N° D'ORDRE : Année 1998

UNIVERSITE DE PARIS-SUD
U.F.R. SCIENTIFIQUE D'ORSAY

THÈSE

présentée

Pour obtenir le grade de

DOCTEUR EN SCIENCES
DE L'UNIVERSITÉ PARIS XI ORSAY

Discipline : Informatique

Par

Olivier FERRET

Titre :

ANTHAPSI : un système d'analyse thématique et d'apprentissage de connaissances pragmatiques fondé sur l'amorçage

Soutenue le 22 décembre 1998 devant la Commission d'examen

MM.	Brigitte GRAU	Examinateur
	Daniel KAYSER	Examinateur
	Yves KODRATOFF	Examinateur
	Maria Teresa PAZIENZA	Rapporteur
	Gérard SABAH	Directeur
	Pierre ZWEIGENBAUM	Rapporteur

Sommaire

Partie III ROSA

Introduction

Il est devenu évident que le nombre de ressources textuelles présentes sous forme électronique est très important et continuera à l'avenir de croître de façon exponentielle. Cette affirmation est même devenue une sorte de poncif, voire un argument publicitaire. Cette transformation n'en altère cependant pas la vérité initiale. Il est également assez visible que les outils informatiques disponibles pour faire face à cette masse sont encore rudimentaires. Que l'on soit en présence des vérificateurs orthographiques ou grammaticaux des traitements de textes, ou des moteurs de recherche pour le WEB ou le Minitel, il est difficile de trouver dans ces outils des capacités renvoyant à ce que comprendre un texte peut signifier, au sens le plus intuitif du terme. Même si toutes les tâches relevant du traitement automatique des langues ne nécessitent pas une compréhension approfondie des textes traités pour produire des résultats exploitables, celle-ci apporterait dans la plupart des cas une amélioration très significative de ces résultats, voire indispensable dès que la tâche devient un peu complexe. Les scores globalement assez faibles des différents systèmes testés dans des évaluations comme MUC, pour l'extraction d'informations, ou SUMMAC, pour le résumé automatique de textes, sont une illustration de ce besoin.

Cependant, comme l'ont montré les travaux de Schank à la fin des années 70 et au début des années 80, la compréhension automatique de textes achoppe en particulier sur l'épineuse question des connaissances nécessaires à sa mise en œuvre. Le problème est encore maîtrisable en ayant recours à un travail de codage "à la main" lorsque l'on travaille dans des micro-domaines comme le faisaient des programmes comme PAM, SAM ou BORIS. Il ne l'est plus dès lors que l'on se situe en environnement ouvert. Or, le développement d'outils comme le WEB ou l'intégration d'outils plus avancés dans les traitements de texte, orientent les besoins de plus en plus dans ce sens.

Parmi les connaissances intervenant dans le cadre de la compréhension automatique de textes, les connaissances pragmatiques, c'est-à-dire les connaissances décrivant le monde et son fonctionnement, présentent une difficulté toute particulière : elles forment en effet plus que les autres un ensemble ouvert. Les connaissances lexicales ou les connaissances sémantiques sont bien entendu très vastes et ne peuvent être cernées une fois pour toutes. Toutefois, il ne semble pas irraisonnable, au prix certes d'un travail important, d'en modéliser une part suffisante pour disposer en toutes circonstances d'un fond commun sur lequel s'appuyer. En l'occurrence, ce fond pourrait être l'équivalent du contenu d'un dictionnaire de la langue considérée.

Une telle approche n'est en revanche pas envisageable en ce qui concerne les connaissances pragmatiques. L'équivalent du dictionnaire pourrait être ici l'encyclopédie mais le parallèle est trompeur. Le contenu d'une encyclopédie renvoie en effet à un ensemble de savoirs globalement assez spécialisés alors que les connaissances pragmatiques faisant le plus défaut aux programmes de compréhension sont les connaissances usuelles formant le fond commun à tous les lecteurs faisant partie du même contexte culturel. Or, ces connaissances ont un tel caractère d'évidence pour tout à chacun vivant dans ce contexte qu'il est difficile d'en trouver la moindre formalisation, y compris dans les encyclopédies.

Le parallèle entre dictionnaire et encyclopédie, aussi discutable qu'il soit, met tout de même en avant la différence considérable de volume existant entre les connaissances présentes dans un dictionnaire et des connaissances pragmatiques couvrant un champ assez large. En effet, les connaissances lexicales et sémantiques tirent par essence un certain caractère de généralité de leur attachement fort à la langue alors que les connaissances pragmatiques sont par définition liées à la réalité qu'ils représentent. Or, l'étendue des réalités à décrire peut être considérée comme potentiellement infinie.

Face à cette impossibilité de circonscrire les connaissances pragmatiques, leur apprentissage automatique apparaît comme une nécessité impérative si l'on souhaite mettre en œuvre des systèmes de compréhension de textes opérant sur une large échelle. Bien que la solution la plus adaptée serait sans doute de plonger un "agent intelligent" dans le monde réel[1], l'état actuel des travaux sur la perception artificielle réserve cette approche à un avenir au terme encore incertain. Les textes présentent pour leur part le double avantage de conserver une trace de ces connaissances pragmatiques et d'être facilement accessibles dès à présent. C'est pourquoi nous nous sommes intéressé, dans le cadre de cette thèse, à l'apprentissage automatique de connaissances pragmatiques à partir de textes.

Ce type d'apprentissage se heurte néanmoins à un blocage en apparence insurmontable, ce qui explique peut-être pourquoi il a été assez peu abordé : pour mettre en évidence les connaissances pragmatiques évoquées par les textes, il est nécessaire de mettre en œuvre des processus de compréhension de texte; or, ces derniers ne peuvent opérer que s'ils disposent des connaissances pragmatiques relatives à la réalité évoquée, en l'occurrence celles que l'on cherche précisément à apprendre.

La seule solution pour sortir de cette interdépendance paralysante consiste selon nous à faire appel à un mécanisme d'amorçage : les capacités d'analyse des textes disponibles à

[1] Nous nous garderons soigneusement de définir la notion de monde réel et de nous aventurer sur le terrain de la controverse philosophique entre idéalisme et matérialisme.

un moment donné, pour aussi imparfaites qu'elles puissent être, sont tout de même utilisées afin de construire une représentation des textes, axée sur la dimension pragmatique. Cette analyse est couplée à un processus d'apprentissage incrémental travaillant à partir de ces représentations au fur et à mesure de leur production dans le but de produire de nouvelles connaissances pragmatiques. Dans ce mode de fonctionnement, le traitement d'un grand nombre de textes à propos d'un même sujet permet de compenser en partie les imprécisions de l'analyse de chacun des textes. Les connaissances ainsi produites sont ensuite réutilisées par le processus de compréhension afin d'améliorer ses capacités d'analyse. Le processus se poursuit ainsi de suite, conduisant à une extension et à une amélioration progressive des connaissances disponibles.

Même si ce processus d'amorçage met l'accent sur la progressivité de l'apprentissage, il ne résout pas pour autant le problème de l'état initial. En l'absence de toute connaissance, les capacités d'analyse sont réduites à zéro et le processus ne peut tout simplement pas démarrer. Deux solutions sont envisageables pour sortir de ce blocage initial. L'une d'elle consiste à modéliser manuellement un vaste ensemble de connaissances considérées comme élémentaires en espérant atteindre la masse critique permettant d'acquérir toutes les autres connaissances pragmatiques. C'est l'approche retenue par le projet Cyc.

L'autre solution se fonde sur l'application du concept d'amorçage à différents niveaux de représentation. Elle exploite le fait que les traitements portant sur les textes ne nécessitent pas tous des connaissances pragmatiques ou peuvent se contenter de connaissances acquises de façon statistique à partir de gros corpus. En dépit des limites de leurs résultats, ces traitements peuvent tout à fait être mis au service de l'acquisition de connaissances pragmatiques. Bien entendu, les connaissances obtenues par ce moyen ne sont pas caractérisées par un niveau de représentation très élevé : les mots y tiennent lieu de concepts et les structures se résument bien souvent à de simples ensembles.

Leur présence donne cependant la possibilité de développer un processus d'analyse plus élaboré, puisque pouvant s'appuyer sur cette première forme de connaissance. Un tel processus est capable de produire des représentations de texte elles-mêmes plus avancées qui, après exploitation par un processus d'apprentissage spécifique, donnent lieu à des connaissances pragmatiques plus précises et mieux structurées. Cet amorçage peut se décliner ainsi sur toute une succession de niveaux jusqu'à atteindre des formes de représentation semblables aux schémas largement utilisés par Schank.

Dans le cadre de cette thèse, nous avons choisi de développer plus spécifiquement la seconde solution et donc, de pousser l'amorçage le plus loin possible. Il n'était néanmoins pas possible, dans un temps assez limité, de traiter l'intégralité du problème.

Cela est d'autant plus vrai qu'en toute généralité, un tel amorçage devrait ne pas se limiter à la seule dimension pragmatique et inclure au moins la dimension conceptuelle. Compte tenu de notre objectif global, l'apprentissage de connaissances pragmatiques, nous nous sommes focalisé sur la dimension pragmatique.

Dans ce contexte, nous avons choisi de définir en priorité ce à quoi nous souhaitions aboutir, c'est-à-dire le point d'arrivée de l'amorçage, ainsi que ce dont nous partions. Le premier est incarné par le système MLK; le second par le système ROSA. Les deux forment le système ANTHAPSI (ANalyse THématique et APprentissage de SItuations). En vertu de l'amorçage développé ici, chacun de ces deux niveaux fonctionne suivant les mêmes principes. En raison du degré élevé de structuration et de précision des connaisances qu'il manipule, MLK permet de définir avec toute la finesse nécessaire ces principes de fonctionnement. C'est pour cette raison que nous avons choisi de commencer la présentation d'ANTHAPSI par celle de MLK.

ROSA remplit pour sa part deux fonctions dans ce travail : il représente d'abord le premier étage du processus d'amorçage, destiné en final à permettre le fonctionnement de MLK. Il est ensuite un moyen de valider les principes sous-tendant l'activité de chacun des niveaux d'ANTHAPSI. Les principes de ROSA étant semblables à ceux de MLK, les expérimentations permises par ROSA constituent en effet une forme de validation des principes généraux appliqués dans MLK.

Plan de l'exposé

Le manuscrit se compose de trois grandes parties. La première, qui regroupe les trois premiers chapitres, est chargée d'introduire le système ANTHAPSI dans son ensemble, aussi bien au niveau de ses principes que de son architecture. La deuxième rassemble les chapitres 5 à 8 et décrit en détail le système MLK. Les deux derniers chapitres forment la troisième partie du manuscrit et exposent le système ROSA ainsi que les expérimentations dont il a fait l'objet.

Chaque chapitre est introduit par un bref chapeau et suivi par un récapitulatif assez détaillé. Le lecteur pressé ou ne désirant pas aborder tel ou tel aspect en détail pourra donc se contenter de lire le chapeau et le récapitulatif du chapitre concerné.

Les différents chapitres se définissent plus précisément comme suit. Le *chapitre 1* est formé de deux parties : la première cerne précisément le problème de l'apprentissage de connaissances pragmatiques à partir de textes tandis que la seconde expose les principes généraux de l'approche retenue pour le résoudre. Les deux parties reprennent en l'approfondissant l'argumentaire présenté dans cette introduction. Le *chapitre 2* rend

compte des travaux les plus importants spécifiquement dédiés à ce type d'apprentissage et en tire un certain nombre de réflexions sur les orientations intéressantes à suivre. Le *chapitre 3* marque pour sa part le début de la présentation détaillée de la solution adoptée en donnant une vue d'ensemble du système ANTHAPSI. Cette vue présente l'architecture générale d'ANTHAPSI et définit les spécifications fonctionnelles de chacune de ses composantes. Par ailleurs, elle expose les principes de l'amorçage sous-tendant l'ensemble du système.

Le *chapitre 4* marque le début de la description de MLK. Il décrit les connaissances sur lesquelles MLK s'appuie, autrement dit les connaissances sémantiques, formalisées ici au travers des graphes conceptuels, ainsi que les connaissances qu'il cherche en final à construire, en l'occurrence des connaissances pragmatiques stables prenant la forme de schémas. Les familiers des graphes conceptuels pourront laisser de côté la première partie du chapitre jusqu'au paragraphe 1.3. Ceux qui ne s'intéressent pas à l'abstraction des schémas exposée au chapitre 7 pourront ne pas lire la seconde partie. Le *chapitre 5* est dédié quant à lui à la description des représentations de texte manipulées dans MLK. Leur rôle central – elles sont produites par l'analyse thématique de MLK et servent de matière première à son processus d'apprentissage – justifie le fait de les présenter avant les deux grandes composantes fonctionnelles de MLK.

Le *chapitre 6* aborde précisément l'une de ces deux grandes composantes : la mémoire épisodique, chargée de faire émerger de nouvelles connaissances pragmatiques à partir des représentations de texte construites par l'analyse thématique de MLK. Le chapitre détaille la structure de cette mémoire ainsi que les deux opérations fondamentales qui lui sont associées : le rappel des connaissances et la mémorisation de nouvelles représentations de texte. Le *chapitre 7* est consacré pour sa part à l'abstraction de schémas à partir du contenu de la mémoire épisodique, c'est-à-dire le stade ultime de l'apprentissage des connaissances pragmatiques dans MLK. Le *chapitre 8*, enfin, clôt la présentation de MLK en exposant sa seconde grande composante fonctionnelle : l'analyse thématique, chargée de construire les représentations de texte décrites au chapitre 5 à l'aide des connaissances contenues dans la mémoire épisodique détaillée au chapitre 6.

L'exposé de ROSA commence avec le *chapitre 9,* dédié à la partie de ROSA ayant un statut d'amorce initiale en fournissant un mécanisme de segmentation thématique des textes axé sur la robustesse. Le *chapitre 10* , qui marque la fin du manuscrit, décrit quant à lui le cœur de ROSA, constitué à l'image de MLK de deux composantes en étroite interaction, l'une mettant en œuvre un apprentissage de connaissances pragmatiques et l'autre, une segmentation thématique des textes. Le chapitre s'achève sur la présentation

des processus d’amorçage intervenant à la fois au sein de ROSA, entre son amorce initiale et son niveau principal, ainsi qu’entre ROSA et MLK.

Partie I

Préambule

Chapitre 1

Exposé du problème et principes de la solution retenue

Dans ce chapitre, nous commençons par présenter le problème auquel nous nous intéressons, l'apprentissage automatique de connaissances pragmatiques à partir de textes. Nous exposons ensuite en quoi ce problème nous oblige à aborder la compréhension et l'apprentissage selon une approche spécifique mettant en avant la notion d'expérience. Cette notion est ensuite plus amplement définie et son impact tant au niveau de l'apprentissage que de la compréhension est étudiée plus finement.

1. Problématique

1.1. Définition et intérêt du problème

1.1.1. Quelles connaissances sur le monde?

Le problème que nous abordons ici est celui de l'apprentissage automatique de connaissances pragmatiques à partir de textes. Sans sombrer dans l'exégèse, il est nécessaire d'apporter quelques précisions sur ce que cette formule recouvre exactement. La notion de connaissance pragmatique est, au premier chef, une source potentielle d'ambiguïtés. La définition qui en est donnée dans [Sabah 1988] est une façon globale de la cerner : il s'agit de "connaissances décrivant le fonctionnement du monde de référence". On parle aussi de connaissances sur le monde.

Cette définition nous permet déjà de lever l'ambiguïté pouvant résulter de l'emploi du mot pragmatique. Il n'est en effet pas question ici de connaissances contribuant à la description d'une situation particulière d'énonciation où l'on s'attacherait à rendre compte de la position respective des locuteurs, de leur statut vis-à-vis d'une tâche dans laquelle ils seraient impliqués ou de leurs intentions.

Il s'agit au contraire de représenter des situations prototypiques du monde de référence. Dans un souci de simplification, nous considérerons que ce monde de référence s'identifie, dans la suite de notre exposé, au monde "réel" dans lequel nous sommes plongés et nous laisserons donc de côté les univers imaginaires construits par la Science Fiction, la littérature fantastique ou encore les mythes et légendes.

Dans ce cadre, la notion de situation prototypique s'incarne dans des événements aussi divers qu'une *manifestation de protestation* , un *assassinat politique* ou bien *aller au cinéma, prendre le train* ou *faire des courses*. Elle peut être cernée qualitativement en reprenant les principes du théâtre classique pour la voir comme un bloc d'actions plus ou moins ordonnées faisant intervenir un ensemble de protagonistes (au sens large) selon une unité de temps, de lieu et d'action. C'est plus formellement ce que Schank a tenté de représenter au travers de la notion de scénario ("script") dans [Schank & Abelson 1977] et qui a évolué vers celle de MOP ("Memory Organization Packet") dans [Schank 1982].

