventuellement multiplier ce dernier par un coefficient, modérant ainsi l'importance de la relation de généralisation/spécialisation par rapport à la relation d'isotopie (dans les expérimentations exposées plus bas, nous avons ici utilisé un coefficient 2).

Après cette étape de renforcement des descripteurs, la distance entre les segments peut être calculée de la même façon que dans la méthode de base.

Cette méthode permet également, comme il est évoqué plus haut, de caractériser la relation entre deux descripteurs (synonymie, isotopie ou généralisation/spécialisation). On peut donc tenter d'exploiter ces informations pour caractériser de la même façon les relations entre les segments. On comptabilise à cette fin, pour chaque couple de segments, le nombre de relations de chaque type relevées lors du calcul du coefficient A. Si un type de relation donné se révèle proéminent relativement aux autres, on pourra considérer qu'il peut s'appliquer également aux segments eux-mêmes.

Résultats obtenus

La méthode présentée ci-dessus a été implémentée et testée en utilisant un dispositif du domaine de l'informatique (il s'agit d'un dispositif fourni avec l'application Anadia, légèrement modifié). La figure B.2 montre les résultats obtenus sur le même texte que précédemment.

En les comparant avec les résultats obtenus avec la méthode initiale, on observe comme on pouvait s'y attendre un plus grand nombre de regroupements, ceux déjà présents précédemment étant pour la plupart élargis ou plus saillants. On remarquera en particulier les regroupements des paragraphes 10 à 15, puis 16-17 qui n'apparaissaient pas dans la première expérimentation, et qui paraissent très pertinents au vu du texte puisqu'ils correspondent exactement au découpage logique de l'auteur. On remarquera également l'apparition du regroupement des paragraphes 28 à 31, qui paraît lui-aussi cohérent. Le résultat obtenu semble donc assez satisfaisant en comparaison avec celui de la méthode de base.

En revanche, la caractérisation des relations entre les descripteurs sur ce texte paraît difficilement propageable aux segments, car on ne peut en général distinguer aucun type proéminent par rapport aux autres. Cela peut en partie s'expliquer par le fait que, les paragraphes étant relativement courts, le nombre de descripteurs significatifs est en général trop faible pour faire émerger une relation caractéristique.

Conclusion

D'après les expérimentations réalisées, et en partie exposées ci-dessus, il apparaît que l'utilisation d'Anadia permet améliorer très sensiblement les résultats produits par une méthode de segmentation thématique distributionnelle classique. Il convient cependant de noter que si elle en améliore les résultats, elle ne permet pas d'en contourner la contrainte principale, à savoir que ce type de méthode

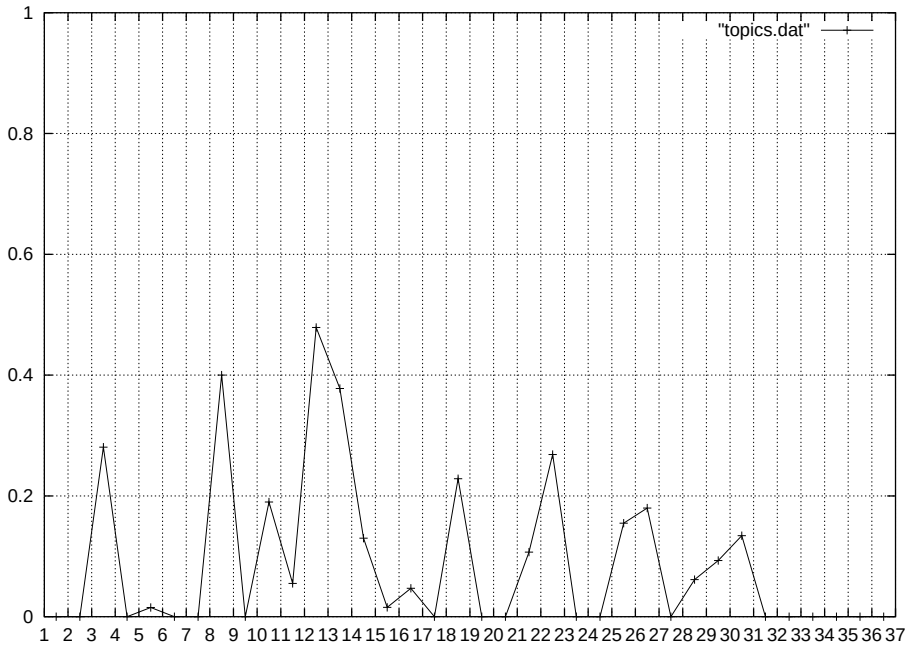


FIG. B.2 – Analyse thématique avec Anadia

ne s'applique qu'à certains types de textes particuliers, où la redondance terminologique est forte. De plus, elle nécessite une connaissance du domaine non négligeable, représentée par le dispositif Anadia.

Mais cette méthode bénéficie néanmoins d'un certain nombre de qualités. Tout d'abord, elle permet d'établir des relations entre les descripteurs possédant une réelle valeur sémantique. Cette valeur peut en outre être quantifiée assez finement en utilisant différentes notions de distance au sein d'un dispositif. Enfin, cette méthode a pour avantage de rester relativement pertinente pour le traitement des portions du texte qui ne se rapportent pas directement au domaine du dispositif, puisque l'analyse distributionnelle reste indépendante du champ thématique.

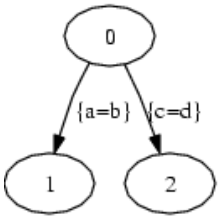
Annexe C

Déterminisation d'automates et structures de traits

Introduction

Nous présentons ici des algorithmes permettant de manipuler un cas particulier d'automates où l'alphabet est l'ensemble des structures de traits. Ce travail a été réalisé en collaboration avec Arnaud Soulet, et fait suite à (Soulet, 2002). La première particularité de notre alphabet Σ est qu'il n'est pas fini. La seconde, plus importante, réside dans la relation d'équivalence particulière que nous définissons entre les éléments de cet ensemble. Cette dernière interdit de déterminiser un tel automate avec l'algorithme standard.

En effet, nous considérons ici que le critère de déclenchement d'une transition est donné par l'ensemble des traits requis, sans exclure l'éventuelle présence d'autres traits. Ainsi, une structure σ_1 pourra déclencher une transition (q_1, σ_2, q_2) dés l'instant où σ_1 contient σ_2, ce qui n'exclut pas qu'elle puisse déclencher une transition alternative σ_3, même si $\sigma_2 \neq \sigma_3$. On pourra ainsi se trouver face à des configurations non-déterministes qui ne seraient pas résolues par l'algorithme standard. Par exemple :



Cet exemple présente un cas de non déterminisme car toute structure incluant $[a : b, c : d]$ pourrait déclencher l'une ou l'autre des deux transitions. Or, comme $[a : b] \neq [c : d]$, l'algorithme standard de déterminisation ne lèverait pas cette ambiguïté. La solution proposée ici est de procéder à une phase préliminaire permettant de lever ce cas particulier d'ambiguïté.

Représentation des transitions

Il s'agit en premier lieu de modifier la représentation des transitions : à la place d'un simple élément de Σ, le critère déclencheur d'une transition sera représenté par :

– Une structure de traits dite « positive », contenant les traits qui *doivent tous* être présents pour que la transition soit déclenchée.

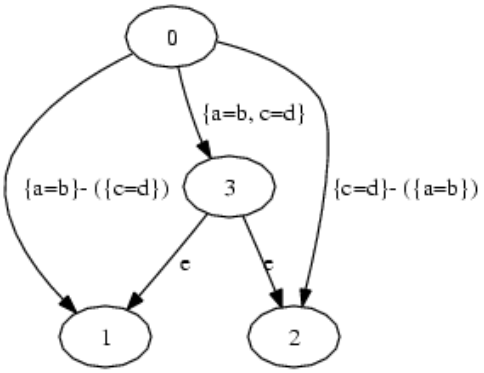
– Un ensemble de structures de traits dites « négatives », contenant les structures qui *ne doivent pas* être présentes pour que la transition soit déclenchée.

Ainsi, le critère de déclenchement d'une transition sera représenté par un couple de la forme $(p, \{n_1, n_2, ..., n_k\})$ où $p \in \Sigma$, $k \geq 0$, $i \in [0; k]$ et $n_i \in \Sigma$. Ce critère sera nommé « déclencheur », par opposition des lettres de l'alphabet Σ qui restent représentées par de simples structures de traits. Pour plus de commodité, ces déclencheurs seront par la suite désignés en utilisant une notation « objet », où la partie positive et la partie négative d'un déclencheur σ seront désignées respectivement par σ.pos et σ.neg.

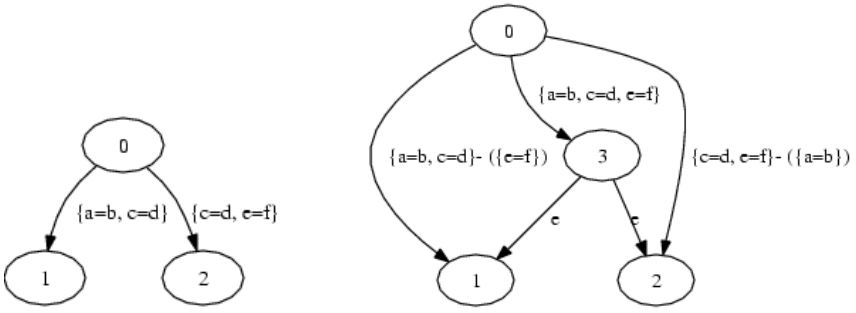
Il convient de noter que, la représentation des mots et des déclencheurs étant différentes, le choix d'une transition étant donnée une lettre d'un mot à reconnaître devra faire appel à une fonction d'équivalence particulière, qui sera donnée plus loin.