Cette définition nous permet d'opérer deux distinctions supplémentaires visant à restreindre le champ de l'idée intuitive de connaissance sur le monde. Distinction d'abord par rapport aux connaissances dites conceptuelles. En affirmant qu'une voiture est un objet manufacturé auto-propulsé comportant une caisse, une carrosserie, un moteur et quatre roues, on exprime une forme de connaissance sur le monde. Mais on cherche là à spécifier les propriétés générales d'objets ou d'actions du monde que l'on considère sans pour autant préciser la façon dont ils interagissent habituellement dans ce même monde. L'exemple suivant permet de percevoir plus nettement cette différence.

> On considère d'une part, les concepts *aller* et *cinéma* et d'autre part, la situation *aller_au_cinéma* (les caractérisations données se veulent informelles et n'ont qu'un caractère illustratif)
>
> *aller* : c'est une action telle qu'une entité animée change de lieu en passant d'un lieu source à un lieu cible
>
> *cinéma* : c'est un lieu public dans lequel on projette des films
>
> *aller_au_cinéma* : c'est une situation se décomposant en événements plus élémentaires tels que choisir_un_film, choisir_un_cinéma, faire_la_queue, acheter_un_billet, s'installer_à_une_place, acheter_des_friandises, voir_les_publicités, voir_le_film (la liste ne se veut pas exhaustive).
>
> Tous ces événements ne surviennent pas systématiquement et leur ordre d'apparition est plus ou moins fixé mais ils sont caractéristiques de ce à quoi on peut s'attendre lorsqu'on va au cinéma.

La seconde distinction relève de connaissances sur le monde plus abstraites et plus générales cherchant à rendre compte des mécanismes causaux profonds régissant l'apparition des événements. Dans le cas des êtres animés d'intentions, les connaissances de ce type mettent en évidence les buts des personnages d'une situation et les relient aux plans que ceux-ci sont susceptibles d'appliquer pour atteindre ces buts compte tenu d'un

état du monde. Les actions observées sont alors interprétables en tant que résultat de la mise en œuvre de certains de ces plans. Au contraire des connaissances que l'on considère ici, ces buts et ces plans ne sont pas liés à une situation particulière et représentent plutôt une connaissance générale sur les motivations, le comportement de certaines catégories de personnages ainsi que sur la façon dont ces personnages interagissent. On cherchera ainsi à décrire comment réaliser des buts simples tels que *se nourrir, se rendre en un lieu, s'informer* ou bien expliciter des configurations plus complexes telles que *la concurrence entre plusieurs personnages pour accéder à une ressource limitée* ou *la préférence pour un plan plus facile à réaliser même s'il ne garantit pas le succès*. Les TOPs ("Thematic Organization Packets") décrits dans [Schank 1982] constituent une proposition de formalisation de ce type de connaissances pragmatiques.

Le souci de représenter la causalité sous-jacente aux phénomènes s'est également appliquée en dehors des êtres pourvus d'une intentionnalité. Le même type de connaissance existe en effet pour décrire le fonctionnement des machines, les interactions entre objets et rendre compte plus généralement des phénomènes physiques. On parle alors de modèles, et plus précisément de modèles causaux, d'où l'appellation de "Model-Based Reasoning" pour faire référence aux approches exploitant ces représentations [Forbus 1988].

1.1.2. L'intérêt des connaissances pragmatiques

La nécessité de faire appel à des connaissances de diverses sortes pour résoudre les problèmes relevant de l'Intelligence Artificielle n'est plus guère à démontrer. La pratique en l'espèce a également révélé qu'il est heureusement possible de circonscrire dans beaucoup de cas le champ des connaissances qu'il est nécessaire de modéliser. Dans le domaine du traitement automatique des langues, qui est celui qui nous intéresse ici, on a montré que cette possibilité de circonscription est difficile dès lors que l'on n'impose pas l'usage d'une sous-langue très contrainte. Même si le sujet du discours est thématiquement bien cerné, le locuteur ou le rédacteur fait naturellement, notamment au travers de métaphores (qui sont des phénomènes très courants), des références à des connaissances que tout à chacun possède sur le monde qui nous entoure et qui se trouvent en dehors du champ thématique principal du discours. On voit donc la nécessité, pour qui veut essayer de construire un système automatique de compréhension d'une langue, de disposer d'une base de connaissances sur le monde.

Certes, il est possible d'élaborer des programmes efficaces de traitement automatique d'une langue même s'ils ne font aucun usage de connaissances pragmatiques. C'est le cas, par exemple, de programmes de filtrage d'informations, d'indexation automatique de documents ou même de découpage thématique de textes. Ces programmes se fondent

généralement sur la présence d'indices de surface – des marques linguistiques spécifiques telles que certains connecteurs, la récurrence de mots non grammaticaux – qu'ils exploitent pour accomplir une tâche spécialisée.

Dans une optique moins radicale, certains travaux reposent sur une utilisation implicite de connaissances pragmatiques. C'est le cas en particulier de ceux qui utilisent comme source de connaissances un réseau de co-occurrences lexicales calculées à partir d'un vaste corpus de textes. [Kozima 1993] montre ainsi comment un tel réseau peut servir à dégager la structure thématique de textes narratifs. Il met en évidence à cette occasion le fait qu'un tel réseau rend compte à la fois des relations de nature sémantique (relations d'hyperonymie et d'hyponymie, de méronymie, de synonymie et d'antonymie) et des relations de nature pragmatique (proximité des mots dans les textes résultant de l'appartenance des entités qu'ils désignent à une même situation). On peut même ajouter que les réseaux de ce genre contiennent également des relations de nature syntaxique.

De ce fait, ils permettent d'utiliser facilement des types de connaissances différents car ceux-ci se retrouvent amalgamés de façon homogène. Ce mélange, qui se révèle efficace pour certaines tâches, est également le principal reproche que l'on peut adresser à cette approche. Lorsqu'on s'appuie sur la co-occurrence de deux mots, on ne sait en effet jamais quelle est la raison sous-jacente à cette co-occurrence. Il est donc difficile, dans ces conditions, de faire usage de ces co-occurrences pour résoudre des problèmes précis. Ceux-ci étant identifiés, on cherche à leur appliquer des procédures de résolution spécifiques qui reposent sur des connaissances elles-mêmes bien cernées. Cette modularité apparaît comme une nécessité dès lors que l'on s'attache à traiter des problèmes complexes et de fait, s'accommode mal de l'amalgame que constitue une source de connaissances telle qu'un réseau de co-occurrences lexicales.

La compréhension de textes est un de ces problèmes complexes et cela apporte une première justification à son besoin de connaissances sur le monde, présentes sous une forme explicite. Nous ne chercherons pas ici à mener une réflexion approfondie sur ce que recouvre la notion de compréhension de textes. Nous nous contenterons d'une définition opérationnelle, sans doute réductrice, mais qui présente l'avantage d'une possible objectivation. Un système de compréhension de textes se caractérise ainsi par sa capacité à se prêter à des tests de type DQR, où D représente les phrases déclaratives d'un texte, Q, des questions portant sur ce texte et R, les réponses apportées aux questions Q par le système [Sabatier 1997]. Une telle approche a ainsi été préconisée dans le cadre du projet FraCas [FraCaS 1996]. On trouve d'autre part une forme dégradée de cette conception au travers des systèmes d'extraction d'informations évalués dans le cadre des conférences MUC [ARPA & Agency 1996]. La compréhension se restreint alors au remplissage d'un schéma prédéfini spécifiant les attentes spécifiques du système. Même

s'ils n'interprètent pas de question et qu'ils n'engendrent pas de réponse, ces systèmes réalisent néanmoins de façon plus ou complète certaines des tâches que l'on peut attendre d'un système de compréhension de textes telles que la résolution des co-références ou la réalisation d'inférences visant à mettre en évidence des éléments implicites.

Les travaux de Schank [Schank & Abelson 1977] notamment ont bien mis en évidence que le type de tâche qui nous intéresse ici ne peut être accompli en l'absence de connaissances sur le domaine concerné mais également en l'absence de connaissances plus générales sur la façon dont le monde de référence fonctionne. Cette nécessité intervient aussi bien pour expliciter pleinement ce qui est dit dans un texte que pour mettre au jour ce à quoi il fait implicitement référence.

Une illustration typique de la première dimension est le problème de la résolution des co-références. Beaucoup d'entre elles peuvent être résolues en se fondant sur des contraintes syntaxiques. L'utilisation de connaissances sémantiques élargit encore significativement le spectre des cas traités mais ne suffit pas à le couvrir complètement. Dans l'exemple suivant, l'ambiguïté existant à propos de l'antécédent du "l'" ne peut être ainsi levée si l'on ne sait pas que les voleurs s'intéressent généralement davantage aux coffres-forts qu'aux caissiers. Or, c'est là typiquement une connaissance pragmatique.

> Le voleur a assommé le caissier mais n'a pu ouvrir le coffre-fort. Il l'a donc emporté avec lui.

Le rôle des connaissances sur le monde est encore plus essentiel lorsqu'il s'agit de faire le lien entre les différentes parties de ce qui a été dit. Dans le paragraphe qui suit, il n'est pas possible de déterminer de façon précise pourquoi le personnage retrouve la liberté au bout de dix ans si l'on ignore quelles peuvent être les conséquences d'une attaque à main armée, en l'occurrence être arrêté par la police, jugé et jeté en prison pour une durée assez longue. De même, le lien avec le désir de vengeance de ce même personnage ne peut être fait que si l'on connaît le scénario un peu stéréotypé du malfrat trahi par ses acolytes.

> À la suite du hold-up de la rue S^{t} Georges, il ne retrouva la liberté qu'au bout de dix ans. Seul son désir de vengeance à l'égard de ses anciens comparses l'avait maintenu en vie durant toutes ces années.

L'importance des connaissances sur le monde est donc bien établi pour la compréhension de textes. Elle est particulièrement évidente en ce qui concerne les connaissances décrivant les mécanismes causaux régissant la réalisation des événements. Elles permettent en effet de répondre à la question qui vient sans doute le plus naturellement à l'esprit quand on parle de compréhension : pourquoi?

Les connaissances sur les situations, celles auxquelles nous nous intéresserons par la suite, n'en sont pas moins nécessaires à la compréhension. Mais leur rôle se situe sur un plan un peu différent. Étant chargées de caractériser les situations prototypiques du monde considéré, elles ont vocation à servir de point de référence. Elles sont en quelque sorte le dépositaire du fonctionnement normal et habituel de ce monde. À ce titre, elles remplissent une double mission dans le cadre de la compréhension.

Elles permettent d'abord d'expliciter un certain implicite. Le rédacteur d'un texte, ayant naturellement une intuition du fond commun de connaissances sur les situations possédé par ses lecteurs potentiels, ne fait qu'évoquer les situations qu'il met en scène en n'en faisant apparaître que quelques événements marquants. Dans l'exemple ci-dessus, l'épisode judiciaire du personnage principal n'est révélé qu'au travers de son début, le hold-up, et de sa fin, la libération de ce personnage. Seules les connaissances dont on dispose sur la situation d'un homme se faisant arrêter à la suite d'un acte délictueux permettent de compléter la chaîne des événements entre le hold-up et la libération.

La seconde mission de ces connaissances, dans le cadre de la compréhension, s'articule avec celle qui est dévolue aux connaissances sur la causalité des événements. Du fait de leur statut de référentiel, les connaissances sur les situations contribuent à faire la part entre ce qui est du ressort d'un fonctionnement normal du monde de référence et ce qui est un phénomène spécifique que l'on doit s'attacher à expliquer. Elles sont donc impliquées dans le déclenchement de traitements mettant en œuvre des connaissances telles que les buts et les plans. À cet égard, elles jouent un rôle un peu similaire à celui que peuvent jouer les structures casuelles associées aux concepts prédicatifs vis-à-vis du déclenchement de processus de traitement des métonymies [Chibout 1993] ou des métaphores [Ferrari 1993] lors de la construction de la représentation sémantique d'une proposition. Dans le passage ci-dessous, posséder une représentation de la situation *prendre_l'avion* est ainsi nécessaire pour détecter que le fait de sortir d'un aéroport pour prendre un taxi juste avant d'embarquer n'est pas un acte habituel et qu'il doit donc être expliqué à la lumière de ce s'est produit avant ou de ce qui surviendra après.

> Il se rendit à l'aéroport en taxi. Il arriva comme prévu deux heures avant le départ de son vol. Il fit enregistrer ses bagages sans encombre puis s'installa à une table de la cafétéria pour boire un café et lire son journal en attendant l'embarquement. Quelques minutes plus tard, il se leva, paya sa consommation et sortit avec précipitation de l'aéroport pour plonger dans un taxi.

1.1.3. Pourquoi chercher à apprendre automatiquement les connaissances sur les situations?

Montrer que les connaissances sur le monde sont nécessaires pour la compréhension automatique de textes est une chose, en disposer sous une forme manipulable par un ordinateur en est une autre. Compte tenu de l'étendue de ces connaissances lorsqu'on s'intéresse à la compréhension en monde ouvert, répondre à cette exigence nous place en effet devant l'alternative suivante : effectuer un vaste travail d'ingénierie des connaissances de façon à coder manuellement toutes ces connaissances ou essayer de faire en sorte que la machine puisse elle-même les apprendre de façon automatique à partir de données qu'elle est capable d'assimiler.

En y regardant de plus près, le choix se présente de façon moins tranché si l'on s'attache à l'importance de l'intervention de la machine. Il existe de fait toute une gradation en la matière. À l'échelon le plus bas, on trouve le simple outil venant comme support d'une méthode d'acquisition de connaissances. À un stade plus avancé, le système est capable de faire des propositions brutes qui seront ensuite retravaillées par un intervenant humain chargé de produire la connaissance sous sa forme finale. On classera notamment dans cette catégorie les systèmes d'extraction de relations sémantiques à partir de textes qui se contentent d'extraire les phrases ou les morceaux de phrase susceptibles de contenir les informations intéressantes [Béguin et alii 1997].

En avançant dans l'automatisation, on trouve les systèmes capables de produire des connaissances qu'ils pourront directement réutiliser mais qui demandent une validation de ces connaissances auprès d'un expert humain. C'est le cas en particulier d'un système tel que APT [Nédellec 1994] qui effectue cette démarche en proposant à son utilisateur de donner son avis sur des exemples et des contre-exemples engendrés à partir des connaissances apprises. Mais c'est plus généralement le cas, même si l'intervention humaine y est souvent moins directe, de tous les systèmes fonctionnant en apprentissage supervisé, c'est-à-dire en sachant quel est le statut des données qu'ils considèrent vis-à-vis de ce qu'ils doivent apprendre. Au stade ultime de l'automatisation, l'apprentissage est non supervisé et se fait sans aucune intervention humaine.

Ce petit panorama montre en fait qu'il existe un critère plus important que le degré global d'implication d'un expert humain. Il est en effet plus intéressant de distinguer les cas où la connaissance est apportée par un homme directement sous la forme sous laquelle cette connaissance sera utilisée par un système automatique, des cas où le système construit une représentation qui lui est propre, de connaissances qui lui sont fournies de façon plus ou moins directe par un opérateur humain.

La première approche est celle que l'on applique typiquement lorsqu'on intervient dans le cadre de micro-domaines. Les connaissances impliquées sont alors suffisamment limitées pour que le problème soit appréhendable par un homme seul sur une durée raisonnablement faible. Ce n'est évidemment plus le cas dès lors qu'on souhaite s'atteler à la modélisation d'un très vaste ensemble de connaissances sur le monde.

Le projet Cyc [Lenat et alii 1990] est l'une des rares tentatives, si ce n'est la seule, allant dans ce sens. Outre que certaines incertitudes planent encore quant au succès réel d'une telle entreprise, on remarquera que ce codage manuel à vaste échelle est conçu par leurs auteurs comme l'amorçage nécessaire pour aller vers des modes d'acquisition plus automatisés et non comme une tentative de modélisation exhaustive des connaissances sur le monde. En examinant à première vue le contenu de Cyc, on notera d'ailleurs que les connaissances concernées sont essentiellement des connaissances conceptuelles, au sens où nous l'avons défini précédemment, et non des connaissances sur les situations. Ces connaissances doivent constituer le fond minimum pour qu'un système de compréhension de l'anglais, CycNL, soit capable de construire la représentation en CycL, le langage de représentation de Cyc, d'énoncés en anglais et puisse par la suite venir enrichir les connaissances de Cyc.

Même placé dans une telle perspective, le codage manuel d'un si vaste ensemble de connaissances se heurte à plusieurs difficultés, plaidant en faveur de la construction par la machine de ses propres connaissances :

- la difficulté la plus évidente est celle de l'échelle : ce que l'on est capable de faire dans un micro-domaine ne préjuge en effet en rien de ce qui peut être fait concernant l'ensemble des connaissances sur le monde. Une certaine cohérence peut être maintenue par un seul homme sur une durée limitée. On peut être certain que tel n'est pas véritablement le cas lorsque toute une équipe opère sur plusieurs années. La modélisation d'un micro-domaine n'est d'ailleurs pas sans soulever des difficultés car isoler un bloc de connaissances de l'ensemble se heurte à l'alternative suivante : soit on cherche véritablement à découper un morceau dans un ensemble, ce qui suppose la préexistence de ce dernier, soit on modélise entièrement en fonction du domaine, sans s'occuper d'un quelconque critère de généralité. La réalité est le plus souvent un compromis entre les deux. Mais en tout état de cause, ce n'est assurément pas une démarche transposable à l'échelle de la modélisation d'un très vaste ensemble de connaissances sur le monde.

- dans le projet Cyc, le fait de fournir au système des connaissances sous une forme que l'on peut qualifier de prédigérée est vue comme une phase d'amorçage ouvrant la voie vers la capacité du système à s'auto-alimenter. Néanmoins, il n'existe pas véritablement de moyen permettant de savoir à partir de quel moment il pourrait

assimiler de la nourriture structurée plutôt que de la simple bouillie. Autrement dit, il semble difficile d'estimer la taille de l'amorce à fournir, en termes de connaissances, pour qu'il puisse ensuite s'auto-entretenir.

De même, cerner les connaissances possédées par Cyc s'avère être l'une des difficultés importantes montrées par son utilisation. Ce souci est à mettre en relation à la fois avec le problème de la cohérence – peut-on réellement garantir la cohérence de connaissances dont on ne sait pas cerner l'étendue exacte? – et celui de l'évaluation de l'ampleur du travail de codage manuel – comment définir une limite en la matière si l'on ne sait pas spécifier un état du système?

- le problème précédent soulève la question de la définition de moyens d'accès aux connaissances qui ont été codées manuellement. Cependant, peut-on fournir de tels moyens d'accès indépendamment de la tâche considérée, donc de la façon dont ces connaissances vont être utilisées? Cela paraît en réalité peu vraisemblable, d'autant plus que cette difficulté est en pratique plus profonde : on peut même douter que ces connaissances se présentent de la même façon dans leurs différents contextes d'utilisation. La compréhension de textes et la génération de textes sont ainsi assez proches l'une de l'autre, notamment dans le sens où elles utilisent les mêmes types de connaissances. Elles font par exemple appel toutes deux à des connaissances grammaticales. Il est cependant reconnu que les formalismes syntaxiques retenus en analyse sont généralement différents de ceux utilisés en génération. On pourrait bien entendu envisager l'existence de méta-connaissances permettant de spécifier comment passer d'une forme de connaissance à une autre [Pitrat 1990]. C'est là néanmoins un problème qui reste suffisamment difficile à traiter pour que cette pratique ne soit pas largement répandue.

  On en vient donc à penser qu'il est peut être possible de modéliser un vaste ensemble de connaissances sur le monde pour fournir la connaissance résultante à un système mais qu'il n'est pas sûr que celui-ci puisse en tirer véritablement profit si cette modélisation est réalisée sans le souci de l'usage qui sera fait de cette connaissance. Or, vu l'effort impliqué par ce travail de modélisation, il semble déraisonnable de le répéter, même si ce n'est que partiellement, pour différentes tâches. On en vient donc à douter, sur ce point, de la validité des hypothèses qui sous-tendent un projet tel que Cyc.

- en considérant comme illusoire la possibilité de fournir à un système l'ensemble des connaissances sur le monde, et en particulier l'ensemble des connaissances sur les situations, il devient nécessaire de doter ce système d'une dimension apprentissage si l'on souhaite le faire travailler en monde ouvert. Aussitôt, se pose la question de la cohabitation entre connaissances fournies initialement et connaissances apprises. Les premières ont été construites par un être humain qui y a investi un sens qui

échappe pour une bonne part à la machine [Grumbach 1994]. Comment dès lors celle-ci peut-elle faire évoluer et étendre ces connaissances en conservant la logique qui présidait à leur construction si celle-ci lui échappe? Une façon de répondre à cette interrogation consiste à restreindre le type d'apprentissage mis en jeu. En se limitant à l'application de méthodes du type Explanation-Based Learning (EBL), on se contente de spécialiser la connaissance que l'on possède déjà en figeant sous forme déclarative des connaissances qu'il serait possible de reconstituer à l'aide de raisonnements[1]. On ne risque pas de cette façon de remettre en cause la cohérence de la base de connaissances mais il est également clair qu'on ne pourra pas de la sorte la développer pour y intégrer de nouveaux domaines.

1.1.4. Pourquoi apprendre les connaissances sur les situations à partir de textes?

Les auteurs défendant une vision constructiviste de la cognition humaine [Vygotski 1962] [Harnad 1987] [Grumbach 1994] font généralement la distinction entre deux types d'apprentissage chez l'homme :

- un apprentissage s'effectuant à partir des interactions que nous avons avec le monde qui nous entoure. Il conduit à la formation de catégories qui ont ensuite la capacité d'être évoquées par des symboles;

- un apprentissage résultant de la manipulation et de l'association de symboles. Il suppose que ces symboles possèdent un ancrage au sein des catégories produites par le premier type d'apprentissage.

L'apprentissage des connaissances sur les situations peut revêtir ces deux formes :

- il peut résulter des situations dans lesquelles nous nous trouvons impliqués quotidiennement ou plus exceptionnellement. Nous allons travailler, nous prenons le train ou l'avion, nous allons au cinéma, nous partons en vacances. Nous vivons toutes ces situations et la représentation que nous en construisons est le fruit de cette expérience;

- il peut également être le produit de l'évocation de situations que nous n'avons jamais vécues et que nous ne vivrons jamais. Lorsque nous lisons un roman nous faisant débarquer sur une planète inconnue ou nous faisant suivre les pérégrinations

[1] C'est d'ailleurs une réponse possible au point précédent car cette spécialisation s'effectue dans le cadre d'une tâche donnée. Cela ne dispense pas toutefois de l'existence de moyens minimaux d'utilisation de la connaissance fournie pour être capable de résoudre les problèmes posés. L'EBL ne viendra en effet que fournir un surcroît d'efficacité mais ne palliera pas un manque total dans ce domaine.

d'un pilleur de banques, nous bâtissons et mémorisons une représentation de situations ne faisant pas partie de notre vécu. Cette évocation s'appuie sur des symboles qui sont soit directement ancrés dans notre expérience de l'interaction avec le monde, soit le produit de combinaisons de symboles possédant eux-même un tel ancrage.

Compte tenu de l'état actuel des travaux en Intelligence Artificielle, le premier mode d'apprentissage nous est assez clairement inaccessible pour moment. Des réalisations vont certes dans ce sens (cf. travaux relevant du courant de recherches de la vie artificielle et des animats) mais elles se situent à un niveau qui reste trop élémentaire vis-à-vis du niveau des connaissances que nous considérons ici.

Nous devons donc nous contenter du seul second mode d'apprentissage, avec de ce fait la contrainte d'une absence d'ancrage des symboles. Au lieu que leur signification soit donnée par leur ancrage au sein des catégories issues de notre immersion dans le monde, elle prend une dimension relationnelle en s'incarnant dans les relations qu'un symbole entretient avec d'autres symboles. On se rapproche ainsi de la thèse fonctionnaliste [Fodor 1981], davantage pour des raisons pratiques que par adhésion. Nous montrerons néanmoins dans la seconde partie de ce chapitre comment réintroduire une vision plus constructiviste en dépit de ces limitations.

Puisqu'il n'est actuellement pas possible à un système de s'immerger dans le monde au même titre qu'un être humain, il faut lui trouver une source équivalente de situations dans le monde symbolique. Les textes se présentent à cet égard comme une solution intéressante.

Sur un plan pratique, ils sont abondants et facilement accessibles à nos ordinateurs : l'affirmation selon laquelle de plus en plus de ressources textuelles sont disponibles sous forme électronique a ainsi été promue au rang de poncif bien établi.

Sur le fond, les textes contiennent suffisamment de traces de situations prototypiques du monde "réel" pour être exploités en vue de l'apprentissage de ces dernières. Ainsi que nous l'avons vu au §1.1.2, un texte se contente en général d'évoquer une situation en fournissant suffisamment d'informations sur celle-ci pour qu'un lecteur, compte tenu de ses connaissances, soit à même de la reconnaître et de situer les spécificités propres à la situation particulière du texte. Il est en revanche assez rare de trouver dans les textes, mis à part peut-être quand ils s'adressent aux jeunes enfants, une description détaillée d'une situation considérée comme faisant partie du fond commun culturel. Il faut donc se contenter de traces lorsqu'on adopte un tel support d'apprentissage. Fort heureusement, celles-ci varient en fonction du texte traité : suivant le contexte dans lequel la situation s'inscrit, le type de texte, le style de l'auteur, les éléments explicités ne sont pas les mêmes. Il faut donc poser comme principe que la construction de la représentation

complète d'une situation ne pourra intervenir que par recoupement de traces provenant de différents textes.

1.2. Difficultés présentées par le problème

1.2.1. Comprendre pour apprendre

Jusqu'au début des années 80 [Michalski 1977] [Mitchell 1982], les travaux touchant à l'apprentissage se sont principalement centrés autour d'une approche inductive : à partir de la description d'un ensemble d'exemples, on cherche à construire une description plus générale et plus concise qui recouvre au moins les exemples considérés initialement. On passe ainsi d'une définition en extension à une définition en compréhension de l'objet des exemples. Une variante de cette approche consiste à donner également des contre-exemples. La description généralisatrice doit alors s'attacher à couvrir tous les exemples sans inclure les contre-exemples. Mais l'induction est confrontée à une difficulté intrinsèque. Étant donné un ensemble d'exemples, il n'existe pas une seule généralisation possible. On peut généraliser à des degrés divers et selon des perspectives différentes, auquel cas on s'attache à des caractéristiques différentes des exemples.

> Soit un ensemble de pièces de monnaie provenant de plusieurs pays. Si l'on est plutôt intéressé par leurs caractéristiques physiques, on pourra les envisager comme un ensemble d'objets en métal. À un niveau très général, elles forment simplement un ensemble d'objets. En revanche, si l'on s'attache à leur fonction, on dira qu'il s'agit d'un ensemble de valeurs monétaires.

Il faut donc un moyen pour choisir entre toutes ces généralisations possibles. Ce moyen est appelé le biais d'apprentissage. Il vient contraindre la généralisation de façon à garantir le respect d'un certain critère. Celui-ci peut être purement syntaxique : on impose par exemple que les généralisations se présente sous la forme de formules logiques réduites à des conjonctions de prédicats du 1er ordre. Il s'avère plus sélectif et plus pertinent quand il traduit une certaine forme de connaissance sur le domaine concerné. Cette connaissance permet en effet de faire la part entre ce qui est important et ce qui ne l'est pas, ce qui est vrai et ce qui est faux. Elle peut se manifester de façon très implicite, à la façon d'un oracle, comme dans le cas de l'algorithme de l'espace des versions [Mitchell 1982] où ce sont les contre-exemples qui jouent ce rôle. Elle peut également être présente de façon déclarative au travers d'une théorie du domaine bien formalisée.

Les premiers travaux sur l'apprentissage ont plutôt mis l'accent sur des méthodes inductives applicables à tout domaine en dehors d'une modélisation quelconque de

celui-ci. La notion de similarité entre exemples y était donc prépondérante. C'est pourquoi on qualifie cette approche de Similarity-Based Learning (SBL). Dans [DeJong 1981] et [DeJong 1983], DeJong a néanmoins montré que ce type d'apprentissage n'était pas adapté au problème de l'apprentissage de connaissances pragmatiques à partir de textes. Cette inadaptation est visible au travers des trois points suivants :

- les textes en tant que suites de caractères ne peuvent être directement l'entrée d'un système d'apprentissage de connaissances pragmatiques. Il est indispensable de leur appliquer des traitements afin de faire apparaître ce qu'ils recèlent. Or ces traitements, assimilables à ce que nous avons défini précédemment comme étant de la compréhension, nécessitent des connaissances sur le domaine considéré. Il est dès lors évident que ces connaissances doivent également intervenir au niveau de l'apprentissage afin de le guider plus efficacement que ne le font les méthodes fondées sur la détection de similarités.

- les notions d'exemple et de contre-exemple sous-entendent l'existence d'une volonté délibérée guidant le processus d'apprentissage. On parle alors d'un apprentissage supervisé. Ceci est envisageable lorsque l'étendue des entités que l'on cherche à caractériser est bien définie. Tel n'est pas le cas avec les connaissances pragmatiques. On ne peut chercher raisonnablement à donner des exemples de chaque situation pour en construire une représentation. Outre le travail colossal que cela impliquerait, cela reviendrait à segmenter l'apprentissage en un ensemble de micro-inductions sans que l'on sache comment gérer la cohérence du tout. Par ailleurs, les textes sont des entités composites et chacun d'entre eux peut bien souvent servir d'exemple pour plusieurs situations.
  En fait, la philosophie de l'apprentissage prôné par DeJong est toute autre. Il s'agit d'exploiter les résultats fournis par un système de compréhension sur des textes qui lui sont donnés à analyser afin d'augmenter et d'améliorer les connaissances que ce système possède. D'une certaine façon, cet apprentissage peut être qualifié d'opportuniste. C'est en tout cas un apprentissage non-supervisé, rendu possible par le rôle de guide tenu par les connaissances contribuant à la compréhension.

- les méthodes reposant sur la détection de similarités entre les exemples se fondent sur le pré-requis évident et implicite que l'on dispose d'un ensemble d'exemples et non d'un seul. Or, par définition, un apprentissage opportuniste doit entrer en action dès qu'il y a lieu de le faire, c'est-à-dire à chaque traitement d'un nouveau texte dans le cas qui nous occupe. Une généralisation doit donc intervenir dès la première rencontre avec une situation afin que le système de compréhension puisse en tirer profit pour la suite. En l'absence d'un oracle, on voit d'ailleurs mal quelle serait l'utilité de stocker des représentations de texte en ignorant totalement quand

déclencher la généralisation ainsi qu'en ne sachant pas sur quelles représentations elle devrait porter.

Compte tenu de ces spécificités de l'apprentissage de connaissances pragmatiques, DeJong a proposé la notion d'apprentissage à base d'explications (EBL) : un système de compréhension de textes commence par produire une explication, c'est-à-dire qu'il explicite la chaîne causale reliant les différents événements relatés, et c'est cette explication qui est ensuite généralisée dans les limites définies par la connaissance sur le domaine ayant permis de produire l'explication. On laisse ainsi de côté les éléments ne faisant pas partie de la chaîne causale, car considérés comme anecdotiques, et l'on généralise ceux qui en font partie en s'arrêtant lorsque les relations causales ne sont plus valides.

Bien entendu, ce type d'apprentissage ne se restreint pas à l'apprentissage de connaissances à partir de textes. Il s'applique en fait dès que l'on est capable d'expliciter les relations profondes unissant les phénomènes observés, autrement dit d'en construire une explication, que ce soit pour un texte, une partie d'échecs ou un plan d'ordonnancement.

1.2.2. Un apparent cercle vicieux

Dans la conception défendue par DeJong, concrétisée notamment au travers du système GENESIS [Mooney & DeJong 1985], la compréhension est placée comme un préalable indispensable à l'apprentissage des connaissances pragmatiques. Parallèlement, nous avons montré que ces mêmes connaissances sont indispensables pour comprendre les textes au sens où DeJong l'entend, c'est-à-dire pour déterminer en quoi différents événements présents dans un même texte sont liés entre eux.

Posé de cette façon, le problème de l'apprentissage de connaissances pragmatiques à partir de textes semble enfermé dans un cercle vicieux. Il faut en effet que ces connaissances soient présentes en entrée du système alors que ce sont précisément elles que l'on souhaite voir en sortie. Or, nous ne sommes pas dans la situation où une partie d'entre elles permettraient d'en acquérir une autre partie puisque de façon évidente, les connaissances servant à la construction de la représentation d'un texte appartiennent au même domaine que celles que l'on peut élaborer à partir de cette représentation.

Cette constatation contribue à mettre en lumière le fait que l'EBL n'est pas un type d'apprentissage conduisant à construire de nouvelles connaissances. En opérant par généralisation d'explications, il s'attache plutôt à spécialiser une connaissance existante en fonction de l'utilisation qui en est faite. Il permet ainsi d'améliorer l'efficacité de systèmes déjà en fonctionnement. L'application de cette connaissance spécialisée à une situation

particulière est en effet plus directe et donc moins coûteuse en termes d'inférences. Dans ces conditions, le cercle infernal exposé précédemment ne pose plus problème dans la mesure où les connaissances en entrée ne sont effectivement pas les mêmes que celles en sortie. Néanmoins, le lien existant entre elles est si fort que l'on ne peut considérer ces dernières comme de nouvelles connaissances. Il n'y a pas extension de la théorie du domaine mais plutôt explicitation de certaines assertions qui n'étaient présentes auparavant qu'en compréhension. C'est en fait le contraire de l'induction et dans cette opération, l'étendue de ce qui est vrai et l'étendue de ce qui est faux ne changent pas.