Méthode de pré-déterminisation

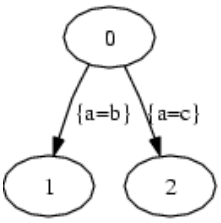
En s'appuyant sur cette représentation, on peut transformer les cas d'ambiguïté pour obtenir une nouvelle configuration qui puisse être déterminisée par l'algorithme standard. Pour l'exemple précédent, on générera l'automate suivant :



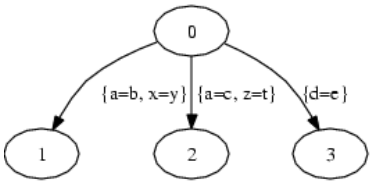
L'ambiguïté est ainsi levée puisque tout structure ne peut plus déclencher qu'une seule transition : soit on a à la fois $[a : b]$ et $[c : d]$, soit $[a : b]$ mais pas $[c : d]$, soit l'inverse. Une nouvelle ambiguïté a certes été introduite au niveau de l'état 3 (transitions nulles), mais celle-ci sera levée par l'algorithme standard. Pour mieux se convaincre du fonctionnement de l'algorithme, on pourra observer le cas suivant qui présente un cas où les deux alternatives partagent un même trait $[c : d]$ (l'automate prédéterminisé se trouve à droite) :



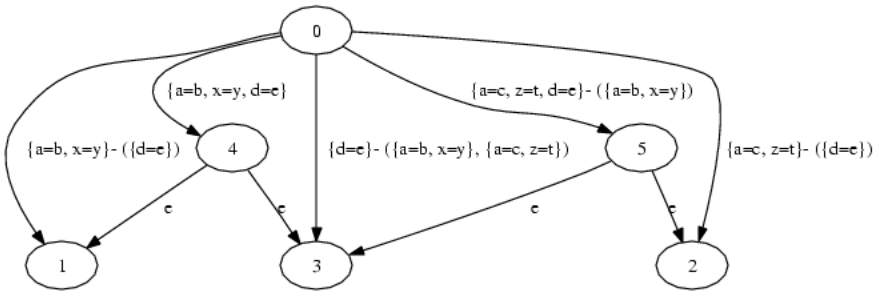
Il est évident que dans la configuration ci-dessous n'a pas lieu d'être déterminisée puisque les deux alternatives s'excluent mutuellement (de même, deux alternatives s'excluront mutuellement si la partie négative de l'une est incluse dans la partie positive de l'autre) :



On s'interrogera peut-être sur la nécessité de représenter la partie négative de chaque transition par un ensemble de structures de traits. En effet, une approche plus simple consisterait à utiliser une seule structure représentant l'ensemble des traits « négatifs ». Dans ce cas, la transition serait interdite à une entrée donnée dés l'instant où l'intersection de celle-ci avec la partie négative serait non vide. Mais nous allons observer sur l'exemple suivant que cette approche ne peut fonctionner :



Observons l'automate pré-déterminisé ci-dessous. La transition $0 \rightarrow 1$ ne doit pas se déclencher en présence de $[a : b, x : y]$ ni de $[a : c, z : t]$. En revanche, elle doit bien se déclencher en présence d'une structure du type $[d : e, a : b]$. Pour cette raison, on ne peut représenter la partie négative de cette transition simplement par l'ensemble « plat » que serait $[a : b, x : y, a : c, z : t]$. En effet, son intersection avec l'entrée $[d : e, a : b]$ serait égale à $[a : b]$ et donc non vide, ce qui interdirait le déclenchement de la transition.



Ajoutons que pour rendre le système plus flexible, on peut introduire une notion de priorité entre les transitions. Ainsi, on ne déterminisera que les ambiguïtés où les deux transitions sont de priorité égale. Dans tous les autres cas, la différence de niveau de priorité suffit à lever le non-déterminisme. Par la suite, la priorité d'un déclencheur σ sera notée σ.priorité.

La section suivante détaille l'algorithme de pré-déterminisation que nous venons de présenter, avant de rappeler l'algorithme de déterminisation standard qui pourra être appliqué à l'issue de cette phase préliminaire.

Phase préliminaire à la déterminisation

Comme nous l'avons explicité plus haut, nous présentons ici un algorithme permettant de pré-déterminiser un automate fonctionnant sur des structures de traits, afin de le rendre déterminisable par

l'algorithme standard (ce dernier sera rappelé plus loin). L'algorithme sera décomposé en plusieurs procédures. La première fonction retourne vrai si l'alternative entre les deux structures introduit un non-déterminisme. On vérifie tout d'abord qu'elles ne sont pas équivalentes. Si leur priorité est différente, le non-déterminisme est résolu d'office par le jeu des priorités (la priorité la plus haute sera choisie). Enfin, on teste qu'elles ne n'excluent pas mutuellement.

Algorithm 1 Fonction $Conflit(\sigma_1, \sigma_2)$

retourner $(\sigma_1 \neq \sigma_2) \wedge (priorite(\sigma_1) = priorite(\sigma_2)) \wedge \neg Exclusion(\sigma_1, \sigma_2)$

La fonction suivante retourne vrai si deux structures s'excluent mutuellement. C'est le cas si leur parties positives contiennent un même trait avec des valeurs différentes, ou si l'une la partie positive de l'une contient l'une des parties négatives de l'autre :

Algorithm 2 Fonction $Exclusion(\sigma_1, \sigma_2)$

retourner

$(\exists f_1 \in pos(\sigma_1), f_2 \in pos(\sigma_2) | nom(f_1) = nom(f_2) \wedge val(f_1) \neq val(f_2))$

$\vee (\exists f \in neg(\sigma_1) | pos(\sigma_2) \subseteq f)$

$\vee (\exists f \in neg(\sigma_2) | pos(\sigma_1) \subseteq f)$

L'algorithme en lui-même repose intégralement sur la procédure suivante, qui effectue un seul pas de la pré-déterminisation, en retournant faux si aucune modification n'a été effectuée (i.e. le traitement est terminé) :

Algorithm 3 Fonction $Predeterminise_{aux}(Q, \delta, q_0, F)$

for all $q \in Q$ **do**
  for all $((q, \sigma_1, q_1), (q, \sigma_2, q_2)) \in \delta \times \delta | \sigma_1 \neq \sigma_2 \wedge \sigma_1 \neq \varepsilon \wedge \sigma_2 \neq \varepsilon$ **do**
    if $Conflit(\sigma_1, \sigma_2)$ **then**
      if $pos(\sigma_1) = pos(\sigma_2)$ **then**
        $neg(\sigma_1) \leftarrow neg(\sigma_1) \cup neg(\sigma_2)$
        $neg(\sigma_2) \leftarrow neg(\sigma_2) \cup neg(\sigma_1)$
      else
        $Q \leftarrow Q \cup q_3$
        $pos(\sigma_3) \leftarrow pos(\sigma_1) \cup pos(\sigma_2)$
        $neg(\sigma_3) \leftarrow \{neg(\sigma_1), neg(\sigma_2)\}$
        $priorite(\sigma_3) \leftarrow priorite(\sigma_1)$
        $\delta \leftarrow \delta \cup \{(q, \sigma_3, q_3), (q_3, \varepsilon, q_1), (q_3, \varepsilon, q_2)\}$
        $neg(\sigma_1) \leftarrow pos(\sigma_2) - pos(\sigma_1)$
        $neg(\sigma_2) \leftarrow pos(\sigma_1) - pos(\sigma_2)$
      end if
      retourner VRAI
    end if
  end for
end for
retourner FAUX

Enfin, l'algorithme principal consiste simplement en la répétition du pas de pré-déterminisation tant qu'une modification a été effectuée au pas précédent :

Algorithm 4 Procédure $Predeterminise\,(Q, \delta, q_0, F)$

$b \leftarrow$VRAI
while b **do**
  $b \leftarrow Predeterminise_{aux}\,(Q, \delta, q_0, F)$
end while

Algorithme de déterminisation standard

A la suite de la phase présentée dans la section précédente, l'algorithme standard peut être appliqué. Notons que comme dans notre cas l'alphabet est infini, l'implémentation effective de la boucle sur Σ nécessite une passe préliminaire permettant de déterminer l'alphabet effectivement composé de l'ensemble des déclencheurs des transitions partant de l'un des états compris dans l'état courant q'.

Algorithm 5 Fonction $Determinise_{aux}\,(\Sigma, Q, \delta, q_0, F)$

$\delta' \leftarrow \varnothing$
$q_0 \leftarrow \{q_0\} \cup \{q \in Q | \,(q_0, \varepsilon, q) \in \delta\}$
$Q' \leftarrow \left\{q'_0\right\}$
for all $q' \in Q'$non encore considéré **do**
  for all $\sigma \in \Sigma - \{\varepsilon\}$ **do**
    $q'' \leftarrow \{q_2 \in Q | \,(\exists q_1 \in q' | \,(q_1, \sigma, q_2) \in \delta)\}$
    if $q'' \neq \varnothing$ **then**
      for all $q_1 \in q''$non encore considéré **do**
        $q'' \leftarrow q'' \cup \{q_2 \in Q | \,(q_1, \varepsilon, q_2) \in \delta\}$
      end for
      $\delta' \leftarrow \delta' \cup \{(q', \sigma, q'')\}$
      $Q' \leftarrow Q' \cup \{q''\}$
    end if
  end for
end for
$F' \leftarrow \{q' \in Q' | q' \cap F \neq \varnothing\}$
retourner $(\Sigma, Q', \delta', q_0,', F')$

Aspect dynamique

Une fois l'automate déterminisé, il peut peut être utilisé de façon standard pour tester si un mot est accepté (c'est-à-dire si une séquence de structures de traits correspond à un chemin de l'état initial à un état final). Il convient toutefois d'appliquer une procédure particulière pour tester si une structure donnée permet de déclencher une transition donnée. Cette procédure doit en effet être cohérente avec le procédé de pré-déterminisation et la représentation choisie pour les transitions (sous la forme d'une structure « positive » et d'un ensemble de structures « négatives ». Pour qu'une structure s puisse déclencher une transition σ, il faut qu'elle contienne la partie positive de σ, sans contenir aucune de ses parties négatives :

Algorithm 6 Fonction $Match\,(s, \sigma)$

retourner $(s \subseteq pos\,(\sigma)) \wedge \neg\,(\exists f \in neg\,(\sigma)\,| s \subseteq f)$

Annexe D

Notes

D.1 Terminologie et analyse thématique

En se positionnant par définition entre forme et concept, le terme constitue, comme nous l'avons vu à plusieurs reprises, un objet important dans le domaine de l'analyse thématique. (Rastier, 1995b) pose un regard critique sur certains aspects de la pratique terminologique, et discute ses liens avec l'analyse thématique et la représentation des connaissances en intelligence artificielle.