Dès lors, la question qui s'impose, compte tenu de nos préoccupations, est : peut-on faire de l'apprentissage de connaissances pragmatiques à partir de textes en dehors du cadre de l'EBL afin d'acquérir véritablement de nouvelles connaissances?

1.2.3. Détail du cercle vicieux

Une première réponse à la question finale du point précédent consiste à examiner de plus près notre domaine d'intérêt. Nous avons montré au § 1.1.1 que les connaissances pragmatiques sont multiples et que nous ne cherchons à apprendre qu'un sous-ensemble d'entre elles, les connaissances sur les situations. Au travers de ce type de connaissances, on ne cherche pas à expliquer de façon profonde pourquoi les événements surviennent et quels liens ils entretiennent entre eux. On essaie seulement de cerner les situations prototypiques du monde de référence et de définir quels sont les événements et les personnages les composant.

Pour mettre en évidence cette dimension des textes, on imagine aisément que la tâche d'analyse est bien moins profonde que le processus de compréhension tel que nous l'avons défini plus haut. Il s'agit en effet principalement de découper les textes en fonction des situations auxquelles ils font référence. Nous appellerons cette tâche par la suite *analysethématique*. En qualifiant cette analyse de moins profonde, on sous-entend le fait qu'elle nécessite moins de connaissances et surtout, des connaissances moins sophistiquées que ne le sont les plans et les buts par exemple.

Néanmoins, cela n'implique en rien la rupture de notre cercle vicieux. Dans [Grau 1983], Grau a montré que les connaissances sur les situations sont justement nécessaires pour mener une analyse thématique fine telle que celle que nous souhaitons pratiquer. Mais du point de l'apprentissage, cette constatation nous bloque encore une fois puisque ce sont les connaissances que l'on veut acquérir qui servent à construire les représentations à partir desquelles l'apprentissage devrait opérer.

L'approche proposée par Kozima dans [Kozima 1993] et que nous avons esquissée au §1.1.2 est de ce point de vue une solution de rechange intéressante car elle repose sur une

source de connaissances, un réseau de co-occurrences lexicales, assez facile à constituer. Elle ne peut cependant être unique en ce qui nous concerne car elle présente deux caractéristiques gênantes du point de vue de l'apprentissage de connaissances sur les situations :

- son pouvoir de résolution, c'est-à-dire la taille des unités qu'elle permet de détecter, n'est ni très fort – les segments formés sont de l'ordre d'un paragraphe – ni très ajustable en fonction du texte ou du passage considéré – la taille de chacun des segments est sensiblement la même.

- elle n'est pas capable de rassembler en un même ensemble tous les segments relatifs à une même situation dès lors qu'ils ne sont pas contigus. Les textes se retrouvent donc avec une structure thématique linéaire alors qu'on reconnaît généralement le caractère hiérarchique de la structure des textes [Dahlgren 1993] [Grosz & Sidner 1986]. Ainsi, dans la configuration d'un emboîtement de thèmes, cas particulièrement courant de l'exposé d'une situation entre-coupé par l'exposé d'une situation plus détaillée, on aura trois segments alors qu'on ne devrait n'en distinguer que deux si l'on se fonde sur le critère des situations évoquées.

Les travaux de Kozima tendent malgré tout à montrer que notre objectif initial n'est pas hors d'atteinte. Il reste néanmoins à proposer une solution plus globale au problème posé.

1.3. Une solution possible

1.3.1. Les principes de la solution

Comme nous avons pu le voir au §1.2.3, un problème tel que celui de l'analyse thématique peut être résolu par des méthodes de différents types. Certaines apportent une solution complète et précise mais leur mise en œuvre pose des contraintes importantes, tandis que d'autres offrent un résultat plus approximatif mais imposent des exigences moins fortes. Dès lors que l'on prend en compte la dimension apprentissage, c'est-à-dire la possibilité pour un système de faire évoluer ses capacités, il est tentant de déterminer s'il est possible de passer progressivement du second type de solution au premier type. Il s'agit en fait d'utiliser les résultats fournis par des méthodes peu fines afin de satisfaire les contraintes posées par les méthodes plus sophistiquées et de permettre ainsi l'application de ces dernières.

Cette démarche, consistant à s'appuyer sur un niveau pour bâtir le niveau supérieur tout en conservant le même objet, est du ressort de ce que l'on appelle l'amorçage [Pitrat 1990]. L'idée sous-jacente est que l'activité d'un processus produit des éléments capables

d’alimenter un processus ayant un objectif similaire mais qui, aidé par ces moyens plus importants, peut prétendre à des résultats plus avancés.

Introduire cette notion d’amorçage au niveau de l’apprentissage de connaissances sur les situations conduit à métamorphoser notre cercle vicieux précédent en spirale ascendante. Cette transformation s’opère de la façon suivante : étant donnés un certain nombre de moyens fournis initialement, le système possède un niveau donné d’analyse qui lui permet de construire des représentations thématiques des textes. Ces représentations servent alors de support à un apprentissage, lequel produit des connaissances sur les situations. À un stade peu avancé du processus, ces connaissances sont bien entendu assez éloignées du résultat d’une modélisation manuelle : elles sont incomplètes, assez peu générales et l’importance relative de leurs constituants n’est pas très bien établie. Ces connaissances sont malgré tout réutilisées lors du traitement ultérieur de nouveaux textes afin de faire évoluer les capacités de l’analyse thématique dans le sens d’une plus grande acuité. Les représentations ainsi produites participent elles-mêmes à un processus d’apprentissage et viennent s’intégrer aux connaissances déjà élaborées lors des phases d’apprentissage précédentes. Le processus se poursuit ainsi jusqu’à obtenir des connaissances sur les situations suffisamment stables pour être généralisées et prendre une forme plus abstraite. Cette généralisation n’intervient pas globalement mais s’applique sélectivement aux connaissances accumulées qui sont jugées suffisamment établies.

1.3.2. Les contraintes pesant sur la solution

Les conséquences du principe d’amorçage

Les principes que nous avons fixés au §1.3.1 conduisent à dégager trois grandes contraintes étroitement mêlées que la solution proposée se doit de respecter :

- l’apprentissage doit se faire de façon incrémentale,
- l'apprentissage ne doit pas dépendre de fortes capacités explicatives existant a priori; autrement dit, il doit être préférentiellement de type SBL,
- le processus de compréhension doit être capable d’utiliser les résultats de l’apprentissage à mesure que ceux-ci sont produits.

La première contrainte est assez évidente au vu des principes de la solution. On suppose que les textes forment une sorte de flux continu que le processus de compréhension traite au fur et à mesure. Pour qu'un effet d'amorçage puisse intervenir, il faut que ce processus s'améliore dans le même temps. Il est donc nécessaire qu'un

apprentissage ait lieu après le traitement de chaque nouveau texte, sans attendre de disposer d'autres représentations de texte concernant les mêmes situations du monde de référence. La construction de la représentation d'une situation s'effectue en conséquence de manière incrémentale, avançant à chaque fois que cette situation est évoquée dans un nouveau texte. Seule la dernière étape de généralisation des connaissances jugées stables est réalisée en une seule étape.

La deuxième contrainte est quant à elle une conséquence de notre objectif de base : nous souhaitons créer des connaissances sur les situations qui soient véritablement nouvelles. Il ne s'agit pas de spécialiser des connaissances qui seraient de même nature que celles qui sont apprises mais plus générales. Or, seule une approche de type SBL peut être envisagée dès lors que l'on souhaite aboutir à l'acquisition de nouvelles connaissances. Dans les approches déductives tel que l'EBL, l'apprentissage est guidé par des connaissances qui sont par nature liées très fortement aux connaissances que l'on veut apprendre. On y gagne une plus grande validité du résultat mais c'est au prix d'une concession importante faite à son originalité.

En optant en faveur d'un apprentissage reposant sur la similarité, nous prenons donc le risque d'obtenir des connaissances à la validité moins bien assurée. Toutefois, nous faisons l'hypothèse que la mise en place d'un couplage très étroit entre apprentissage et compréhension ainsi que le fait de retarder le plus possible la décision finale de généralisation, au sens de la construction d'une représentation plus abstraite, peuvent atténuer significativement ce problème. Dans la mesure où ce qui est appris, en partant du produit de la compréhension, influence la façon dont le système analyse les textes, l'apprentissage est implicitement guidé par ce processus d'analyse. Les connaissances construites sont donc le résultat d'un équilibre entre apprentissage et compréhension. Retarder la décision de généraliser, ainsi que cela est mis en avant dans [Lebowitz 1988], permet à cet équilibre de s'installer.

La troisième contrainte découle comme la première assez naturellement du principe d'amorçage. Le processus d'analyse constitue l'autre moitié indispensable du couple mis en scène par ce principe. Pour qu'il joue son rôle, il doit être capable d'évoluer afin de produire des représentations de texte de plus en plus fines et complètes. On suppose en effet que ses moyens initiaux sont suffisamment réduits pour être véritablement opérationnels sur une très large étendue de textes mais son objectif final reste identifié au produit d'une analyse thématique telle que celles fondées sur les connaissances pragmatiques que nous souhaitons justement apprendre [Grau 1983]. Pour qu'une telle évolution se réalise, le processus d'analyse n'a guère d'autre choix que de s'appuyer sur ce qui s'enrichit par nature au sein du système considéré, autrement dit les connaissances en cours d'apprentissage. Cela implique aussi pour lui une capacité à utiliser des

connaissances au statut par essence instable tant du point de vue de leur consistance que du point de vue de leur complétude.

L'importance de la notion de mémoire

Du fait de notre approche, les connaissances que nous souhaitons apprendre ne se construiront que de manière progressive. Durant cette longue période de gestation, elles seront constamment utilisées à la fois par le processus de compréhension et par le processus d'apprentissage. Il sera donc nécessaire de les stocker dans une mémoire spécifique facilitant le plus possible les interactions avec ces deux processus.

L'une des caractéristiques importantes imposées par ce contexte concerne la façon dont l'accès aux connaissances sur les situations est réalisé. On a vu que dans les textes, les situations ne sont généralement qu'évoquées. Autrement dit, les textes ne font que donner quelques éléments de la situation, éléments supposés suffisants pour la faire resurgir dans son intégralité. La mémoire que nous considérons devra se conformer à ce comportement. Ce sera donc une *mémoire associative*. Cette contrainte est par ailleurs renforcée par le fait qu'elle est supposée abriter un très vaste ensemble de connaissances. Il est de fait inenvisageable de prévoir un accès de type séquentiel qui obligerait à parcourir une bonne partie de la mémoire avant de trouver l'élément recherché.

L'autre des caractéristiques majeures de cette mémoire est liée au fait que son contenu est destiné à évoluer. Constamment, de nouvelles connaissances y seront ajoutées et des connaissances plus anciennes seront modifiées plus ou moins profondément. Cela signifie que les propriétés de cette mémoire, telle son associativité par exemple, ne devront pas être seulement valides à un moment donné mais qu'elles devront se conserver en dépit de l'évolution de son contenu. Cette mémoire devra donc être *dynamique.* En particulier, il est nécessaire que les moyens qui permettent l'accès aux connaissances changent parallèlement à la transformation de ces dernières.

1.3.3. La théorie de la mémoire dynamique

Lorsque l'on se situe dans une approche mêlant étroitement compréhension de textes et apprentissage de connaissances avec un rôle important accordé à la notion de mémoire, il est difficile de ne pas évoquer la référence en la matière que constitue la théorie de la mémoire dynamique de Schank [Schank 1982]. Nous la présenterons donc rapidement dans ce qui suit (pour une présentation plus approfondie des travaux de Schank, on pourra se reporter à [Bichindaritz 1994]) puis nous mettrons en évidence ce qui nous en différencie et ce qui nous en rapproche.

Un aperçu

Les travaux sur la mémoire dynamique s'inscrivent dans le prolongement de ceux présentés dans [Schank & Abelson 1977]. Ces derniers mettaient en évidence comment différents types de connaissances, en l'occurrence les thèmes, les buts, les plans et les scénarios (cf. §1.1.1), s'articulent entre eux afin de comprendre des textes, au sens où nous l'avons défini au §1.1.2. En mettant l'accent sur le rôle des connaissances dans la compréhension, Schank et Abelson ont également fait apparaître l'importance de la notion de mémoire dans cette même tâche de compréhension. Les connaissances sont en effet abritées au sein d'une mémoire qui les organise et conditionne ainsi la façon dont on y accède. La dépendance de la compréhension vis-à-vis de ces connaissances la rend de fait étroitement dépendante de la mémoire qui contrôle leur accessibilité.

La théorie de la mémoire dynamique pousse cette logique encore plus loin en plaçant la mémoire comme objet premier d'étude. La compréhension est alors vue comme un processus véritablement fondé sur la mémoire, presque une sous-partie d'une théorie de la mémoire. Cette logique pousse Schank à s'intéresser non seulement à la façon dont la mémoire est utilisée mais également à comment elle se construit et comment elle évolue face aux différentes expériences auxquelles elle est confrontée. De fait, compréhension et apprentissage se retrouvent étroitement mêlés par l'intermédiaire de leur support commun, la mémoire.

Les relations qui unissent mémoire, compréhension et apprentissage sont illustrées dans leurs grandes tendances au travers de la figure 1.1. Lorsque de nouvelles données sont appréhendées[1], la mémoire, compte tenu de son état et des données en question, active un certain nombre d'attentes. Selon les cas, le processus de compréhension vient confirmer ces attentes ou au contraire les infirmer. Schank s'est tout particulièrement intéressé à la seconde situation dans la théorie de la mémoire dynamique car il souligne son importance vis-à-vis du problème de l'apprentissage.

Pour lui, la constatation d'une dissonance entre les attentes engendrées par la mémoire et les données qui sont traitées est en effet le signe de la nécessité pour la mémoire de s'adapter à une expérience qui lui est inédite. Cette adaptation s'effectue à partir de la construction d'une explication de cette différence. Il faut préciser que l'élaboration d'une explication n'est pas réservée au traitement des seuls échecs de la compréhension. Cette tâche est également accomplie lorsque les attentes s'accordent avec les données traitées. Plus généralement, le processus d'explication permet de mettre en évidence les éléments

[1] Ces données peuvent bien entendu prendre la forme d'un texte. Toutefois, les principes posés vont au-delà de l'unique compréhension de textes. Le terme compréhension est à prendre ici au sens large.

qui permettront de “ranger” de façon adéquate l’expérience traitée au sein de la mémoire. Cette notion d’explication est d’ailleurs une dimension que Schank développera plus particulièrement dans la suite de son travail [Schank 1986], notamment au travers du système SWALE [Schank & Leake 1989].

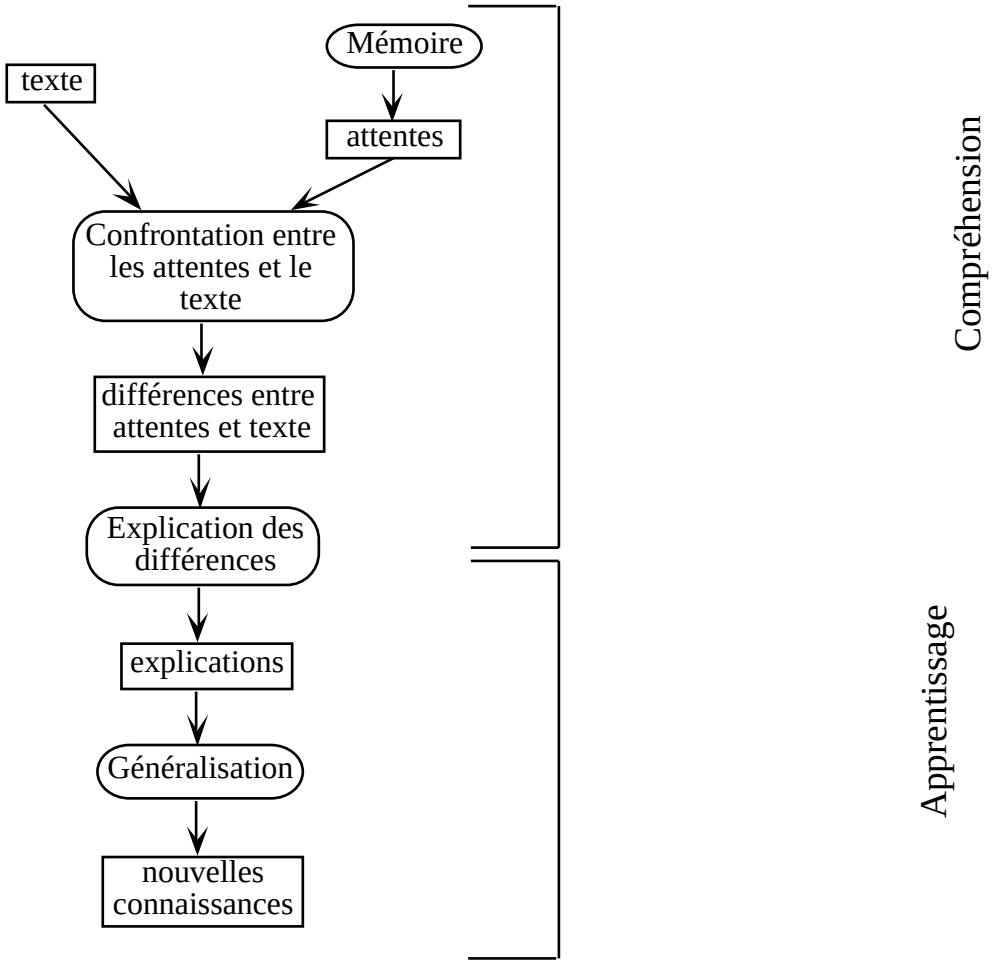


Fig. 1.1 - Relations entre mémoire, compréhension et apprentissage dans la théorie de la mémoire dynamique

En ce qui concerne l’adaptation de la mémoire face à une expérience inédite, la construction d’une explication ouvre la porte sur une première phase : l’expérience inédite est indexée au sein de la mémoire par les connaissances générales dont elle constitue une exception ou tout du moins, une variation importante. Une seconde phase prend place lorsque l’exception en question devient récurrente. Il y a alors généralisation pour faire apparaître de nouvelles connaissances générales et à la suite, réorganisation des connaissances qui servaient auparavant de référence vis-à-vis de ces exceptions.

La théorie de la mémoire dynamique spécifie également les types de connaissances organisant une telle mémoire. Ainsi que nous l’avons vu au §1.1.1, ils sont dominés par les MOPs et les TOPs. Ces structures s’inscrivent dans le prolongement de celles définies

dans [Schank & Abelson 1977]. Les TOPs apparaissent ainsi comme une tentative pour rassembler en un tout cohérent ce qui apparaissaient auparavant de façon plus dispersée sous la forme des thèmes, des buts et des plans. Ils caractérisent un type de connaissances ne dépendant pas des situations. Plus précisément, ils rendent compte des relations permettant d'articuler entre eux les buts et les plans. C'est ainsi qu'un TOP se compose de la description d'un but, ou d'une configuration de buts, ainsi que d'un ensemble de conditions portant sur les plans susceptibles de le, ou les satisfaire.

Les MOPs, quant à eux, cherchent au contraire à représenter des configurations d'événements propres aux situations. En cela, ils ont le même objet que les scénarios mais tentent de remédier à la rigidité de ces derniers qui constituaient de gros blocs de connaissances non modulaires. Les MOPs n'organisent donc plus directement des actions mais des scènes. Une scène rassemble un ensemble d'actions présentant une certaine unité de but. Par exemple, dans un MOP *visite chez le médecin,* on aura des scènes telles que *prendre rendez-vous* ou *attendre dans la salle d'attente*. Une scène peut d'autre part être partagée par plusieurs MOPs. Ce sont donc des unités de connaissance plus petits que les anciens scénarios mais également plus généraux. La notion de scénario est conservée mais elle renvoie à une instanciation particulière d'une scène. La dimension des structures organisatrices est complétée par la notion de méta-MOP qui joue, vis-à-vis des MOPs, le même rôle que les MOPs vis-à-vis des scènes.

Parallèlement à cette dimension de composition, on trouve une dimension de généralisation afin de caractériser ce qui est commun à plusieurs situations proches. Les MOPs *visite chez le médecin* et *visite chez le dentiste* peuvent par exemple être généralisés en un MOP *visite chez un professionnel de la santé* ou, à un niveau de généralité supérieur, en un MOP *visite chez un professionnel*. Ces MOPs plus généraux sont appelés des U-MOPs (Universal MOPs). Ils sont composés de scènes généralisées qui sont aux scènes ce que les U-MOPs sont aux MOPs. On obtient donc à la fois une hiérarchie de composition et une hiérarchie de généralisation. Schank précise que toutes ces structures ne sont que des discrétisations de continuums existant pour chacune de ces deux dimensions.

Position vis-à-vis de la théorie de la mémoire dynamique

Dans [Schank 1982], Schank fait la distinction, en préambule de son exposé sur l'apprentissage et la généralisation en liaison avec la mémoire dynamique, entre un apprentissage conduisant à la création de structures entièrement nouvelles à la suite d'une première rencontre avec une situation et un apprentissage se caractérisant par la modification de structures existantes par l'apport de nouvelles informations sur des situations déjà connues. Dans ce dernier cas, la modification peut être de plus ou moins grande ampleur : indexation des cas spécifiques à partir de la structure générale lorsqu'il

ne s’agit que de simples variations autour de celle-ci, ou réorganisation plus profonde faisant intervenir la création de nouvelles structures lorsque les variations deviennent récurrentes.

Le type d’apprentissage qui nous intéresse ici, celui touchant à la création de nouvelles connaissances, n’est pas celui sur lequel Schank a mis l’accent dans sa théorie. Il constate sa nécessité en tant que point départ obligé et identifie deux sous-problèmes le concernant :

- la reconnaissance de la similarité d’une nouvelle expérience avec des expériences antérieures. Schank met ici en avant la connexion forte existant chez les jeunes enfants entre une situation et leur état émotionnel et physique dans cette situation. Il considère que la reconnaissance d’une scène à ce stade, car il suppose l’apprentissage des scènes comme premier, est donc liée pour l’enfant au fait de se retrouver dans l’état qu’il a précédemment associé à cette scène;

- la focalisation de l’attention sur les traits des situations pertinents du point de vue de leur similarité. Schank avance que ce problème ne se pose pas en tant que tel puisque le jeune enfant peut développer plusieurs représentations d’une même scène, chacune étant associée à une dimension particulière pouvant être considérée comme un axe de focalisation. Il distingue ainsi la dimension personnelle (ce que ressent l’individu physique), la dimension sociale (interaction avec les autres individus) et la dimension physique (interaction avec le monde physique). On revient donc plutôt au problème de la reconnaissance abordé ci-dessus.

Bien qu’il souligne le fait que les premiers scénarios sont bâtis sur le critère de la récurrence d’un ensemble d’actions, Schank n’approfondit pas cette notion et préfère centrer l’apprentissage sur le modèle présenté au paragraphe précédent, même lorsqu'il l'applique aux jeunes enfants : l’apprentissage est vu avant tout comme un moyen pour réduire la différence existant entre les attentes que l’on peut avoir et les données auxquelles on est confronté.

De notre point de vue cependant, cette approche est insuffisante à elle seule. Ainsi que Schank le constate, la mettre en œuvre implique la possibilité d'expliquer d’où provient la différence observée entre les attentes et les données. Or, il nous semble que ces capacités explicatives ne peuvent être présentes aux stades les plus précoces, notamment du fait de l’absence des connaissances requises. Le processus d'apprentissage décrit par Schank s'inscrit dans un cadre où des connaissances de haut niveau (TOPs, MOPs, méta-MOPs, U-MOPs) existent déjà et permettent la production d'explications. Il rend compte de la façon dont ces connaissances évoluent pour s'adapter aux situations particulières qui sont rencontrées. C'est en fait une théorie de l'apprentissage tel qu'il existe chez un adulte

possédant la représentation d'un grand nombre de situations et qui n'est plus confronté qu'à des variantes de celles-ci.

Nous pensons néanmoins qu'elle n'est pas applicable à l'acquisition de situations radicalement nouvelles, que ce soit chez le jeune enfant, où c'est le cas le plus fréquent, ou chez l'adulte, où ce cas est naturellement plus rare. Par ailleurs, on peut se demander dans quelle mesure il est véritablement possible de décrire ce qui se passe à un certain stade de développement si l'on n'explicite pas finement la façon dont on peut y accéder. Il n'est en effet pas certain que les conditions fixées de la sorte corresponde à un état de développement quelconque. Cette remarque s'adresse également à des travaux tels que ceux de DeJong pour lesquels l'état de départ est le résultat d'une modélisation manuelle, donc contrôlable, sans que l'on sache cependant si celle-ci pourrait être réalisée de la même façon à large échelle.

Nous retiendrons néanmoins de la théorie de la mémoire dynamique de Schank le rôle central accordé à la notion de mémoire. Celle-ci servant de support premier aux tâches de compréhension comme aux tâches d'apprentissage, elle permet de mêler étroitement ces deux dimensions et de faire ainsi que l'une puisse s'enrichir des résultats de l'autre et vice versa.

Cette théorie illustre également l'idée importante selon laquelle l'apprentissage n'est pas une tâche cognitive spécifique et identifiable en tant que telle vis-à-vis des autres tâches cognitives que sont la perception ou le langage par exemple. Elle constitue plutôt une partie intégrante de chacune d'entre elles.

1.3.4. Une première vue d'ensemble de la solution

À ce stade de la présentation, on peut synthétiser la solution retenue par la figure 1.2 ci-après. Il faut préciser que nous nous situons là à un niveau fonctionnel ne présupposant rien sur la façon dont ces différentes fonctions seront réparties lors de leur implantation.

Au centre de cette solution, on trouve une mémoire dynamique et associative destinée à accueillir des connaissances sur les situations. Par l'action d'un processus de mémorisation, elle est capable d'intégrer la représentation que le système construit des expériences auxquelles il est confronté. Dans notre cas, il s'agit de représentations de textes et le processus les produisant est une analyse thématique, une dimension de la compréhension de textes mettant en évidence les situations auxquelles les textes font référence. Ce processus d'analyse fait lui-même appel à la mémoire pour accéder aux connaissances dont il a besoin. Cet accès est géré par un mécanisme d'extraction agissant

suivant un mode associatif. Les connaissances présentes au sein de la mémoire sont le produit d'un processus d'apprentissage incrémental opérant en sortie de l'analyse thématique afin de confronter les nouvelles représentations de texte produites avec les connaissances abritées par la mémoire, elles-mêmes construites à partir de représentations de textes plus anciennes. La mémorisation de nouvelles expériences est donc systématiquement précédée d'une phase d'apprentissage.

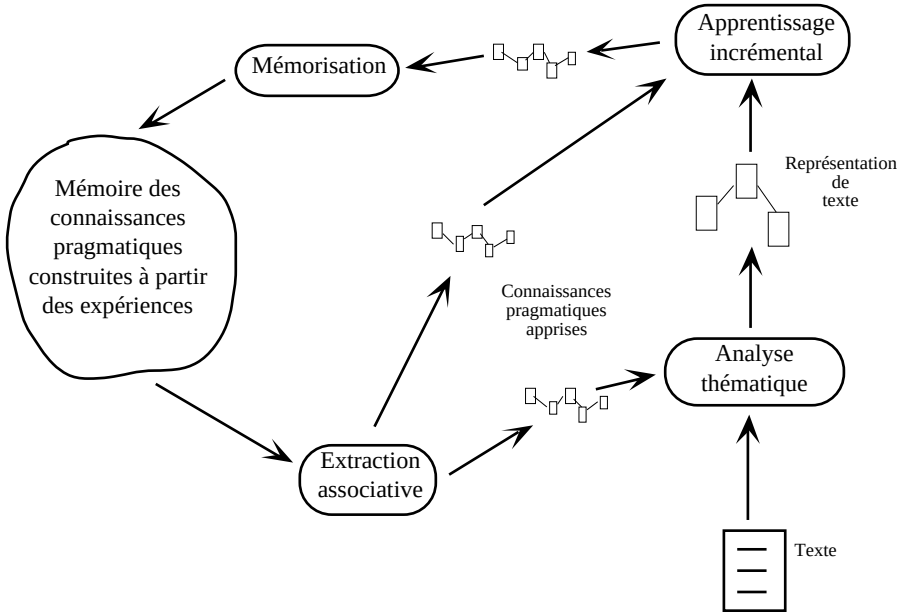


Fig. 1.2 - Vue d'ensemble de la solution retenue

Comme on peut le voir concrètement au niveau de la figure 1.2, l'ensemble forme une boucle. Celle-ci doit en réalité être interprétée en intégrant la dimension supplémentaire qu'est le temps. On retrouve alors la notion d'amorçage que l'on avait mis en avant initialement : le contenu de la mémoire au temps T est utilisé afin d'élaborer ce que ce contenu sera au temps T+1.

La mise à plat de la solution toute entière met également en évidence le rôle central qu'y occupe la notion d'expérience. C'est en effet la confrontation avec de nouvelles expériences, dans le cas présent, ce sont de nouveaux textes, qui alimente la spirale évoquée ci-dessus. Le système construit une représentation de ces expériences qui devient ensuite une source de connaissances nouvelles après passage par le processus d'apprentissage et mémorisation. Ces connaissances seront à leur tour utilisées pour construire la représentation de nouvelles expériences.

La notion d'expérience est donc particulièrement importante puisqu'elle est manipulée plus ou moins directement par tous les processus explicités ici. C'est pourquoi nous chercherons à la cerner plus précisément dans ce qui suit ainsi qu'à définir l'impact d'une

approche reposant sur cette notion tant sur le plan de l'apprentissage que sur celui de la compréhension.

2. Une approche centrée sur la notion d'expérience

2.1. La notion d'expérience

2.1.1. Définition

Dans le cours de ce chapitre, nous avons à plusieurs reprises fait usage du terme expérience mais sans chercher à préciser son sens en dehors de celui que tout à chacun possède intuitivement. Or, ce dernier présente une ambiguïté : une expérience désigne à la fois le fait de se confronter à quelque chose, en l'occurrence des textes, et le résultat à long terme de cette confrontation, ce qui correspondrait pour nous aux connaissances sur les situations accumulées dans la mémoire dont nous avons défini les caractéristiques précédemment.

Afin de clarifier le discours, nous adopterons dorénavant les conventions suivantes. Le terme *expérience* sera réservé au résultat immédiat, en termes de représentations internes, de l'interaction du système avec son environnement. Dans notre cas, une expérience s'identifiera donc à une représentation de texte. L'environnement est en effet constitué de textes et l'interaction avec ceux-ci s'effectue via le processus de compréhension qui en élabore une représentation thématique. Pour désigner le résultat à long terme de l'accumulation de ces expériences, nous utiliserons l'appellation *agrégatd'expériences*. Nous verrons plus précisément quelle est l'origine d'une telle appellation mais nous pouvons dès à présent souligner que le terme agrégat rend compte assez intuitivement du cumul d'un ensemble d'expériences.

La notion d'expérience étant définie, nous pouvons nous attacher à en cerner les caractéristiques. Ces dernières apparaissent par contraste vis-à-vis de connaissances abstraites telles que les MOPs par exemple. Globalement, le trait principal des expériences est leur contextualisation extrême. Elles sont en effet le produit d'un processus spécifique opérant sur des données particulières. De ce fait, elles sont influencées à la fois par les caractéristiques de ce processus et par les propriétés des entités dont les données considérées sont les instances. La représentation thématique d'un texte rapportant un détournement d'avion est ainsi liée conjointement à la tâche d'analyse thématique de textes et au domaine du terrorisme.

Cette contextualisation se traduit plus précisément par deux différences par rapport aux connaissances abstraites. D'abord, les éléments constituant les expériences sont plus spécifiques que ceux servant à décrire le domaine. Dans une représentation de texte à propos d'un détournement d'avion, il sera ainsi question de palestiniens du FPLP ou de membres du Djiad Islamique alors qu'au niveau des connaissances abstraites correspondantes, il ne sera fait mention que de preneurs d'otages.

Ensuite, un certain nombre de constituants des expériences ne se retrouvent sous aucune forme au niveau des connaissances abstraites pouvant exister à propos du même domaine. Il s'agit de caractéristiques propres à la situation particulière considérée mais qui ne présentent qu'un caractère anecdotique d'un point de vue général. Pour reprendre l'exemple du détournement d'avion, il se peut très bien que dans un cas, il soit question d'un Boeing 747 alors que dans un autre, il s'agisse d'un Airbus A330. Ce type de détail fait intrinsèquement partie de la situation spécifique considérée, donc de l'expérience qui lui est associée, mais n'a cependant pas d'influence sur le cours de ce type de situations et n'apparaît pas de ce fait au niveau des connaissances générales. Cet aspect est d'ailleurs un point important à prendre en compte par les processus manipulant les expériences. Ils doivent en effet intégrer le fait que les données qu'ils utilisent sont en quelque sorte bruitées.

2.1.2. Expérience et cas

La définition que nous avons donnée ci-dessus de la notion d'expérience la rapproche de façon évidence de la notion de cas distinguée en Intelligence Artificielle. Raisonner à base de cas, c'est essayer de réutiliser les expériences que l'on a construit dans le passé pour les appliquer, moyennant une certaine adaptation, à une situation présente. Les cas, comme les expériences, représentent donc une connaissance à un niveau opérationnel, connaissance dont l'exploitation nécessite la définition de moyens spécifiques. Dans ce qui suit, nous considérerons que cas et expérience sont synonymes. On se contentera d'introduire une légère nuance : un cas peut recouvrir aussi bien le résultat d'un processus que la trace de son déroulement, c'est-à-dire le détail des données et des actions que celui-ci a mises en jeu. Une expérience se limitera ici au premier aspect.