Dans le courant de la terminologie théorisé au début du XXe siècle par Eugen Wüster (représentant de l'Ecole de Vienne) et D.S. Lotte (créateur de l'Ecole Soviétique), le terme est défini comme une forme permettant de désigner une notion (ou concept) de façon univoque et non contextuelle. Felbert définit le terme comme « un symbole conventionnel représentant une notion définie dans un certain domaine du savoir ». La tradition terminologique pose donc le terme comme doté d'une signification autonome, réalisant la dénotation parfaite, que l'insertion dans la proposition peut seulement influencer sans participer à la définir. En s'éloignant du principe sémiotique de la détermination de la chose sur le mot, le terme se place ainsi dans un système (qu'on peut qualifier d'idéaliste) où la référence est stable et absolue, et où signifié et concept tendent à se confondre. Ce point de vue trouve bien sûr sa justification dans la notion de langue de spécialité, où les diverses ambiguïtés propres au langage naturel tendent à s'estomper.

Une caractéristique forte de la vision traditionnelle du terme est son approche purement nominaliste. La désignation en effet est un rôle attribué aux noms, que la norme ISO « Terminologie » définit d'ailleurs comme la « désignation d'un objet par une unité linguistique ». Et la première génération des systèmes automatiques d'extraction de termes (comme LEXTER (Bourigault, 1992)) ne prenaient en effet en compte que les syntagmes nominaux (des systèmes plus récents comme celui basé sur Syntex (Bourigault et Fabre, 2000) lèvent toutefois cette limitation).

Rastier décrit ainsi le procédé de la terminologie normative permettant au mot d'accéder au statut de terme :

- La *nominalisation* permet de d'obtenir une forme canonique dite substantive (puisque le nom est censé représenter la « substance » des choses).
- La *lemmatisation* permet de se débarrasser des variations de graphies des mots en se basant sur une forme canonique (masculin singulier pour les noms).
- La *décontextualisation* vise à donner une définition du terme indépendamment du contexte linguistique de ses occurrences.
- La constitution du mot en type (*abstraction*), produite par sa définition, permet de le dégager de ses occurrences particulières, séparant la notion de l'objet individuel.

Le terme ainsi construit se soustrait à l'interprétation, et permet alors d'exprimer univoquement

un concept, que Rastier décrit alors ici comme « le signifié d'un mot dont on décide de négliger la dimension linguistique », ou encore comme « le produit de l'instauration du terme ».

La critique par Rastier de cet aspect normatif de la signification du terme repose sur le fait qu'elle suppose que le sens d'un signe puisse exister indépendamment de son interprétation en contexte. Rastier considère au contraire qu'« un signe en tant que tel ne peut être interprété, puisque l'isoler le coupe précisément de ses conditions d'interprétation, de son contexte, c'est-à-dire du texte ». Il prône donc une approche interprétative de la terminologie, où le mot se définit par rapport à son contexte et au texte, et reconsidérant le système de la constitution des termes (par exemple en étudiant le rapport entre termes et non-termes). Enfin, il propose de relier la sémantique lexicale à la sémantique textuelle, en rapprochant la notion de terme de celle de « molécule sémique » (cf. section 3.3.2).

D.2 Liens entre RST et GST

Bien qu'initialement motivées par des objectifs distincts (cf. sections 3.4.4 et 3.4.3), ces deux théories peuvent se montrer complémentaires. En effet, différents travaux comme (Moore et Pollack, 1992) et (Moore et Paris, 1993) montrent l'intérêt d'intégrer la dimension intentionnelle au sein du modèle RST. A cette fin, Moser et Moore proposent dans (Moser et Moore, 1996) un modèle basé sur la notion de *structure linguistique intentionnelle*, décrivant comment les intentions du locuteur déterminent la structure du discours[1].

La mise en oeuvre de ces modèles en traitement automatique de la langue concerne majoritairement le modèle RST appliqué à la génération automatique de texte. D'autres travaux abordent toutefois le problème de l'analyse automatique du discours, visant la structuration des textes conformément aux modèles évoqués plus haut. Dans ce domaine, on peut principalement citer les divers travaux de Marcu, qui propose tout d'abord dans (Marcu, 1996) une formalisation de RST en logique du premier ordre[2]. Dans (Marcu, 2000), cette formalisation est étendue afin de prendre en compte la dimension intentionnelle de GST, en se basant sur (Moser et Moore, 1996). Cette dernière est cependant difficilement applicable concrètement en TAL, car elle suppose l'existence d'une « fonction intentionnelle » capable de traduire les intentions du locuteur en objets logiques du premier ordre.

D.3 Les adverbes de phrase

La classe des adverbes définit un ensemble dont les contours sont relativement difficiles à définir. Faute de mieux, on peut définir un adverbe comme un terme invariable sans pour autant appartenir aux classes des prépositions, déterminants ou conjonctions. Par exemple, « lentement » ou « beaucoup ». Un syntagme adverbial a pour tête un adverbe qui peut être suivi d'un complément (par ex. « conformément à X ») et/ou précédé d'un autre adverbe (par ex. « très lentement »).

Un syntagme adverbial peut modifier un verbe, et constitue alors un complément circonstanciel, par ex. « X parle souvent à Y ». Il peut également modifier un syntagme prépositionnel, par ex. « immédiatement après son départ ».

Le syntagme adverbial peut également porter sur l'ensemble d'une proposition ou d'une phrase. Dans ce cas, il s'agit d'un *adverbe de phrase*. Celui-ci peut avoir un emploi « scénique », en précisant le

[1] Le concept de structure linguistique intentionnelle, parfaitement explicite dans GST, n'est présent qu'implicitement dans RST, au sein de la relation noyau-satellite. Le lien entre les deux théories reposerait donc dans l'équivalence entre les relations de nucléarité de RST et de dominance de GST.

[2] Cette formalisation est ensuite utilisée pour élaborer un algorithme permettant de construire automatiquement l'ensemble des arbres rhétoriques valides à partir de la donnée des relations existant entre les segments (l'analyse du discours doit donc être préalablement amorcée).

cadre notionnel ou spatio-temporel au sein duquel doit être interprétée la proposition ou la phrase. Par exemple : « Ici, on apprécie beaucoup les plats épicés ». Un adverbial de phrase peut également avoir une fonction modale, en précisant le degré de réalité que le locuteur assigne au contenu propositionnel du reste de la phrase, ou l'évaluation qu'il en fait. Par exemple : « Très probablement, Jean sera absent ». Il peut également s'agir d'un connecteur de discours, par exemple : « Corrélativement, nous devons considérer que X ».

Dans le modèle de l'encadrement du discours, certains adverbes dits « de phrase » pourront en fait avoir une portée plus large, s'étendant sur plusieurs phrases ou propositions (cf. section 3.4.1).

Annexe E

Règles et scripts LinguaStream

Dans cette section sont regroupés les divers scripts, grammaires ou règles LinguaStream que nous avons eu l'occasion d'évoquer dans ce manuscrit. Ils se réfèrent tous à un formalisme (ou outil) particulier, désignés par les acronymes suivants :

CDML Constraint-based Discourse Modeling Language (formalisme de description de structures discursives par contraintes).

EDCG Extended Definite Clause Grammar (grammaire locale d'unification).

Groovy Script Groovy/EBMS.

LSA LinguaStream Automation (scripts d'automatisation).

MRE Macro-Regular-Expressions (transducteurs exprimés sous la forme de macro-expressions régulières).

XSD XML Schema.

E.1 Extraction de syntagmes nominaux complexes – EDCG

```
syntagmeNominal(type:Type..length:Length..key:Key) -->
    introSyntagme(Type, VP, Key1),
    suiteSyntagme(VP, Length, Key2).

introSyntagme(Type, nvp) --> @tag:det..stag:Type.
introSyntagme('def', nvp) --> @tag:pre..stag:det.
introSyntagme('nom', vp) --> @tag:nom..lemma:L.

suiteSyntagme(_, Length, Key) -->
    suiteSyntagmeAux(L1, Key1), !,
    suiteSyntagme(vp, L2, Key2).

suiteSyntagme(vp, 0) --> [].

suiteSyntagmeAux(0) --> @tag:adj..lemma:L.
suiteSyntagmeAux(0) --> pre, @tag:det.
suiteSyntagmeAux(1) --> pre, @tag:nom..lemma:L.
suiteSyntagmeAux(0) --> @tag:det, @tag:adv..lemma:L.
suiteSyntagmeAux(1) --> @tag:con..stag:coo), @tag:nom..lemma:L.
```

```
suiteSyntagmeAux(0) --> @tag:con..stag:coo), @tag:adj..lemma:L.
suiteSyntagmeAux(1) --> @tag:nom..lemma:L.
suiteSyntagmeAux(0) --> @tag:det.

pre --> @lemma:'de'.
pre --> @lemma:'de+le'.
pre --> @lemma:'à+le'.
```

E.2 Extraction d'expressions temporelles – EDCG

```
exprTemp(X) --> periode(X), !.
exprTemp(X) --> ponctuel('O', X), !.


periode(periode:(type:df..debut:D..fin:F)) -->
    $DD, ponctuel(_, D),
    $DF, ponctuel(_, F),
    { delimiteursPeriode(DD, DF) }.

periode(periode:(type:d..debut:D)) --> debutPeriode, ponctuel(_, D).
periode(periode:(type:f..fin:F))   --> finPeriode, ponctuel(_, F).

periode(periode:(type:f..duree:D..grain:1)) -->
    duree, $Tmp, $'ans',
    { atom_to_int(Tmp,Q), number(Q), D is Q }.

periode(periode:(type:f..duree:D..grain:1)) -->
    duree, $Tmp, $'ans',
    { nombre(Tmp,Q), D is Q }.

periode(periode:(type:df..duree:D)) --> duree, ponctuel(_, D).


ponctuel('O', R) -->
    $Mot, ponctuel('N', R),
    { prefixePonctuel(Mot) }.

ponctuel('O', approx:A) -->
    ls_token(_, lemma:vers), ponctuel(_, A).

ponctuel('O', numeral:(quantif:Q..unite:U)) -->
    $Tmp, $X, ponctuelPeriode(V1),
    { atom_to_int(Tmp,Q), number(Q),numer(X,V2), U is V1*V2 }.

ponctuel('O', periode:(type:df..duree:Q..grain:U)) -->
    $N, $X, ponctuelPeriode(V1),
    {nombre(N,Q),numer(X,V2), U is V1*V2}.
```

```
ponctuel('N', D) --> datation(D).

ponctuel('N', periode:(type:f..duree:Q..grain:U)) -->
    $Tmp, $X,
    { atom_to_int(Tmp,Q),number(Q),periode(X,U) }.

ponctuel('N', periode:(type:f..duree:Q..grain:U)) -->
    $Tmp, $X,
    { nombre(Tmp,Q),periode(X,U) }.

ponctuel(_, contemp:oui) --> contemporain.

ponctuelPeriode(V) --> $'d\'', $X, { periode(X,V) }.
ponctuelPeriode(V) --> $'de', $X, { periode(X,V) }.
ponctuelPeriode(V) --> $'des', $X, { periode(X,V) }.

datation(type:enum..debut:P..suite:P2) -->
    precis(P), $X,
    datation(P2), { enumeration(X) }.

datation(P) --> precis(P), !.
datation(F) --> flou(F).

flou(S) --> annees(S).
flou(S) --> siecle(S).

siecle(siecle:(partie:R..id:contemp)) -->
    relatif(R), $'du', $'siècle'.


annees(enum:(debut:annees:(type:global..annee:A)..suite:P)) -->
    $'années', anneeSouple(A), $X,
    suiteAnnees(P), { enumeration(X) }.

annees(annees:(type:global..annee:A)) -->
    $'années', anneeSouple(A).

annees(annees:(type:R..annee:A)) -->
    relatif(R), $'des',
    $'années', anneeSouple(A).

annees(annees:(type:contemporain)) -->
    $'dernières', $'années'.

suiteAnnees(enum:(debut:annees:(type:global..annee:A)..suite:P)) -->
    anneeSouple(A), $X,
    suiteAnnees(P), { enumeration(X) }.

suiteAnnees(annees:(type:global..annee:A)) -->
    anneeSouple(A).
```

```
precis(jour:J..mois:M..annee:A) --> jour(J), mois(M), annee(A), !.
precis(mois:M..annee:A) --> mois(M), annee(A), !.
precis(jour:J..mois:M) --> jour(J), mois(M).
precis(annee:A) --> annee(A).
precis(mois:M) --> mois(M).
precis(annee:A) --> $'l\'', $'an', annee(A).
precis(annee:A) --> $'l\'', $'année', annee(A).

jour(N) --> $X, { atom_to_int(X, N), N>0, N<31 }.

mois(N) --> $X, { nomMois(X, N) }.

annee(N) --> $X, { atom_to_int(X, N), N>=1900, N=<2100 }.

anneeSouple(N) -->
    $X, { atom_to_int(X, N), N>0, atom_length(X, L), member(L, [2, 4]) }.
```

E.3 Recherche Google avec extraction des hypertermes – LSA

```
query = automation.getUserInput("Query:");

print("Google query: " + query);
google = automation.getGoogle("xxx");
google.maxResults = 3;
google.languages = "lang_fr";
results = google.search(query);

stream = automation.openStream("analysis.ls");
for(r : results)
{
    outputFile = automation.createTemporaryFile("png");
    r.userData = outputFile;
    stream.setParameter("input", r.URL);
    stream.setParameter("output", outputFile.absolutePath);
    print("Processing: " + r.URL +  " to " + outputFile);
    stream.run();
}

outputFile = automation.createTemporaryFile("xml");
print("Search results saved as " + outputFile);
xml = automation.getXMLWriter(outputFile);

xml.startDocument();
xml.startElement("search");
xml.textElement("query", query);
for(r : results)
{
```

```
        xml.startElement("result");
        xml.textElement("title", r.title);
        xml.textElement("snippet", r.snippet);
        xml.textElement("url", r.URL);
        xml.textElement("hyperterms", r.userData.toString());
        xml.endElement("result");
    }
    xml.endElement("search");
    xml.endDocument();

    stream = automation.openStream("viewSearchResults.ls");
    stream.setParameter("input", outputFile.absolutePath);
    stream.run();
```

E.4 Portée des introducteurs de cadres temporels (version scriptée) – Groovy

```
import fr.unicaen.util.Period;
import fr.unicaen.linguastream.tfs.FeatureSet;
import fr.unicaen.linguastream.markup.ebms.Token;
import fr.unicaen.linguastream.markup.ebms.Anchor;
import fr.unicaen.linguastream.scripting.ScriptingDocumentHandler;
import fr.unicaen.geosem.phenoloc.semantics.TemporalSemanticsMatcher;

class FramingDocumentHandler extends ScriptingDocumentHandler
{
    /* Token */ lastSentenceStart = null;
    /* Token */ lastSentenceEnd = null;
    /* Token */ frameStart = null;
    /* Token */ introducer = null;
    /* String */ currentTense = null;
    /* SemanticsMatcher */ temporalMatcher = null;

    void startDocument()
    {
        current = new Period(1994, 1994);
        temporalMatcher = new TemporalSemanticsMatcher(current);
    }

    void startAnchor(Anchor anchor)
    {
        if(anchor.type == "sentence")
            lastSentenceStart = anchor;
    }

    void endAnchor(Anchor anchor)
    {
        if(anchor.type == "sentence")
```

```
            lastSentenceEnd = anchor;
        else if(anchor.type == "paragraph")
            closeFrame("paragraph", "structure");
    }

    void token(Token token)
    {
        if(introducer == null)
        {
            if(token.type == "intro")
            {
                introducer = token;
                frameStart = lastSentenceStart;
            }
        }
        else
        {
            if(token.type == "intro")
            {
                closeFrame("new frame", "newFrame");
                introducer = token;
                frameStart = lastSentenceStart;
            }
            else if(token.type == "tm")
            {
                tense = token.getFeature("temps");
                if(currentTense == null)
                    currentTense = tense
                else if(currentTense != tense)
                    closeFrame("tenses", "tenses");
            }
            else if(token.type == "temporal")
            {
                FeatureSet introSem =
                    introducer.featureSet.getFeature("sem").getNodeValue();
                if(temporalMatcher.match(introSem, token.featureSet) == 0.0)
                    closeFrame("semantics", "semantics");
            }
        }
    }

    void closeFrame(String reason, String closureCriterion)
    {
        if(introducer == null)
            return;
        if(lastSentenceEnd != null && lastSentenceEnd.isAfter(frameStart))
        {
            log("Frame closure, cause=" + reason);
            fs = introducer.featureSet.clone();
            fs.addFeature("closureCriterion", closureCriterion);
```

```
                createMarkup(frameStart, lastSentenceEnd, "frame", 50, true, fs);
            }
            else
            {
                log("Ignoring sub-sentence frame (removing introducer)");
                deleteMarkup(introducer);
            }
            frameStart = null;
            introducer = null;
            currentTense = null;
        }
    }
```

E.5 Schéma « LinguaStream Document » – XSD

```
<?xml version="1.0">

<schema xmlns="http://www.w3.org/2001/XMLSchema"
        targetNamespace="http://www.linguastream.org/LSD/2.0">

    <element name="a">
        <annotation>
            <documentation>LinguaStream anchor-type markup</documentation>
        </annotation>
        <complexType>
            <attributeGroup ref="markupAttributes"/>
            <attribute name="anchor" type="anchorType" use="required" />
        </complexType>
    </element>

    <element name="b">
        <annotation>
            <documentation>LinguaStream block-type markup</documentation>
        </annotation>
        <complexType mixed="true">
            <sequence>
                <any minOccurs="0" maxOccurs="unbounded"/>
            </sequence>
            <attributeGroup ref="markupAttributes"/>
        </complexType>
    </element>

    <attributeGroup name="markupAttributes">
        <attribute name="type" type="token" use="required"/>
        <attribute name="id" type="positiveInteger" use="required"/>
        <attribute name="layer" type="integer" use="optional" default="0"/>
    </attributeGroup>
```

```
        <simpleType name="anchorType">
            <restriction base="NMTOKEN">
                <enumeration value="start"/>
                <enumeration value="end"/>
            </restriction>
        </simpleType>

    </schema>
```

Annexe F

Divers

F.1 Sources des extraits de corpus

AF D. Finley, *Autour d'un fleuve*. Le Niger, éditions Gamma.

AFM Site de l'Association Française contre les Myopathies du département de l'Isère, dossier intitulé « L'intégration scolaire ». `http://members.aol.com/delegation38`

AFT Site de l'Ambassade de France en Thaïlande, section consacrée aux « Relations culturelles et de coopération ». `http://www.ambafrance-th.org`

APH M. Oria et J. Raffin, *Anatomie, Physiologie, Hygiène*. Hatier, 1973.

ARL O. Kirch et T. Dawson, *Administration réseau sous Linux*. O'Reilly, 2001.