En fait, la différence entre expérience et cas ne réside pas tant dans leur nature ou leur fonction que dans la façon dont ils s'organisent au sein de la mémoire qui les abrite. Ainsi que l'expose Kolodner dans [Kolodner 1993], le schéma traditionnel d'organisation d'une base de cas fait intervenir deux composantes. L'une représente ce que l'on peut considérer comme les cas les plus fréquents ou encore, les cas de référence. L'autre est destinée à rendre compte des variantes significatives de ces cas de référence ou bien de cas plus franchement différents. La première composante prend la forme d'une

connaissance générale, souvent organisée en une hiérarchie de généralisation, tandis que la seconde est concrètement représentée par des cas rattachés aux connaissances générales dont ils expriment une spécialisation particulière. La mémoire dynamique de Schank est un exemple typique d'une telle organisation. Elle est en effet structurée par un ensemble de MOPs présentant des degrés de généralité différents, chaque MOP étant la représentation générale d'une situation concrète. À chaque MOP sont attachés des scénarios spécifiant des variantes observées de la situation dont ce MOP rend compte. Si l'on possède un MOP Restaurant[1] dans lequel on précise que l'on paye la note à la fin du repas, la rencontre d'un restaurant dans lequel on règle la note en début de repas se traduira ainsi par la création d'un scénario qui sera associé au MOP Restaurant et qui illustrera cette possible variation.

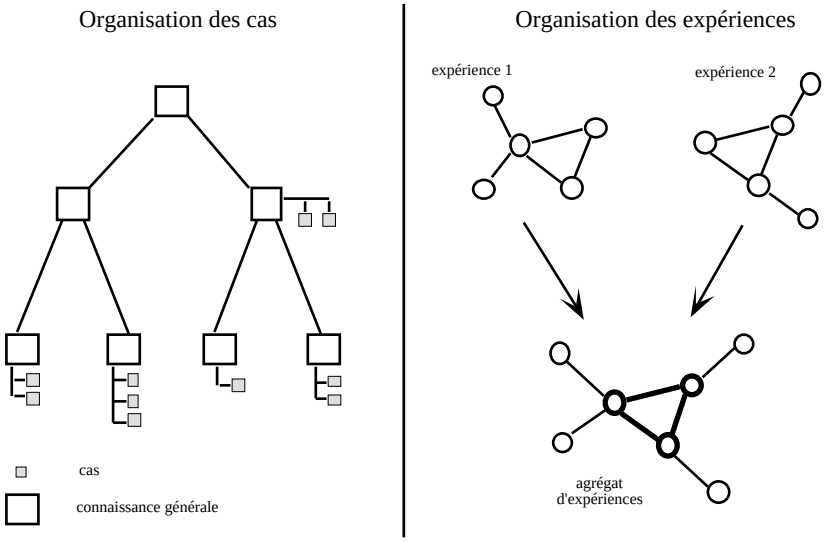


Fig. 1.3 - Différences entre la structuration des cas et celle des expériences

Ce type de structuration repose néanmoins sur l'implicite essentiel que l'on dispose des connaissances permettant de représenter les cas généraux. Or, nous avons vu que nous souhaitions ne pas nous reposer sur cette hypothèse. Une telle organisation ne peut donc être retenue pour la mémoire des expériences. Notre objectif étant précisément l'émergence des situations générales, il semble plus intéressant d'organiser les expériences en fonction de ce qui les rapproche de façon récurrente plutôt qu'en fonction de ce qui les différencient plus ponctuellement. La mémoire, dans le cas des expériences, doit donc être structurée en ensembles d'expériences ayant été jugées similaires. Ces ensembles s'identifient aux agrégats d'expériences mentionnés précédemment comme représentant l'expérience du système sur le long terme. Au sein de ces agrégats, il faut chercher à représenter de façon unique ce qui est commun aux différentes expériences formant l'agrégat tout en conservant les différences qui caractérisent chacune d'entre

[1] Toute ressemblance avec un restaurant de poissons schankien est tout à fait volontaire

elles. La figure 1.3 illustre cette différence d'organisation entre base de cas et mémoire des expériences.

2.1.3. Une source d'inspiration de nature psychologique

La notion d'agrégat d'expériences ne répond pas seulement à une apparente nécessité résultant de contraintes posées comme principes. Elle plonge également ses racines dans les travaux menés au début des années 30 par le psycholinguiste Vygotski [Vygotski 1962] à propos des rapports entre pensée et langage. Dans ce cadre, Vygotski s'est plus particulièrement intéressé au problème de la formation des concepts chez l'enfant. Il a fait à cet égard la distinction entre concepts spontanés et concepts scientifiques. Les premiers sont construits par l'interaction naturelle de l'individu avec son environnement tandis que les seconds s'élaborent à partir de définitions verbales.

La notion d'agrégat est inspirée des travaux de Vygotski sur la genèse des concepts spontanés. Celle-ci est appréhendée de façon résolument inductive comme le passage d'une notion de proto-concept conçu en tant que regroupement d'éléments défini en extension à une notion de concept s'incarnant dans une définition en compréhension de cet ensemble. Cette transformation repose sur l'explicitation des traits, prenant eux-mêmes la forme de concepts, caractérisant les regroupements d'objets.

La genèse des concepts passe par plusieurs stades que nous résumerons au travers de trois grandes étapes. L'étape initiale est celle de la formation des regroupements, que Vygotski dénomme *tas* à ce stade. Cette formation s'opère sur des critères de ressemblance plus ou moins précis mais qui n'ont dans la majeure partie des cas que peu de justification en regard des critères que pourrait appliquer un adulte. La seconde de ces étapes est une phase de maturation des premiers regroupements formés. Sous l'influence des interactions avec le monde extérieur et des interactions avec d'autres individus, les tas s'affinent pour devenir des *complexes*. Ce sont toujours des regroupements d'objets en extension mais les critères qui les réunissent se sont progressivement précisés sans pour autant avoir été explicités. La troisième et dernière étape correspond à l'explicitation des traits qui contribuent à réunir les objets d'un complexe et donc, à l'émergence d'une définition en compréhension de ce dernier. C'est le stade du *concept*.

Avec toute la prudence qui s'impose lors de telles transpositions, les agrégats d'expériences peuvent être assimilés aux complexes vygotskiens. Ils présentent la même caractéristique de regrouper un ensemble d'expériences sur des critères de similarité, ensemble dont les limites s'affinent progressivement à mesure que de nouvelles expériences viennent l'enrichir et le préciser. De même que les complexes doivent devenir des concepts par un processus d'abstraction intervenant au terme de cette phase de

maturation, les agrégats doivent à terme, c'est-à-dire lorsqu'ils seront devenus suffisamment stables, être abstraits de à façon prendre la forme de connaissances générales. Il est à noter que dans un cas comme dans l'autre, cette transformation est éminemment locale et ne concerne pas l'ensemble de la mémoire. Des entités à des stades différents y cohabitent donc. Cela contribue à la progressivité de l'apprentissage.

Pour être complet sur les relations entre agrégat d'expériences et complexe, il nous faut souligner une nuance entre les deux notions. Alors qu'un complexe est une collection d'objets liés entre eux mais où chacun conserve son indépendance, les agrégats fusionnent les différents objets qui les composent de façon à ne faire apparaître qu'une seule fois leurs parties communes tout en conservant distinctement ce qui les différencie. Cette opération s'accompagne d'une pondération des éléments constituant les agrégats en fonction de leur degré de récurrence au sein de ceux-ci. Cette pondération contribue ainsi à faire apparaître leur importance relative. On se rapproche de la position exprimée par Bordeaux dans [Bordeaux 1993] qui tente, dans le domaine des perceptions visuelles, de synthétiser les positions de Vygotski d'une part, et celles de Langacker à propos de la gradualité de la maîtrise des structures cognitives d'autre part. La gradualité caractérisant l'élaboration des agrégats n'est cependant pas conservée lors du passage vers les connaissances générales. Cette étape marque, comme chez Vygotski, une rupture nette explicable par la différence de nature assez franche entre les deux types de connaissances.

2.2. L'apprentissage à partir d'expériences

2.2.1. Expérience et apprentissage

Liens avec les méthodes générales d'apprentissage

Comme on a pu le voir au paragraphe précédent, la notion d'expérience n'est pas détachée de notions existant déjà au sein de l'Intelligence Artificielle. De ce fait, les procédures d'apprentissage qui lui sont liées ont des affinités avec des méthodes déjà existantes. Ce que nous avons déjà laissé apparaître de notre approche permet de dégager quatre grands traits de l'apprentissage à base d'expériences. Il est :

- incrémental,
- non-supervisé,
- intégré étroitement à la tâche qui utilise les expériences;
- il s'appuie le moins possible sur des connaissances de même nature existant a priori.

Ces caractéristiques le rapprochent fortement de ce que [Gennari et alii 1989] nomme formation de concepts (“concept formation”). Ce mode d’apprentissage se donne pour objectif de créer progressivement une hiérarchie de concepts à partir d’instances considérées incrémentalement et de façon non supervisée. Chaque nœud de cette hiérarchie contient une description intensionnelle du concept représenté par ce nœud en même temps qu’il pointe vers les instances ayant conduit à le former. Les méthodes de formation de concept représentent à cet égard une spécialisation des méthodes de regroupement conceptuel (“conceptual clustering”) orientée vers la mise en œuvre d’un apprentissage incrémental et non supervisé.

Afin de concrétiser notre propos et situer plus précisément les particularités de notre approche par rapport à ces méthodes de formation de concept, nous allons détailler l’un de ces premiers représentants, le système UNIMEM [Lebowitz 1986]. Celui-ci s’inscrit dans un cadre plus large, appelé “Generalization-Based Memory” (GBM), lui-même héritier des travaux sur la mémoire dynamique, notamment par filiation avec IPP [Lebowitz 1983].

Le fonctionnement d’UNIMEM est schématiquement le suivant. Lorsqu’une nouvelle instance lui est présentée, il parcourt en profondeur d’abord la hiérarchie des concepts déjà existants à partir de son sommet de façon à trouver le concept le plus spécifique dont la définition s’accorde avec les propriétés de l’instance en question. Cette concordance est évaluée par une distance adaptée au type des propriétés (numérique, symbolique, booléenne, ...). À chaque trait d’un concept est associé un compteur traduisant le degré de confiance que l'on accorde à ce trait. Si la valeur de ce compteur tombe en dessous d’un certain seuil, la propriété est éliminée de la définition du concept. Si toutes les propriétés sont éliminées, c’est le concept lui-même qui disparaît. Lors du parcours de la hiérarchie pour trouver le concept s’accordant avec une nouvelle instance, la traversée d’un nœud, appelé Gen-Node, provoque l’incrémentation du compteur des traits en accord avec ceux de l’instance et au contraire, la décrémentation de ceux qui sont jugés contradictoires par rapport à ceux de l’instance. Lorsque le concept le plus spécifique compatible avec la nouvelle instance a été trouvé, cette instance est rattachée au Gen-Node correspondant. Les Gen-Nodes sont donc caractérisés à la fois par une définition intensionnelle sous la forme de traits pondérés et par une définition extensionnelle sous la forme de l’ensemble des instances dont ils sont la généralisation.

Lors du rattachement d’une instance à un Gen-Node, on examine si les traits de celle-ci qui sont nouveaux ou en désaccord par rapport à ceux du Gen-Node ne sont pas également partagés par d’autres instances de ce même Gen-Node. Si c’est le cas de façon significative, de nouveaux Gen-Nodes sont créés pour chaque regroupement possible. De ce fait, une instance peut être commune à plusieurs Gen-Nodes. Les nouveaux

Gen-Nodes ainsi créés sont reliés à celui qui leur a donné naissance par une relation de spécialisation.

À côté des points de similitude évoqués ci-dessus, cette description laisse apparaître quelques différences par rapport à ce que nous avons esquissé précédemment. Au premier plan de celles-ci se trouve la volonté, ceci est vrai pour UNIMEM mais également plus largement pour toute l'approche formation de concepts, de construire une hiérarchie de concepts. Ce point conditionne assez fortement les méthodes utilisées mais ne constitue en revanche ni un objectif, ni une contrainte que nous nous fixons. Compte tenu du caractère intrinsèquement bruité des expériences, nous cherchons avant tout à faire apparaître la description d'un ensemble de situations en faisant l'hypothèse que les relations de composition (ce qui rend compte du contenu des situations) priment en l'occurrence sur les relations de généralisation. Celles-ci pourront être mise à jour dans un second temps à partir de ce premier ensemble de situations. Par ailleurs, on peut douter fortement de la possibilité, et même de l'intérêt, car les niveaux les plus hauts ne sont alors guère significatifs, de faire apparaître une hiérarchie unique embrassant de façon cohérente toutes les situations. Au mieux, il faudra sans doute se contenter d'une forêt d'arbres, voire d'une forêt de treillis, car il n'est pas évident de privilégier un seul point de vue ainsi que le fait apparaître Lebowitz à propos des instances.

Une autre différence en apparence notable réside dans l'élaboration immédiate par UNIMEM d'une définition intensionnelle pour chaque nouveau concept construit alors que cette étape n'intervient, dans notre conception, qu'à l'issue de la maturation d'un agrégat d'expériences. Cette différence est moins marquée qu'il n'y paraît cependant. La définition intensionnelle d'un concept est constituée pour UNIMEM de la généralisation de l'ensemble des traits communs aux instances qu'il regroupe, chaque trait étant pondéré par un facteur de confiance. Toutefois, cette définition évolue à mesure que le système rencontre de nouvelles instances : des traits supplémentaires peuvent apparaître lorsqu'une nouvelle instance est associée à ce concept, le poids des traits déjà présents évolue en fonction de la constitution des instances qui "traversent" le concept et les traits peuvent même disparaître si leur poids devient trop faible. En utilisant ce mécanisme de poids des traits variant en fonction de leur récurrence, l'évolution de la définition des concepts se rapproche donc fortement de la façon dont les agrégats d'expériences eux-mêmes évoluent.

En ce qui concerne la façon dont les regroupement d'instances sont gérés, la principale différence entre les deux approches se situe au niveau du principe régissant la création de nouveaux regroupements. Dans le cas d'UNIMEM, une telle création s'effectue dans l'optique de spécialiser un concept déjà existant. Ce souci est directement hérité de la volonté de construire une hiérarchie. Étant dégagé d'une telle obligation, nous nous

contenterons dans notre cas de recourir à la mesure de similarité utilisée pour construire les agrégats d'expériences. Le traitement d'une nouvelle expérience conduit à installer un contexte, constitué notamment des agrégats de la mémoire particulièrement liés à l'expérience considérée sur le plan thématique. Si la nouvelle expérience est similaire à l'un de ces agrégats, elle y est intégrée. Dans le cas contraire, elle forme un nouvel agrégat.

Liens avec l'apprentissage caractéristique du raisonnement à base de cas

Nous avons vu précédemment que la notion d'expérience entretient des liens étroits avec celle de cas. En accord avec notre volonté de définir la notion d'apprentissage à partir d'expériences, il nous a donc semblé intéressant de nous pencher sur la façon dont l'apprentissage est considéré dans le raisonnement à base de cas.

Dans [Bichindaritz 1994], Bichindaritz a mis en évidence les spécificités d'un tel apprentissage par rapport au cadre présenté dans [Kodratoff & Michalski 1990]. Elle a en particulier montré qu'essayer de situer le raisonnement à base de cas en termes de type global d'inférence, synthétique ou analytique, n'est pas adapté. Le raisonnement à base de cas se décompose en tout un ensemble de sous-problèmes à propos desquels un apprentissage peut avoir lieu. En fonction du système considéré, cet apprentissage peut être plutôt analytique ou plutôt synthétique. Il peut même arriver qu'au sein du même système, l'apprentissage relatif à un sous-problème soit effectué de façon synthétique alors que pour un autre sous-problème, il est réalisé de manière analytique.

Globalement, on peut localiser l'intervention de l'apprentissage dans le raisonnement à base de cas au niveau des trois dimensions suivantes :

- les cas eux-mêmes. La forme d'apprentissage la plus évidente est ici l'ajout de nouveaux cas, mais il faut également prendre en considération leur généralisation ainsi que leur affinement, c'est-à-dire la mise en évidence de nouveaux traits après leur intégration en mémoire via l'apport d'informations réalisé par de nouveaux cas. Dans ces deux dernières formes, la capacité à expliquer, caractérisée par celle de mettre en relation les éléments constituant les cas ainsi que par celle d'expliciter certains éléments non exprimés, joue un rôle particulièrement important;

- les mécanismes de manipulation et d'évaluation des cas intervenant au cours du raisonnement. Il s'agit par exemple de la distance permettant d'établir la similarité de deux cas. Ce peut être également des facteurs déterminant la façon dont un cas est adapté à une nouvelle situation. L'apprentissage s'effectue généralement par rétroaction des succès et des échecs du système sur les paramètres influençant le mécanisme considéré, ce qui pose le problème bien connu du "credit assignment";

- la structuration de la mémoire abritant les cas. L'objectif est ici d'apprendre à organiser la mémoire de façon à en extraire le plus efficacement possible un cas pertinent vis-à-vis de l'accomplissement de la tâche en cours. La solution passe classiquement par la détermination des index favorisant cette remémoration future. Les méthodes de type regroupement conceptuel sont un moyen de faire émerger ces index, de même que l'assignation de bonus ou de malus à un ensemble d'index prédéterminé en fonction de l'adéquation des cas extraits. Mais la présence de suffisamment de connaissances permet de les sélectionner de façon plus argumentée sur la base des explications que l'on peut alors produire.