ATM *Atlas Transmanche Électronique*. `http://atlas-transmanche.certic.unicaen.fr`

BEA Bulletins *REC Info* du Bureau d'Enquête et Analyses (BEA) pour la sécurité de l'aviation civile. `http://www.bea-fr.org`

EC Étude de « Énergie-Cités », association des autorités locales européennes pour une politique énergétique locale durable. Rapport sur « Lüchow-Dannenberg ». `http://www.energie-cites.org`

KENYA M. Carle, *Au Kenya, Guide Bleu*. Hachette, 1976.

HER R. Hérin, R. Rouault, V. Veshambre, *Atlas de la France scolaire De la maternelle au lycée*, Dynamiques du territoire, Reclus, 1994.

LM Journal *Le Monde*.

LMD Journal *Le Monde Diplomatique*.

LR Revue *La Recherche*.

MB G. Flaubert, *Madame Bovary*.

REC Joël Bockaert, Laurent Fagni et Jean-Philippe Pin, « Des récepteurs activés de l'intérieur », *CNRS-Info* no 395.

ROU Synthèse d'un rapport de la communauté européenne sur les transports en Roumanie.

SA Revue *Scientific American*.

SAL Texte recueilli sur une vignette jointe à une plaque de chocolat, reproduit et analysé dans (Charolles, 1997).

SIL Sébastien Ferré, *Systèmes d'information logiques, un paradigme logico-contextuel pour interroger, naviguer et apprendre*, thèse de doctorat.

VE M. Casse, *La vie des étoiles*. Nathan, 1985.

YR1 Journal *L'Yonne Républicaine* du 10 mai 2003.

WP1 Encyclopédie Wikipedia, article « Technologie ». `http://fr.wikipedia.org/wiki/Technologie`

F.2 Analyse RST du « two frameworks text »

Voir figure F.1.

F.3 Architecture logicielle du moteur de recherche « sémantique et multi-dimensionnel »

Voir figure F.2.

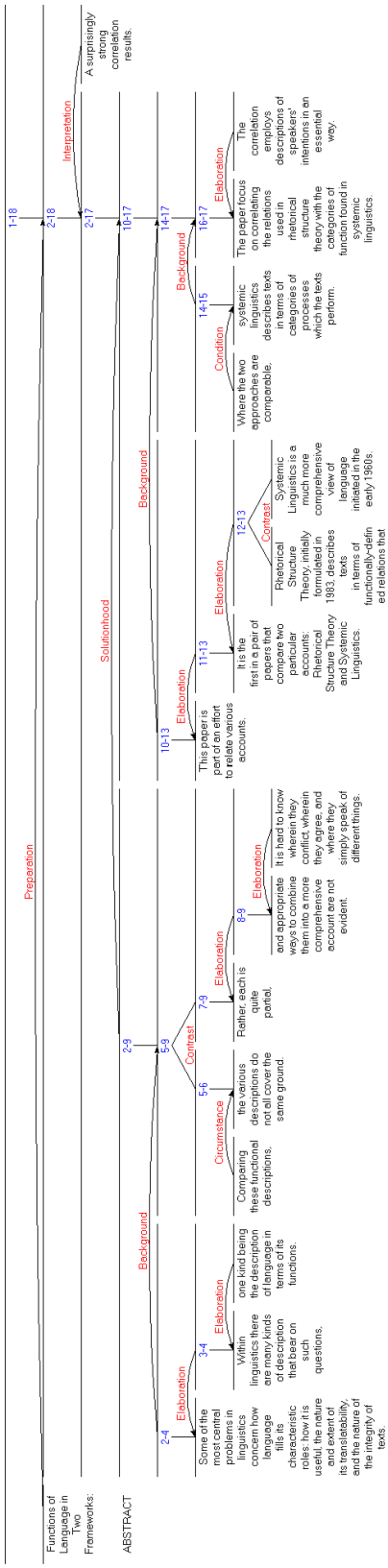


FIG. F.1 – Analyse RST du « two frameworks text »

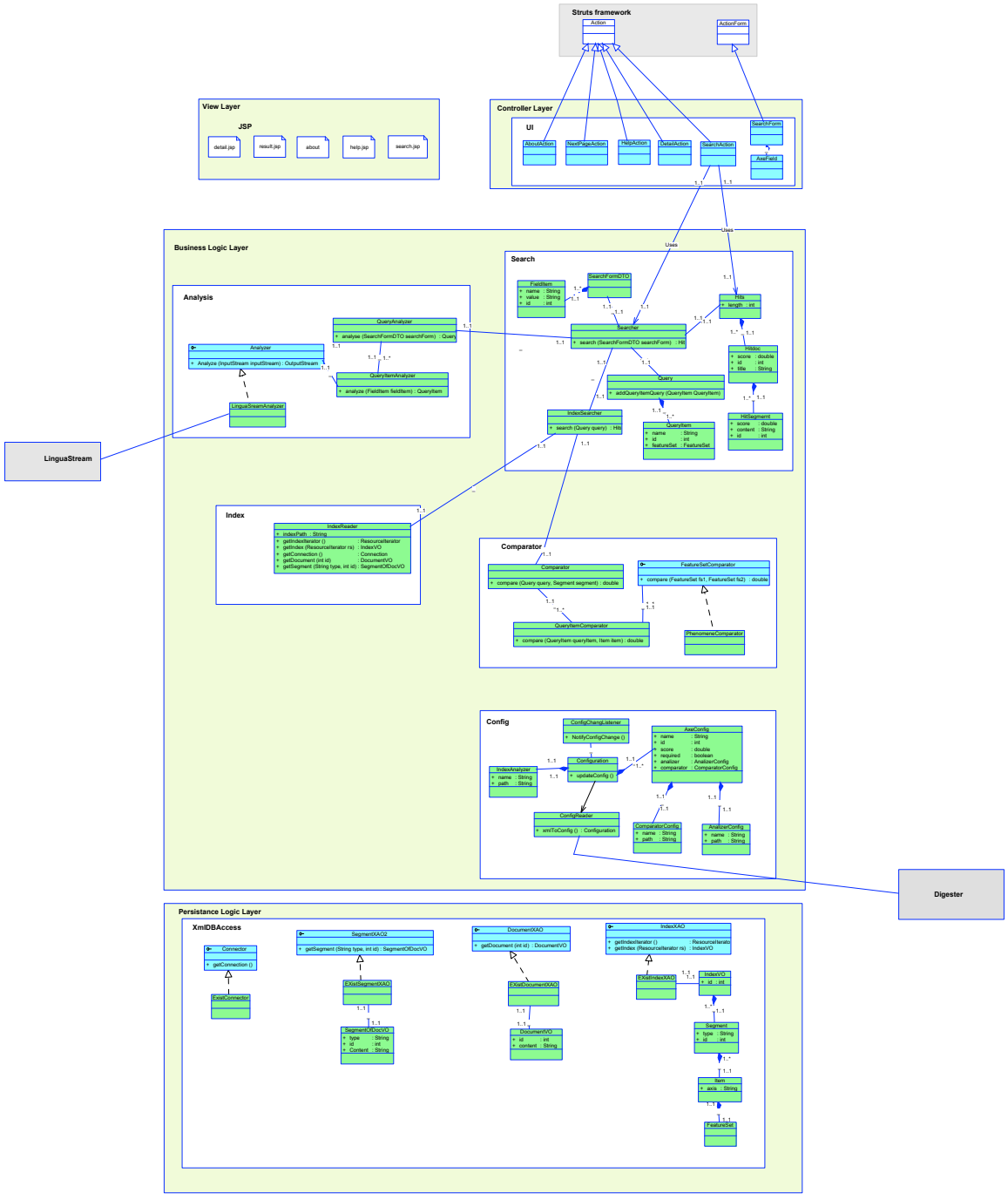


FIG. F.2 – Architecture logicielle du moteur de recherche « sémantique et multi-dimensionnel ». Extrait de (Benallel, 2005).

Index

Table des figures

Bibliographie

Bachta, A. (2002). *L'espace et le temps chez Newton et chez Kant, Essai d'explication de l'idéalisme kantien à partir de Newton*. Paris, L'Harmattan.

Beaudet, S. (2002). Extraction et analyse sémantique automatiques des entités géo-référencées. Mémoire de Maîtrise, Université de Caen, France.

Benallel, D. (2005). Réalisation d'un moteur de recherche "sémantique et multi-axial". Mémoire de D.E.S.S., Université de Caen, France.

Bennai, N. (2003). Application des standards XML à l'interrogation de bases sémantiques. Mémoire de D.E.S.S., Université de Caen, France.

Beust, P. (1998). *Contribution à un modèle interactionniste du sens*. Thèse de doctorat, Université de Caen, France.

Bilhaut, F. (2002). Identification et localisation spatio-temporelle des "phénomènes" dans les textes géographiques. Mémoire de D.E.A., Université de Caen, France.

Bilhaut, F. (2005). Composite Topics in Discourse. *In Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP'05)*, pages 107–111, Borovets, Bulgaria.

Bilhaut, F., Charnois, T., Enjalbert, P. et Mathet, Y. (2003a). Passage Extraction in Geographical Documents. *In Proceedings of New Trends in Intelligent Information Processing and Web Mining (IIPWM'03)*, pages 121–130, Zakopane, Pologne.

Bilhaut, F. et Enjalbert, P. (2005a). Discourse Thematic Organisation Reveals Domain Knowledge Structure. *In Proceedings of the 2nd Indian International Conference on Artificial Intelligence (IICAI'05)*, pages 2815–2831, Pune, India.

Bilhaut, F. et Enjalbert, P. (2005b). Recherche d'information géographique. *In* P. Enjalbert, éditeur : *Sémantique et traitement automatique des langues naturelles*, pages 371–406. Hermès Sciences.

Bilhaut, F., Ho Dac, L.-M., Borillo, A., Charnois, T., Enjalbert, P., Le Draoulec, A., Mathet, Y., Miguet, H., Péry-woodley, M.-P. et Sarda, L. (2003b). Indexation discursive pour la navigation intradocumentaire : cadres temporels et spatiaux dans l'information géographique. *In Actes de la 10e Conférence Traitement Automatique du Langage Naturel (TALN'03)*, pages 315–320, Batz-sur-Mer, France.

Bilhaut, F. et Widlöcher, A. (2006). LinguaStream : An Integrated Environment for Computational Linguistics Experimentation. *In Proceedings of the 11th Conference of the European Chapter of the Association of Computational Linguistics (EACL'06)*, Trento, Italy.

Bilhaut, F., Widlöcher, A., Laignelet, M. et Pimm, C. (2006). *LinguaStream User Manual and Developer Guide*.

Boguarev, B. et Kennedy, C. (1997). Saliance-Based Content Characterisation of Text Documents. *In Proceedings of the ACL/EACL Workshop on Intelligent Scalable Text Summarization*, pages 2–9, Madrid, Espagne.

Bourigault, D. (1992). LEXTER, un logiciel d'extraction de terminologie. *In Actes du symposium TAMA 92*, Avignon, France.

Bourigault, D. et Fabre, C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de grammaire*, 25:131–151.

Bradsford, J. et Johnson, M. (1972). Contextual prerequisites for understanding : Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, 11:717–726.

Bray, T., Hollander, D. et Layman, A. (1999). *Namespaces in XML*. Recommandation W3C.

Brooks, T. A. (1995). People, Words, and Perceptions : A Phenomenological Investigation of Textuality. *Journal of the American Society for Information Science*, 46(2):103–115.

Brown, G. et Yule, G. (1983). *Discourse Analysis*. Cambridge University Press.

Bruza, P. D., Song, D. W. et Wong, K. F. (2000). Aboutness from a commonsense perspective. *Journal of the American Society for Information Science*, 51(12):1090–1105.

Carbonell, J., Harman, D., Hovy, E., Maiorano, S., Prange, J. et Sparck-Jones, K. (2000). Vision Statement to guide Research in Question Answering and Text Summarisation. Rapport technique.

Chafe, W. L. (1976). Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View. *In* Li C. N., éditeur : *Subject and Topic*, pages 25–55. New York : Academic Press.

Chafe, W. L. (2001). The Analysis of Discourse Flow. *In* Schiffrin D., Tannen D. and Hamilton E. H., éditeur : *The Handbook of Discourse Analysis*, pages 673–687. Oxford : Blackwell.

Charnois, T. et Enjalbert, P. (2005). Compréhension automatique. *In* P. Enjalbert, éditeur : *Sémantique et traitement automatique du langage naturel*, pages 267–308.

Charolles, M. (1997). L'encadrement du dicours - Univers, champs, domaines et espaces. *Cahiers de recherche linguistique*, 6.

Charolles, M. (2003). De la topicalité des adverbiaux détachés en tête de phrase. *In* M. Charolles et S. Prévost, éditeur : *Adverbiaux et topiques*, volume 47, pages 11–51. Louvain la Neuve.

Choi, F. Y. (2000). Advances in domain independant linear text segmentation. *In Proceedings of NAACL*, volume 6, Seattle, USA.

Christensen, F. (1967). A Generative Rhetoric of the Paragraph. *Notes Towards a New Rhetoric : Six Essays for Teachers*, pages 52–81.

Ciravegna, F., Lavelli, A., Mana, N., Matiasek, J., Gilardoni, L., Mazza, S., Ferraro, M., Black, W. J., Rinaldi, F. et Mowatt, D. (1999). FACILE : Classifying Texts Integrating Pattern Matching and Information Extraction. *In Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI'99)*, Stockholm, Scandinavie.

Clark, H. et Haviland, S. (1977). Comprehention and the Given-New Contrast. *In* Freedle R., éditeur : *Discourse Production and Comprehension*. Lawrence Erlbaum Associates.

Combettes, B. et Tomassone, R. (1988). *Le texte informatif - Aspects linguistiques*. Bruxelles : De Boeck-Wesmael.

Courdille, T. (2005). Mise en place d'outils d'évaluation sous LinguaStream. Mémoire de D.E.S.S., Université de Caen, France.

Couto, J., Olivier, F., Brigitte, G., Nicolas, H., Agatha, J., Jean-Luc, M. et Sylvie, P. (2004). Régal, un système pour la visualisation sélective de documents. *RIA, La présentation d'information sur mesure*, numéro spécial:481–514.

Covington, M. A. (1994). GULP 3.1 : An Extension of Prolog for Unification-Based Grammar.

Crestan, E., De Loupy, C. et Manigot, L. (2004). Analyses sémantiques pour la navigation textuelle. *In* P. Enjalbert et M. Gaio, éditeur : *Approches sémantiques du document numérique*.

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge University Press.

Cunningham, H. (2000). *Software Architecture for Language Engineering*. Thèse de doctorat, University of Sheffield, UK.

Cunningham, H., Maynard, D., Bontcheva, K. et Tablan, V. (2002). GATE : A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, volume 6, Philadelphia, USA.

Dias, G. et Alves, E. (2005). Discovering Topic Boundaries for Text Summarization Based on Word Co-occurrence. *In Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP'05)*, pages 187–190, Borovets, Bulgaria.

Dik, S. C. (1989). *The Theory of Functional Grammar*. Dordrecht : Foris Publications.

Dik, S. C., Hoffmann, M. E., de Long, J. R., Djiang, S. I., Stroomer, H. et Devries, L. (1981). On the typology of focus phenomena. *In* Hoekstra T., van der Hulst H. et Moortgat M., éditeur : *Perspectives on Functional Grammar*, pages 41–74.

Downing, A. (1991). An alternative approach to theme : A systemic-functional perspective. *Word*, 42:119–143.

Dupont, M. (2003). *Une approche cognitive du calcul de la référence*. Thèse de doctorat, Université de Caen, France.

Déclès, J.-P. et Minel, J.-L. (2000). Résumé automatique et filtrage sémantique de textes. *In* Pierrel, J., éditeur : *Ingénierie des langues*, pages 253–270. Hermes Sciences, Lavoisier.

Enjalbert, P. (2005a). L'extraction d'information. *In* P. Enjalbert, éditeur : *Sémantique et traitement automatique du langage naturel*, pages 309–332.

Enjalbert, P. (2005b). Projet GéoSem : Analyse sémantique de documents géographiques composites pour la recherche d'informations. *In Actes du Colloque CNRS "Société de l'information"*, Lyon, France.

Enjalbert, P., éditeur (2005c). *Sémantique et traitement automatique du langage naturel*. Hermes Sciences, Lavoisier.

Enjalbert, P. et Gaio, M., éditeurs (2004). *Approches sémantiques du document numérique, Actes du 7e Colloque International sur le Document Électronique (CIDE.7)*, La Rochelle, France.

Etcheverry, P., Marquesuzaa, C. et Lesbegueries, J. (2005). Revitalisation de documents territorialisés : Principes, outils et premiers résultats. *In Workshop Met-SI INFORSID'05*.

Fairthorne, R. A. (1969). Content Analysis, Specification and Control. *Annual Review of Information Science and Technology*, 4:73–109.

Felber, H. (1987). *Manuel de terminologie*. Paris, UNESCO.

Ferrari, S. et Beust, P. (2003). Les besoins d'interactions en traitement automatique des langues et en linguistique de corpus : étude de cas. *In Actes des 3èmes Journées de Linguistique de Corpus*, Lorient, France.

Ferrari, S., Frédérik, B., Antoine, W. et Marion, L. (2005). Une plate-forme logicielle et une démarche pour la validation de ressources linguistiques sur corpus : application à l'évaluation de la détection automatique de cadres temporels. *In Actes des 4èmes Journées de Linguistique de Corpus*, Lorient, France. (à paraître).

Ferret, O. et Grau, B. (2001). Utiliser des corpus pour amorcer une analyse thématique. *Traitement Automatique des Langues*, 42(2):517–545.

Ferret, O., Grau, B., Hurault-Plantet, M., Illouz, G., Monceaux, L., Robba, I. et Vilnat, A. (2001a). Finding an Answer Based on the Recognition of the Question Focus. *In Proceedings of the 10th Text Retrieval Conference (TREC 10)*, Gaithersburg, États-Unis.

Ferret, O., Grau, B. et Masson, N. (1997). Utilisation d'un réseau de cooccurrences lexicales pour améliorer une analyse thématique fondée sur la distribution des mots. *In Actes des 1ères Journées du Chapitre Français de l'ISKO*, Lille, France.

FERRET, O., GRAU, B., MINEL, J.-L. et PORHIEL, S. (2001b). Repérage de structures thématiques dans des textes. *In Actes de TALN'01*, pages 163–172, Tours, France.

FIRBAS, J. (1964). On Defining The Theme in Functional Sentence Analysis. *Travaux Linguistiques de Prague*, 2:239–256.

FLEURY, J. (2006). Intégration de l'analyseur syntaxique Syntex dans la plateforme LinguaStream. Mémoire de D.E.S.S., Université de Caen, France.

GOPINATH, M. A. (1976). Colon Specification. *In* MALTBY A., éditeur : *Classification in the 1970s : a second look*, pages 51–80. London : Clive Bingley.

GOUTSOS, D. (1996). A Model of Sequential Relations in Expository Texts. *Text*, 16(4):501–533.

GREFENSTETTE, G. (1994). Corpus-derived first, second and third-order word affinities. *In Proceedings of EURALEX '94*, Amsterdam, Pays-Bas.

GRISHMAN, R. (2005). NLP : An Information Extraction Perspective. *In Proceedings of the 5th International Conference on Recent Advances in Natural Language Processing*.

GROSSO, P., MALER, E., MARSH, J. et WALSH, N. (2003). *XPointer Framework*. Recommandation W3C.