L'apprentissage à partir d'expériences se différencie des approches les plus répandues dans le cadre du raisonnement à base de cas au travers principalement de la première et de la dernière dimension.

Compte tenu du principe de formation des agrégats d'expériences, l'apprentissage concernant les cas eux-mêmes ne fait pas de différence entre l'ajout de nouveaux cas, leur généralisation et leur affinement. Ces trois aspects sont en fait mêlés très étroitement du fait même de la notion d'agrégat. L'ajout se traduit en effet par une agrégation avec des expériences plus anciennes, opération qui donne lieu à une généralisation implicite du fait de la mise à jour de la pondération des éléments constituant l'agrégat. Par ailleurs, chaque nouvelle expérience d'un agrégat lui apporte des éléments nouveaux qui contribuent ainsi à son affinement. La pertinence de ces éléments sera jugée sur le long terme, à la lumière des expériences qui viendront renforcer l'agrégat par la suite.

L'apprentissage à partir d'expériences se singularise également sur le point de vue de la structuration de la mémoire. Puisque nous avons opté en faveur d'une hypothèse minimaliste à propos des connaissances existant a priori sur le domaine, nous ne pouvons nous reposer sur une structuration initiale de la mémoire telle qu'elle existe dans la théorie de la mémoire dynamique de Schank et nous contenter d'apprendre à indexer le mieux possible les cas compte tenu de ce cadre. Il n'est pas non plus question d'avoir recours à un vocabulaire d'indexation résultant d'une modélisation du domaine ou d'une catégorie de domaines comme c'est le cas avec le "Universal Index Frame" (UIF) [Schank & Osgood 1990] pour tout ce qui concerne les interactions entre les agents et leurs buts.

Il faut donc qu'à la fois la structuration de la mémoire et le vocabulaire d'indexation soient les produits d'un processus d'émergence. La notion d'agrégat d'expériences offre naturellement une solution pour ce qui est de l'organisation de la mémoire. On remarquera toutefois qu'on obtient ainsi une structure "à plat". On se contente de réaliser des regroupements d'expériences mais ces regroupements ne sont pas eux-mêmes organisés au-delà de la notion d'ensemble. Une telle structure ne semble pas de prime abord faciliter

l’extraction des expériences, en particulier si l’on s'attache à l’efficacité de cette opération. Kolodner aborde cette question dans [Kolodner 1993] et propose trois pistes pour y remédier :

- l’indexation de surface (“shallow indexing”). Lors de l’opération de recherche dans la base de cas, on s’appuie sur les index existant pour s’orienter vers les cas les plus intéressants compte tenu du contexte courant. Cependant, il est nécessaire en final de calculer une similarité entre les cas ainsi obtenus et la représentation du problème considéré afin de choisir plus précisément celui ou ceux que l’on retiendra. Cette opération est bien entendu coûteuse et justifie la volonté de minimiser le nombre de fois où elle intervient en restreignant le plus possible l’espace de recherche à l’aide des index. L’indexation de surface propose une approche différente en cherchant à minimiser le coût du calcul de similarité, ce qui permet de l’appliquer plus largement. Cette réduction est obtenue en ne faisant intervenir que les traits de surface des cas et non l’intégralité de leur structure. En procédant de la sorte, on peut sélectionner un nombre restreint de cas auxquels on applique ensuite une mesure de similarité plus profonde pour ne retenir en final que celui ou ceux qui seront vraiment utilisés;

- la partition de la base de cas. L’objectif est ici de définir un certain nombre de critères en fonction desquels on divise la base de cas en plusieurs sous-ensembles de taille plus réduite. Il faut d’autre part avoir la capacité de juger, pour chaque nouveau cas considéré, de quel(s) sous-ensemble(s) de la base de cas il est le plus proche. La recherche est alors limitée au(x) sous-ensemble(s) sélectionné(s), ce qui améliore notablement son efficacité;

- l’utilisation du parallélisme. Cette solution ne vise pas à réduire la complexité intrinsèque de la tâche de recherche mais plutôt à exploiter un environnement matériel spécifique. L’existence d’ordinateurs reposant sur un parallélisme massif, notamment les machines SIMD, ouvre la possibilité de mener en parallèle sur tous les cas composant une base de cas une opération telle que l’évaluation de la similarité entre un nouveau cas et un cas de la base. Pour que cette approche soit viable, il est nécessaire que chaque processeur n’ait en charge qu’un nombre limité de cas. Idéalement, on devrait avoir autant de processeurs que de cas dans la base.

Dans l’approche à base d’expériences, on observe une convergence entre la première de ces trois pistes, le problème de l’émergence du vocabulaire d’indexation et celui de l’apprentissage des cas eux-mêmes. On a vu que l’apprentissage des expériences conduit à former des agrégats d’expériences au sein desquels chaque élément constitutif des expériences est pondéré en fonction de son degré de récurrence parmi toutes les expériences formant l’agrégat. Ces éléments pondérés forment naturellement un

vocabulaire pour une indexation de surface. La mémoire doit alors être organisée de façon telle qu'à chaque trait présent dans une expérience au moins, on rattache toutes les expériences possédant ce trait, avec son poids au sein de l'expérience. Lors de la recherche en mémoire d'une expérience similaire à une nouvelle expérience, on collecte toutes les expériences possédant les traits de la nouvelle expérience en exploitant pour cela la structure de la mémoire. On peut ensuite calculer pour chacune une mesure de similarité de surface tenant compte du nombre de traits partagés avec la nouvelle expérience et du poids de ces traits. Un premier ensemble d'expériences proches est ainsi dégagé, expériences auxquelles on pourra appliquer une mesure de similarité plus sophistiquée afin de juger de leur similarité profonde avec la nouvelle expérience.

Du point de vue du raisonnement à base de cas, l'apprentissage à base d'expériences permet donc de faire émerger un vocabulaire d'indexation de surface en même temps que l'apprentissage des cas s'opère et que la structure de la mémoire s'élabore.

2.2.2. Le modèle MoHA

Dans les paragraphes qui précèdent, nous avons présenté l'apprentissage à partir d'expériences par contraste et similitude avec des travaux existant dans des paradigmes proches. Dans ce qui suit, nous décrirons dans ses grandes tendances le modèle MoHA (Modèle Hybride d'Apprentissage), un modèle d'apprentissage qui s'appuie sur la notion d'expérience et qui sert de cadre à notre travail.

Les lignes directrices

MoHA est initialement le fruit des travaux de Françoise Forest et de Brigitte Grau [Forest & Grau 1992] sur le problème de l'apprentissage de concepts et de connaissances pragmatiques à partir de données verbales en s'inspirant des idées de Vygotski [Vygotski 1962] dont nous avons donné un aperçu au §2.1.3. La prise en compte d'expériences perceptives de nature visuelle a été étudiée par François Bordeaux dans [Bordeaux 1993]. Cette voie a été ensuite poursuivie par Françoise Forest en considérant des expériences visuelles de plus haut niveau s'attachant à rendre compte des relations topologiques entre les constituants d'une scène visuelle [Forest 1997]. La composante d'apprentissage de concepts à partir de données verbales a été plus spécifiquement développée par Jean-Pierre Gruselle [Gruselle 1997] à partir du modèle proposé par Françoise Forest dans [Forest 1991]. Celle concernant l'apprentissage de connaissances pragmatiques à partir de textes, conjointement développée avec Brigitte Grau [Ferret & Grau 1997], fait l'objet d'une partie du travail exposé ici.

Depuis son origine, MoHA se fonde sur trois grands principes :

- l'apprentissage des connaissances doit se faire selon une approche constructiviste;
- l'apprentissage doit prendre en compte le fait que les différents types de connaissances sont en interaction;
- il existe une liaison nécessaire entre l'apprentissage des connaissances d'une part, et leur utilisation dans des processus de compréhension, au sens large, d'autre part.

Le premier de ces principes vient s'opposer à l'idée selon laquelle l'apprentissage ne résulterait que de la spécialisation ou de la sélection de structures déjà existantes [Fodor 1975]. Il postule au contraire que ces structures sont construites progressivement à partir de l'interaction de l'apprenant avec le monde dans lequel il se trouve immergé. Ces interactions se déroulent tant au niveau physique qu'au niveau social, au travers des contacts avec d'autres individus. Le langage est un des canaux de ces interactions, canal particulièrement important dans MoHA, tout du moins dans sa forme actuelle. Ses contacts avec le monde laisse chez l'apprenant une trace sous la forme d'expériences. Leur accumulation, leur confrontation et leur appariement sont la source de son apprentissage. Celui-ci se caractérise donc par sa gradualité et par l'importance qu'il accorde à la détection de similarités entre expériences pour bâtir de nouvelles connaissances.

En se fondant sur l'expérience de l'apprenant, cet apprentissage se distingue également par la place qu'il confère à la subjectivité. Les connaissances ainsi acquises seront propres à un individu et dépendantes de son histoire personnelle. Elles pourraient donc être différentes pour un autre individu soumis à des expériences similaires ou pour le même individu ne les rencontrant pas dans le même ordre. Ces caractéristiques sont bien entendu délicates à prendre en compte, voire indésirables, si l'on adopte un point de vue plus formel sur l'apprentissage mais nous pensons qu'elles surviennent inévitablement dès lors que l'on met en avant la notion de mémoire, celle de progressivité de l'apprentissage et que l'on ne cherche pas uniquement à adapter un cadre existant.

Le deuxième principe vient de la constatation suivante. Les différents types de connaissances sont en interaction, que ce soit sur le plan fonctionnel, sur le plan structurel ou sur le plan de leur formation. Les MOPs, par exemple, représentent des connaissances pragmatiques mais sont formés de concepts. À l'inverse, les concepts sont le résultat de l'abstraction d'entités présentes sous des formes plus ou moins diverses dans différentes situations. On a vu d'autre part que la compréhension automatique d'une langue met en jeu des connaissances de natures diverses, aussi bien linguistiques, conceptuelles que pragmatiques.

Dans le cadre de MoHA, on fait l'hypothèse que l'apprentissage ne peut faire abstraction de ces liaisons. C'est évident lorsqu'un type de connaissances sert à l'élaboration d'un autre mais ce principe va au-delà. L'apprentissage ne doit pas s'attacher qu'à un seul type de connaissances mais intégrer au contraire l'ensemble des types de connaissances qui sont en interaction, et de ce fait, complémentaires. Ce principe est d'ailleurs une conséquence du premier principe posé. Il est en effet difficile de prôner une approche constructiviste pour une partie des connaissances auxquelles on s'intéresse et ne pas l'appliquer à une autre partie de celles-ci alors que les unes servent de support aux autres et vice versa.

MoHA est donc un modèle d'apprentissage faisant intervenir des connaissances de natures variées. Celles-ci se définissent par rapport aux trois dimensions suivantes :

- le type des connaissances. On reprend ainsi la distinction opérée au §1.1.1 entre connaissances conceptuelles et connaissances pragmatiques. En fait, il faut préciser que cette distinction s'identifie plutôt à une différenciation progressive dans le cas présent. Lorsque par exemple, le niveau des expériences s'incarne dans un réseau de co-occurrences lexicales, les dimensions pragmatiques et conceptuelles se trouvent de manière évidente étroitement mêlées. La liaison existant entre deux mots peut aussi bien être le résultat d'une propriété générale exprimée à propos d'une entité que traduire l'appartenance de deux entités à une même situation. Les expériences ne sont pas de nature conceptuelle ou de nature pragmatique à la base. Elles servent de support à des processus d'apprentissage qui ont en revanche pour tâche de faire émerger un type de connaissances donné. Il est également assez clair que toutes les dimensions des expériences ne contribuent pas de façon égale à l'émergence de tous les types de connaissances. Cela ne remet toutefois pas en cause la nature globalement neutre des expériences vis-à-vis de la dichotomie entre connaissances conceptuelles et pragmatiques.

- le niveau des connaissances. On fait référence ici à leur degré d'abstraction et de généralité. MoHA fait bien entendu intervenir des expériences mais également le produit de leur abstraction, concrétisé au travers d'un réseau sémantique pour ce qui est des connaissances conceptuelles et d'un réseau de schémas analogues aux MOPs dans le cas des connaissances pragmatiques. Là encore, il existe une certaine gradualité entre expériences et connaissances abstraites. L'abstraction en elle-même n'est pas un processus graduel, conformément d'ailleurs au passage du complexe au concept chez Vygotski. En revanche, l'élaboration des complexes, et dans notre cas des agrégats d'expériences, est progressive. Or, ces agrégats représentent déjà une certaine forme de généralisation, même si elle est largement implicite, des expériences individuelles;

- l'origine des connaissances. Dans son état actuel de développement, MoHA privilégie les expériences de nature verbale. Cette tendance n'est pas toutefois un parti pris résolu mais résulte plutôt de la difficulté évoquée au §1.1.4 de travailler à partir d'autres modalités. Ainsi que nous l'avons évoqué précédemment, des travaux ont d'ailleurs été menés et continuent d'être menés sur l'intégration d'expériences visuelles au sein de MoHA mais leur lien nécessaire avec les composantes verbales reste encore à étudier plus profondément. Sur le plan des principes cependant, la possibilité de traiter des expériences issues de différentes modalités perceptives et surtout, de mettre en rapport pour une même expérience les produits de plusieurs modalités, reste un objectif important d'un modèle tel que MoHA.

La volonté de prendre en compte dans MoHA des connaissances de natures différentes tant au niveau des tâches de "compréhension" que de l'apprentissage implique également de faire cohabiter des paradigmes de représentation que l'on n'associe pas habituellement. On peut ainsi avoir aussi bien des connaissances sous une forme symbolique, classique en Intelligence Artificielle, que sous une forme numérique ou topologique. Cette cohabitation des représentations entraîne aussi la cohabitation des processus qui les construisent ou les utilisent, qui sont donc eux-mêmes de différents types. C'est pourquoi MoHA peut être vu globalement comme un *modèle hybride*.

Le troisième et dernier grand principe fondateur de MoHA reprend la nécessité, que nous avons développée au §1.2, d'une liaison forte entre apprentissage et compréhension, ou plus généralement entre apprentissage et utilisation de ce qui est appris. Cette dépendance apparaît d'ailleurs naturellement. Pour apprendre, il faut d'abord comprendre, c'est-à-dire passer de données brutes à une représentation compatible avec ce que l'on cherche à apprendre. Pour comprendre, un système utilise des connaissances. Or, ce sont ces mêmes connaissances que l'apprentissage contribue à forger ou à faire évoluer. MoHA ne se limite donc pas à un ensemble de mécanismes d'émergence mais illustre également comment ceux-ci s'articulent avec des mécanismes de compréhension. Le modèle de formation des concepts à partir des mots est ainsi en relation étroite avec un processus de désambiguïsation lexicale. De même, le modèle d'émergence de schémas à partir de représentations de textes est en rapport étroit avec un processus d'analyse thématique des textes.

Description générale

Architecture générale

Dans la description que nous tracerons ici de MoHA, nous ne ferons pas apparaître les expériences visuelles. Leur définition et leur traitement ont fait l'objet de travaux mais seule la composante verbale est pour le moment suffisamment développée.