GROSZ, B. J., JOSHI, A. K. et WEISTEIN, S. (1995). Centering : A framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

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

HABERT, B. (2005). Portrait de linguiste(s) à l'instrument. *Revue Texto*.

HAIMAN, J. (1978). Conditionals Are Topics. *Language*, 54:564–589.

HALLIDAY, M. A. K. (1994). *An Introduction to Functional Grammar*. London : Edward Arnold. 2nd edition.

HALLIDAY, M. A. K. et HASAN, R. (1976). *Cohesion in English*. London : Longman.

HEARST, M. (1994). Multi-paragraph segmentation of expository texts. *In Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics*.

HERNANDEZ, N. (2004). *Description et détection automatique de structures de texte*. Thèse de doctorat, Université Paris Sud - Orsay, France.

HJØRLAND, B. (1992). The Concept of 'Subject' in Information Science. *Journal of Documentation*, 48(2):172–200.

HJØRLAND, B. (2001). Towards a Theory of Aboutness, Subject, Topicality, Theme, Domain, Field, Content ... and Relevance. *Journal of The American Society for Information Science and Technology*, 52(9):774–778.

HJØRLAND, B. (2002). Domain analysis in information science. Eleven approaches - traditional as well as innovative. *Journal of Documentation*, 58(4):422–462.

HO DAC, L.-M. (2000). Méthode d'analyse et de représentation de l'encadrement du discours . Mémoire de D.E.A., Université de Toulouse Le Mirail.

HO DAC, L.-M. (2006). Deux modes de segmentation textuelle : univers de discours et chaînes de référence. *Verbum*. À paraître.

HO DAC, L.-M., LE DRAOULEC, A. et PÉRY-WOODLEY, M.-P. (2003). Cohabitation des dimensions temps, espace et "phénomènes" dans un texte géographique. *Cahiers de Grammaire*, 25:125–142.

HUIBERS, T. W. C. (1996). *An Axiomatic Theory for Information Retrieval*. Thèse de doctorat, Université de Utrecht, Pays-Bas.

HUTCHINS, W. J. (1977). On the Problem of 'Aboutness' in Document Analysis. *Journal of Informatics*, 1(1):17–35.

Jackiewicz, A. et Minel, J.-L. (2003). L'identification des structures discursives engendrées par les cadres organisationnels. *In Actes de TALN'03*, Batz-sur-Mer, France.

Jacobs, J. (2001). The dimension of topic-comment. *Linguistics*, 39(4):641–681.

Jacques, M.-P. (2001). La réduction du syntagme terminologique au fil du discours. *Cahiers de grammaire*, 25:93–114.

Jaeger, F. et Oshima, D. (2002). Towards a Dynamic Model of Topic-Marking. Workshop "Information Structure in Context", Stuttgart, Allemagne.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.

Jones, L. (1977). *Theme in English Expository Discourse*. Lake Bluff, Illinois : J. Jupiter Press.

Kan, M.-Y., Klavans, J.-L. et McKeown, K. (1998). Linear Segmentation and Segment Significance. *In Proceedings of the 6th International Workshop on Very Large Corpora (WVLC'98)*, pages 197–205, Montreal, Canada.

Kant, I. (1781). *Critique de la raison pure*.

Keenan, E. O. et Schieffelin, B. E. (1976). Topic as a discourse notion. A study of topic in the conversation of adults and children. *In* Li C. N., éditeur : *Subject and Topic*, pages 335–384. New York : Academic Press.

Kleene, S. C. (1956). Representation events in nerve nets and finite automata. *Automata Studies (Annals of Mathematics Studies)*, 34:3–41.

Kozima, H. (1993). Text Segmentation Based on Similarity Between Words. *In Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics (Student Session)*, Colombus, Etats-Unis.

Laignelet, M. (2003). Les cadres de discours spatiaux et temporels dans les documents géographiques : interactions et croisements. Mémoire de Maîtrise, Université de Toulouse-Le Mirail, France.

Laignelet, M. (2006). Les titres et les introducteurs de cadres comme indices pour le repérage de segments d'information évolutive. *In Actes du Colloque International Discours et Document (ISDD'06)*, Caen, France. Soumis pour publication.

Lalmas, M. (1998). Logical Models in Information Retrieval : Introduction and Overview. *Information Processing and Management*, 1(28):19–33.

Lambrecht, K. (1994). *Information Structure and Sentence Form : Topic, Focus and the Mental Representation of Discourse Referents*. Cambridge Studies in Linguistics. Cambridge University Press.

Landauer, T. K., Foltz, P. W. et Laham, D. (1998). An introduction to Latent Semantic Analysis. *In Discourse Processes*, volume 25, pages 259–284.

Laszlo, M., Kosseim, L. et Lapalme, G. (2000). Goal-driven Answer Extraction. *In Proceedings of the 9th Text REtrieval Conference (TREC 9)*, pages 452–464, Gaithersburg, États-Unis.

Laurini, R. et Milleret-Raffort, F. (1993). *Les bases de données en géomatique*. Hermès, Paris.

Le Draoulec, A. et Péry-woodley, M.-P. (2001). Corpus-based identification of temporal organisation in discourse. *In Actes de Corpus Linguistics (CL'2001)*, pages 159–166, Lancaster, UK.

Le Goffic, P. (1994). *Grammaire de la phrase française*. Paris : Hachette.

Lee, R. et Chang, C. (1973). *Symbolic Logic and Mechanical Theorem Proving*. Academic Press.

Lesbegueries, J. (2005). Des services web destinés à l'indexation et la recherche spatio-temporelle dans un corpus territorialisé. Rapport technique, France.

Li, C. N. et Thompson, S. A. (1976). Subject and Topic : A New Typology of Language. *In* Li C. N., éditeur : *Subject and Topic*, pages 483–485. New York : Academic Press.

Longacre, R. E. (1970). Sentences Structure as Statement Calculus. *Language*, 46:783–815.

Longacre, R. E. (1974). Narrative Versus Other Discourse Genre. *Advances in Tagmemics*, pages 357–376.

Loustau, P. (2005). Traitements sémantiques de documents dans leur composante spatiale. Mémoire de Master, Université de Pau et des Pays de l'Adour, France.

Lucas, N. (2005). Etude linguistique des procédés d'exposition dans un forum de discussion. *In Actes du Symposium Symfonic*, Amiens, France.

Lucas, N. et Giguet, E. (2005). UniTHEM, un exemple de traitement linguistique à couverture multilingue. *In Actes de la 8e Conférence Internationale sur le Document Electronique (CIDE.8)*, Beyrouth, Liban.

Luhn, H. (1957). A statistical approch to mechanised encoding and searching of library information. *IBM Journal of research and development*, 1:309–331.

Maniez, J. (1999). Des classifications aux thésaurus : du bon usage des facettes. *Documentaliste - Sciences de l'information*, 35(4-5):249–262.

Mann, W. C. et Thompson, S. A. (1987). *Rhetorical Structure Theory : A Theory of Text Organization*. Report number ISI/RS-87-190, University of Southern California, Information Sciences Institute.

Marcu, D. (1996). Building up rhetorical structure trees. *In Proceedings of the AAAI annual meeting*.

Marcu, D. (1997). The rhetorical parsing of natural language texts. *In Proceedings of the Meeting of the Association for Computational Linguistics*, pages 96–103.

Marcu, D. (2000). Extending a formal and computational model of rhetorical structure theory with intentional structures à la Grosz and Sidner. *In Proceedings of COLING 2000*, pages 523–529.

Maron, M. E. (1977). On indexing, retrieval and the meaning of about. *Journal of the American Society for Information Science*, 28(1):38–43.

Maron, M. E. et Kuhns, J. (1960). On relevance, Probabilistic Indexing and Information Retrieval. *Journal of the ACM*, 7:216–244.

Marsh, J., Veillard, D. et Walsh, N. (2005). *XML :ID Version 1.0*. Recommandation W3C.

Martin, R. (1983). *La logique du sens*. Presses Universitaires de France.

Masson, N. (1995). An Automatic Method for Document Structuring. *In Proceedings of the 18th Annual International Conference on Research and Development in Information Retrieval*, pages 372–373.

Mathet, Y. (2000). *Étude de l'expression en langue de l'espace et du déplacement : analyse linguistique, modélisation cognitive, et leur expérimentation informatique*. Thèse de doctorat, Université de Caen, France.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. et Miller, K. J. (1990). Introduction to WordNet : An On-Line Lexical Database. *International Journal of Lexicography*, 3(4):235–312.

Molla, D., Schwitter, R., Hess, M. et Fournier, R. (2000). Extrans, an Answer Extraction System. *Traitement Automatique des Langues*, 41(2):495–522.

Moore, J. D. et Paris, C. (1993). Planning text for advisory dialogues : Capturing intentional and rhetorical information. *Computational Linguistics*, 19:652–694.

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

Moser, M. et Moore, J. D. (1996). Toward a synthesis of two accounts of discourse structure. *Computational Linguistics*, 22(3):409–419.

Mourad, G. et Schrepfer-André, G. (2002). Le repérage automatique des segments textuels de discours rapporté au moyen des "selon X" énonciatifs. *In Actes du 5e Colloque International sur le Document électronique (CIDE'02)*.

Muller, C., Royaute, J. et Silberztein, M. (2004). *INTEX pour la Linguistique et le Traitement Automatique des Langues*. Presses Universitaires de Franche-Comté.

Nie, J. Y. (2001). A General Logical Approach to Inferential Information Retrieval. *Encyclopedia of Computer Science and Technology*, 44:203–226.

Paice, C. D. (1981). The Automatic Generation of Literature Abstracts : an Approach Based on the Identification of Self-Indicating Phrases. *In* Oddy, R. N., Robertson, S. E., Van Rijsbergen, C. J. et Williams, P. W., éditeurs : *Information retrieval research*, pages 172–191. Butterworths, Londres.

Perlerin, V. (2004). *Sémantique légère pour le document : Assistance personnalisée pour l'accès au document et l'exploration de son contenu*. Thèse de doctorat, Université de Caen.

Person, C. (2004). *Traitement automatique de la temporalité du récit : implémentation du modèle linguistique SdT*. Thèse de doctorat, Université de Caen, France.

Pinatel, P. (2003). Coloriage thématique à l'intérieur d'un document : approche contextuelle. Mémoire de D.E.A., Université de Caen, France.

Porhiel, S. (2001). Linguistic Expressions as a Tool to Extract Thematic Information. *Corpus Linguistic*, pages 477–482.

Pottier, B. (1964). Vers une sémantique moderne. *In Travaux de sémantique et de littérature*, volume 2, pages 107–137.

Prince, E. (1981). Toward a taxonomy of given-new information. *In* Cole P., éditeur : *Radical Pragmatics*, pages 223–255. New York : Academic Press.

Péry-Woodley, M.-P. (2000a). Cadrer ou centrer son discours ? Introducteurs de cadre et centrage. *Verbum*, 22(1):59–78.

Péry-Woodley, M.-P. (2000b). Une pragmatique à fleur de texte : approche en corpus de l'organisation textuelle. *Carnets de grammaire*, 8.

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

Rambow, O. (1994). Pragmatic Aspects of Scrambling and Topicalization in German. *In IRCS Workshop on Centering in Discourse*.

Ranganathan, S. R. (1963). *Documentation and its facets*. London : Asia Publishing House.

Rastier, F. (1987). *Sémantique interprétative*. Paris : Presses Universitaires de France.

Rastier, F. (1995a). La sémantique des thèmes (ou le voyage sentimental). *In L'analyse des données textuelles*, pages 223–249. Paris : Didier.

Rastier, F. (1995b). Le terme : entre ontologie et linguistique. *La banque des mots*, 7:35–65.

Reynar, J. C. (1994). An Automatic Method of Finding Topic Boundaries. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 331–333.

Reynar, J. C. (2000). *Topic Segmentation : Algorithms and Applications*. Thèse de doctorat, Université de Pennsylvanie, États Unis.

Saggion, H. et Lapalme, G. (2002). Generating Indicative-Informative Summaries with SumUM. *Computational Linguistics*, 28(4):497–526.

Salton, G. (1975). *Dynamic Information and Library Processing*. Prentice-Hall, Englewood Cliffs.

Salton, G., Buckley, C. et Mitral, M. (1996). Automatic Text Decomposition Using Text Segments and Text Themes. *In Proceedings of the UK Conference on Hypertext*, pages 52–65.

SAUSSURE, F. d. (1916). *Cours de linguistique générale*.

SCHMIDT, H. (1994). Probabilistic part-of-speech tagging using decision trees. *In Proceedings of the International Conference on New Methods in Language Processing*.

SMOLCZEWSKA, A. et LALLICH-BOIDIN, G. (2004). Validation par prototypage d'un modèle de segmentation des documents techniques composites. *In* P. ENJALBERT ET M. GAIO, éditeur : *Approches sémantiques du document numérique, Actes du septième Colloque International sur le Document Electronique (CIDE.7)*, pages 75–92, La Rochelle, France.

SOERGEL, D. (1985). *Organizing Information : principles of database and retrieval systems*. London : Academic Press.

SOULET, A. (2002). Système de segmentation du discours par automates. Mémoire de D.E.A., Université de Caen, France.

SRIHARI, R. et LI, W. (2005). A Question Answering System Supported by Information Extraction. *In* STRZALKOWSKI, T. et HARABAGIU, S., éditeurs : *Advances in Open-Domain Question Answering*. Kluwer Academic Publishers, Dordrecht.

STA, J.-D. et YILDIZ, A. (1997). Acquisition automatisée de relations terminologiques à partir de corpus. *In Actes des 4èmes Journées Internationales de Terminologie*, Barcelone, Espagne.

STRAWSON, P. (1964). Indentifying Reference and Truth Values. *Theoria*, 30:93–118.

STRUBE, M. et HAHN, U. (1996). Functional Centering. *In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL'96)*, pages 270–277.

STRUBE, M. et HAHN, U. (1999). Functional Centering : Grounding Referential Coherence in Information Structure. *Computational Linguistics*, 25(3):309–344.

TABOADA, M. T. (1995). Theme Markedness in English and Spanish : A Systemic-Functional Approach. Universidad Complutense de Madrid, Publication Web, www.sfu.ca/~mtaboada.

TANTZEN, S. (2004). Ein Prologparser für temporale und lokale Ausdrücke für das Deutsche. Mémoire de Master, Der Philosophischen Fakultät II, der Friedrich-Alexander-Universität Erlangen-Nürnberg, Deutchland.

ULLMANN, S. (1952). *Précis de sémantique française*. Berne, Francke.

VAN DIJK, T. A. (1972). *Somes Aspects of Text Grammars*. Mouton.

VAN DIJK, T. A. (1977a). Sentence Topic and Discourse Topic. *Papers in Slavic Philology*, 1:49–61.

VAN DIJK, T. A. (1977b). *Text and Context*. London : Longman.

VAN DIJK, T. A. (1985). Semantic Discourse Analysis. *In Handbook of Discourse Analysis*, volume 2, pages 103–136. London : Academic Press.

VERGNE, J. et GIGUET, E. (1998). Regards théoriques sur le tagging. *In Actes de la cinquième conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN'98)*, Paris, France.

VICKERY, B. C. (1963). *La classification à facettes*. Paris : Gauthier-Villars.

VIRTANEN, T. (1990). Temporal Adverbials in Text Structuring : On Temporal Text Strategy. *In* N. ENKVIST ET K. WIKBERG, éditeur : *Nordic Research on Text and Discourse : Nordtext Symposium '90*, pages 185–197. Abo Academic Press.

VOORHEES, E. M. et HARMANN, D. (2001). Overview of the TREC 2001 Question Answering Track. Rapport technique.

WALKER, M. A., JOSHI, A. K. et PRINCE, E. (1998). *Centering theory in discourse*. Oxford University Press.

WIDLÖCHER, A. (2004). Analyse macro-sémantique : vers une analyse rhétorique du discours. *In Actes de RECITAL'04*, pages 183–188.

WIDLÖCHER, A. (2006). Analyse par contraintes de l'organisation du discours. *In Actes de la 13e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2006)*.

WIDLÖCHER, A., FAUROT, É. et BILHAUT, F. (2004). Multimodal Indexation of Contrastive Structures in Geographical Documents. *In Proceedings of the 7th RIAO Conference (Recherche d'Information Assistée par Ordinateur)*, pages 555–570, Avignon, France.

WILSON, P. (1968). *Two kinds of power. An essay on bibliographical control*. Berkeley, CA : University of California Press.

WÜSTER, E. (1974). L'inversion d'un rapport notionnel et les symboles correspondants utilisés en lexicographie. *Nachrichten für Dokumentation*, 26(6):256–263. traduit par Infoterm.

YOUMANS, G. (1991). A New Tool for Discourse Analysis : The Vocabulary-Management Profile. *Language*, 67(4):763–789.

Document préparé avec XEW et mis en page par LaTeX.
Compilé le 16 octobre 2006.

Résumé

Cette thèse s'inscrit dans le domaine du traitement automatique des langues, et concerne l'analyse sémantique de la structure du discours. Nous nous attachons plus particulièrement au problème de l'analyse thématique, qui vise l'étude de la structure des textes selon des critères relatifs à la répartition de leur contenu informationnel. Cette tâche revêt une importance capitale dans la perspective de l'accès assisté à l'information, qui constitue notre principale visée applicative. Le concept même de « thème » étant à la fois complexe et assez rarement considéré en tant qu'objet d'étude dans le domaine de la recherche d'information, la première partie du mémoire est consacrée à une vaste étude bibliographique autour des notions de thème, de topique, de sujet ou encore d'à propos, tant en linguistique qu'en sciences de l'information ou en traitement des langues. Nous en dégageons les lignes de force qui fondent notre approche du thème comme objet discursif, sémantique et structuré. Nous proposons sur cette base différents modèles et procédés s'attachant d'abord au traitement sémantique des documents géographiques, puis à l'analyse automatique des cadres de discours spatio-temporels au sens de Michel Charolles. Nous généralisons ces travaux en introduisant les notions de thème discursif composite et d'axe sémantique. Nous terminons en présentant LinguaStream, environnement d'expérimentation intégré que nous avons conçu pour faciliter l'élaboration de modèles linguistiques opérationnels, et qui nous conduit à proposer des principes méthodologiques originaux.

Title

Automatic Analysis of Discursive Thematic Structures — Application to Information Retrieval

Abstract

This PhD thesis belongs to the Natural Language Processing (NLP) field, and relates to the automated, semantic analysis of discourse structure. More precisely, we address the issue of thematic analysis, which aims at studying the structure of texts with respect to the organisation of their informational content. This task is of particular importance for Information Retrieval, which constitutes the primary application of our work. The concept of « theme » being particularly complex but scarcely studied for itself in the information retrieval literature, the first part of our dissertation is devoted to a large bibliographical study about the notions of theme, topic, subject, and aboutness, within the linguistics, information science and NLP fields. We draw from this study a definition of the theme as a discursive, semantic and structured object. We propose several models and processes, devoted firstly to the semantic analysis of geographical documents, and secondly to the automatic analysis of temporal discourse frames in the sense of Michel Charolles. We generalise this work introducing the notions of composite topic and semantic axis. The last part is devoted to the LinguaStream platform, an integrated experimentation environment that we designed to ease the elaboration of operational linguistic models, and that lead us to propose some original methodological principles.

Discipline

Informatique linguistique

Mots-clefs

Traitement du langage naturel, Analyse du discours, Sémantique, Recherche documentaire.

GREYC – CNRS UMR 6072
Université de Caen, Campus II
BP 5186, F-14032 Caen Cedex, France