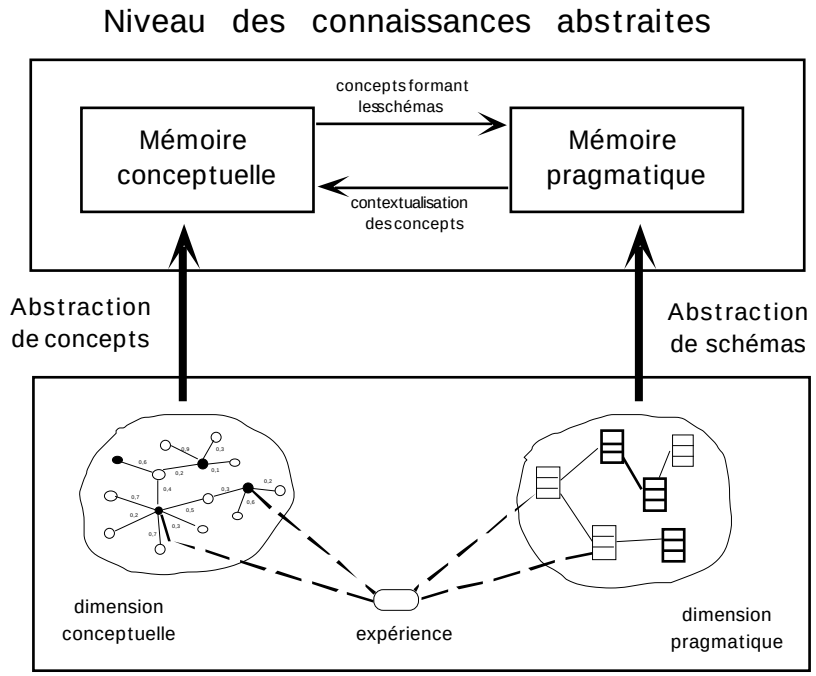


Fig. 1.4 - Architecture de MoHA

Ainsi que le montre la figure 1.4, les connaissances présentes au sein de MoHA se répartissent en deux niveaux distincts. Comme son nom peut le laisser penser, le niveau des expériences rassemble l'ensemble des expériences que MoHA a construites suite aux interactions qu'il a eu avec son environnement. Celui-ci n'est composé pour le moment que de matériaux verbaux : textes ou énoncés oraux retranscrits. Chaque expérience peut comporter plusieurs dimensions, sans que cela soit néanmoins systématique. Nous avons fait apparaître sur la figure les deux possibles actuellement : la dimension conceptuelle de la composante verbale et la dimension pragmatique de cette même composante. La première est le point de départ du processus d'émergence des concepts de la mémoire conceptuelle tandis que la seconde est le point de départ du processus d'émergence des schémas de la mémoire pragmatique.

La structuration du niveau des expériences est donc guidé par les conditions d'apprentissage des connaissances qu'il contient. On lie ce qui a été acquis au cours de la

même expérience, c'est-à-dire dans une même unité de temps, d'espace et d'action. Le niveau des connaissances abstraites se caractérise au contraire par sa relative indépendance vis-à-vis de l'origine des connaissances qui le composent. Il est organisé en fonction du type de ces connaissances. On y fait ainsi la distinction entre les connaissances sur les propriétés générales des objets et des actions et les connaissances sur les situations prototypiques. Mais il y importe peu de savoir qu'un concept a été construit sur la base d'expériences essentiellement visuelles ou au contraire verbales, ou qu'il est le produit de plusieurs agrégats d'expériences ou d'un seul. On cherche principalement, au sein de ce niveau, à expliciter les conditions nécessaires et suffisantes permettant de caractériser les entités que l'on veut représenter.

Cette caractéristique des connaissances abstraites ne signifie pas cependant l'absence de tout lien entre leur niveau et celui des expériences. Les connaissances abstraites conservent des références vers les expériences dont elles sont issues et les manipulations qui les concernent peuvent être répercutées au niveau de ces expériences afin de bénéficier de la plus grande richesse de celles-ci et de leur plus grande adaptation à un contexte spécifique. Les connaissances abstraites représentent une sorte de synthèse, réalisée à un maximum supposé de stabilité, d'un ensemble d'expériences mais elles ne prétendent pas en capter tous les aspects. Elles constituent plutôt une structuration supplémentaire permettant de mettre en œuvre des manipulations à une échelle plus vaste.

Ces manipulations concernent également des tâches d'apprentissage. Le fait d'abstraire des expériences ne signifie pas en effet que le résultat de cette opération soit figé à jamais. Les connaissances abstraites sont elles aussi soumises à une évolution mais celle-ci s'opère différemment de ce qui se passe au niveau des expériences. Alors que dans ce dernier cas, il s'agit essentiellement d'un processus progressif, presque continu, de formation, l'apprentissage touchant les connaissances abstraites s'effectue par restructurations explicites et discrètes.

Le niveau des expériences

Dans son état actuel d'élaboration, le niveau des expériences de MoHA présente une restriction par rapport à la description esquissée ci-dessus. Il n'existe en effet pas de lien entre les différentes composantes des expériences. Plus précisément, les expériences ne comportent qu'une seule dimension. On a donc des expériences "conceptuelles" d'une part, et des expériences "pragmatiques" d'autre part. Néanmoins, rien ne s'oppose à la mise en œuvre de la conception développée. En particulier, les deux types d'expériences s'appuient chacun sur la notion de situation qui doit permettre de réaliser le pont entre elles à un niveau plus fin qu'une simple liaison entre expériences de types différents.

La dimension conceptuelle des expériences verbales

La dimension conceptuelle des expériences verbales est un point de départ du processus d'émergence des concepts formant la mémoire conceptuelle [Gruselle 1997]. Son contenu est donc le reflet du souci de capturer, au travers des mots formant une expérience verbale, les concepts qui sont évoqués, cela afin de pouvoir les expliciter par la suite. Cette capture est réalisée en mémorisant les mots utilisés au cours de l'expérience et en les associant à la représentation de la situation au cours de laquelle ils sont intervenus. Une expérience verbale s'inscrit en effet dans le cadre d'une ou de plusieurs situations.

Les mots et les situations sont représentés chacun par un type de nœud spécifique et l'ensemble des expériences forme un graphe bi-partite mot/situation, étant entendu que chaque mot de la langue n'est présent qu'une seule fois dans ce graphe. Il faut ajouter que la liaison entre un mot et une situation est pondérée afin de rendre compte de la saillance de ce mot dans le contexte de la situation. Ce poids constitue initialement une forme de résumé de la perception des informations non-verbales (prosodie, attitudes gestuelles, ...) qui accompagnent le message linguistique et contribuent à mettre en relief ses différents composants. Par la suite, il évolue en fonction des corrélations trouvées entre ce mot et d'autres mots appartenant à de nouvelles expériences.

Un concept ayant émergé de cette dimension prend la forme d'un sous-graphe du graphe bi-partite évoqué ci-dessus. Sur le principe, le processus d'émergence des concepts opère sur la base de la confrontation, pour un ensemble de mots, des situations dans lesquels ils sont intervenus. Grossièrement, un concept correspond donc à un ensemble de mots présents simultanément dans un nombre significatif de situations. Employé dans le contexte d'une situation, un mot n'est pas ambigu et peut d'une certaine manière être identifié au concept qu'il dénote. On fait à partir de là l'hypothèse que la récurrence d'une configuration de mots entre plusieurs situations est le signe que chacun d'entre eux désigne dans chacune de ces situations le même concept.

Ce processus d'émergence s'appuie sur un mécanisme de propagation d'activation agissant dans le graphe pondéré mot/situation. Un processus de désambiguïsation lexicale, fondé également sur une propagation d'activation, utilise le même principe de récurrence d'une configuration de mots entre situations que le processus d'émergence.

La dimension pragmatique des expériences verbales

Nous ne la détaillerons pas ici puisqu'elle constitue une partie du travail qui sera présenté par la suite. Nous nous contenterons de rappeler qu'elle est organisée autour de la notion d'agrégats de représentations de textes. Ces représentations de textes sont de nature thématique. Elles mettent en évidence les situations évoquées par les textes. Leurs constituants élémentaires sont les concepts de la mémoire conceptuelle. Les agrégats sont

formés sur la base de la similarité entre représentations de textes. En leur sein, celles-ci sont appariées et stockées de façon à ce que leurs parties communes n'apparaissent qu'une seule fois. Ce processus s'accompagne d'une pondération des constituants de ces représentations en fonction du degré de récurrence de ces constituants au sein des différentes expériences.

Le niveau des connaissances abstraites

Nous décrirons ici les grandes caractéristiques fonctionnelles la mémoire conceptuelle et de la mémoire pragmatique du point de vue de MoHA. Nous présenterons plus précisément au chapitre 4 la formalisation, que nous avons adoptée dans le cadre de notre travail, des connaissances qu'elles contiennent.

La mémoire conceptuelle

La mémoire conceptuelle rassemble les connaissances exprimant les propriétés générales des objets et des actions du monde de référence. Elle prend la forme d'un ensemble de concepts liés entre eux par des liens de différents types. Parmi ceux-ci, on relèvera les relations classiques d'hyperonymie et d'hyponymie ou celles exprimant l'usage qu'il est permis de faire d'un concept, c'est-à-dire les concepts auxquels il peut être lié et le type de relation assurant cette liaison. Ces cadres d'usage des concepts s'apparentent aux structures casuelles des grammaire de cas [Fillmore 1968]. Le tout constitue un réseau sémantique similaire sur les plans cités à ceux habituellement utilisés en Intelligence Artificielle et dérivant du modèle décrit dans [Collins & Quillian 1969].

La spécificité de la mémoire conceptuelle de MoHA réside en fait dans la donnée suivante : les concepts qui composent cette mémoire ne sont pas donnés a priori mais résultent de l'abstraction d'expériences. Il en découle que ces concepts ne sont pas définis simplement par les relations qu'ils entretiennent entre eux mais également par les relations qu'ils conservent avec les expériences qui leur ont donné naissance. Il s'agit là d'une forme d'ancrage des connaissances conceptuelles. Cet ancrage n'est pas réalisé pour le moment directement dans un niveau perceptif comme cela pourrait être le cas avec des expériences visuelles. Un concept ne renvoie en effet qu'à une configuration de mots et de situations au niveau de la dimension conceptuelle des expériences verbales. Cette première avancée dans la perspective de l'ancrage des concepts est néanmoins la source de capacités nouvelles. Les relations existant dans le réseau mot/situation offre ainsi la possibilité de rendre les relations entre concepts plus sensibles au contexte dans lequel elles sont envisagées. D'autre part, une liaison entre mots et concepts existe naturellement de par cet ancrage. Celle-ci sert déjà à la désambiguïsation lexicale et pourra donc être utilisée dans la construction de la représentation sémantique des propositions des textes.

Pour achever notre présentation de ce niveau conceptuel, précisons qu'au stade actuel de développement du modèle, les relations entre concepts ayant émergé ne sont pas encore différenciées. Elles sont dotées d'un poids mais ne possèdent pas encore d'étiquettes. Il est donc encore nécessaire de faire intervenir une expertise humaine concernant ce point.

La mémoire pragmatique

La mémoire pragmatique de MoHA regroupe les connaissances sur les situations prototypiques du monde de référence. Elle permet de caractériser le fait qu'un ensemble d'actions et d'états impliquant un certain nombre d'entités sont liés de façon cohérente de par leur appartenance commune à une même situation. Cette mémoire n'a cependant pas pour vocation de spécifier ce qui est toujours vrai dans une situation, ni de la décrire de façon exhaustive, mais cherche plutôt à rendre compte de ce que l'on y rencontre habituellement. On se trouve là dans le domaine couvert par les MOPs de Schank et le contenu de la mémoire pragmatique se veut assez proche d'un réseau de schémas similaires aux MOPs.

Les représentations des situations sont elles-mêmes structurées. Au niveau élémentaire, elles sont formées de concepts et s'appuient plus généralement sur les connaissances de la mémoire conceptuelle afin d'établir les relations devant unir ces concepts pour former la représentation des actions et des états impliqués dans les situations. Ces actions et ces états sont eux-mêmes différenciés suivant la dimension de la situation qu'ils décrivent. On distingue ainsi les états précisant quelles sont les circonstances dans lesquelles la situation prend place, les actions qui se produisent dans le cadre de la situation et qui en constituent le corps et enfin, les états exprimant les modifications apportées à l'état du monde du fait du déroulement de la situation.

Symétriquement à la mémoire conceptuelle, la mémoire pragmatique est le produit de l'abstraction de la dimension pragmatique des expériences verbales. Cette abstraction n'est pas un processus global. Elle intervient ponctuellement et localement. À partir d'un ensemble d'expériences dont les traits ont émergé avec suffisamment de stabilité, on forme un ou plusieurs schémas qui viennent s'insérer dans le réseau déjà existant des connaissances de la mémoire pragmatique. Cette caractéristique est similaire pour la mémoire conceptuelle. Même si l'abstraction est un processus discret à l'échelle d'une situation ou d'un concept, elle conserve un certain caractère de continuité, ou tout du moins de progressivité, à l'échelle de toute une mémoire.

2.3. L'utilisation des expériences pour la compréhension

2.3.1. Un cadre de référence : le raisonnement à base de cas

Nous avons montré au §2.1.2 que les notions de cas et d'expérience ont un ensemble de caractéristiques communes. Cela nous a notamment conduit à examiner au §2.2.1 quelles sont les similarités et les différences de l'apprentissage à partir d'expériences avec l'apprentissage caractéristique du raisonnement à base de cas. De par son origine, la notion d'expérience présente néanmoins une grande spécificité du point de vue de l'apprentissage. Par ailleurs, le raisonnement à base de cas, bien qu'intégrant l'apprentissage de façon assez naturelle, ne donne pas forcément à celui-ci une place centrale. Le raisonnement à base de cas ne pouvait donc être retenu comme cadre de référence unique pour l'apprentissage à partir d'expériences.

La situation est en revanche différente en ce qui concerne l'utilisation que l'on peut faire des expériences, en particulier pour les tâches de compréhension, puisque ce sont elles qui nous intéressent ici. Le raisonnement à base de cas est en effet le seul mode de raisonnement capable de travailler à partir de connaissances exprimées implicitement au travers du produit de l'activité d'un processus et non pas de connaissances formalisées a priori. Il est donc particulièrement adapté à un cadre où l'apprentissage ne privilégie pas une généralisation rapide visant à construire de nouvelles connaissances abstraites.

Afin de mieux cerner les principes présidant à l'utilisation des expériences dans l'accomplissement de tâches particulières, il nous a paru intéressant d'étudier de plus près les différentes composantes du raisonnement à base de cas et de les confronter avec les caractéristiques de la notion d'expérience.

De ce point de vue, les systèmes de raisonnement à base de cas ne forment pas un bloc homogène. Il est possible de les différencier suivant :

- le but dans lequel on fait appel au raisonnement à base de cas
- la façon dont les cas sont utilisés, ou encore le type des tâches dans lesquelles ils sont impliqués.

Le premier point se définit par un continuum entre deux positions extrêmes. Dans les systèmes ayant la capacité de résoudre les problèmes qui leur sont posés sans avoir recours au raisonnement à base de cas, celui-ci est considéré comme un moyen d'accélérer le processus de résolution en évitant de redévelopper à chaque fois tout un raisonnement coûteux. À l'inverse, lorsque les systèmes ne possèdent pas la connaissance leur permettant de produire une solution, le raisonnement à base de cas peut

être vu comme l'unique recours pour élaborer une solution, même si cette dernière est peu sûre du fait de l'impossibilité d'une validation par des connaissances existantes.

Avec l'utilisation des expériences pour l'analyse de textes, nous nous situons résolument dans le giron de la seconde position puisque nous nous imposons la contrainte d'une absence initiale de connaissances pragmatiques sur les domaines considérés. Notons toutefois qu'à mesure du développement de la représentation d'un domaine par le biais de l'apprentissage, on devrait passer progressivement de la seconde position à la première.

Dans [Kolodner 1993], Kolodner propose pour la seconde dimension de séparer les tâches de nature interprétative d'une part et les tâches de nature résolution de problèmes d'autre part. Dans les premières, on cherche à évaluer ou à interpréter un ensemble de faits en confrontant les uns aux autres des cas proches de la situation donnée, cette comparaison s'appuyant à la fois sur les similitudes de ces cas et sur leurs différences. C'est une approche que l'on peut également qualifier d'argumentative. Dans les tâches de type résolution de problèmes, les cas se présentent plutôt comme des solutions apportées à des problèmes proches. Le travail principal consiste alors à adapter ces cas au problème particulier qui est posé. Comme pour la première dimension, il s'agit là des deux positions extrêmes que les systèmes réalisés mêlent plus ou moins.

La tâche d'analyse de texte à laquelle nous nous intéressons se situe assez clairement comme une tâche de nature interprétative. Cette nature intrinsèque de la tâche est renforcée par le fait que les cas sont présents non seulement sous la forme d'expériences individuelles mais que celles-ci sont regroupées en agrégats d'expériences. Or ces agrégats sont le résultat d'une mesure de similarité appliquée entre les expériences. En utilisant ces agrégats, on s'appuie donc implicitement sur une forme de comparaison des expériences entre elles.

Pour en finir avec la typologie des systèmes de raisonnement à base de cas, on notera que les deux critères mis en évidence pour réaliser cette typologie ne sont pas sans lien entre elles. Pour réaliser l'opération d'adaptation dans les tâches de résolution de problèmes, il faut avoir des connaissances sur le domaine afin de savoir si les modifications que l'on opère donnent un résultat valide. Ces connaissances peuvent bien souvent être utilisées également pour réaliser la tâche considérée. L'emploi du raisonnement à base de cas est alors surtout motivé par un souci de plus grande efficacité. À l'inverse, raisonner en se fondant sur la comparaison de cas déjà observés permet dans une certaine mesure de faire l'économie de connaissances en se fondant d'abord sur la notion de similarité. L'approche interprétative est donc assez proche du raisonnement à base de cas vu comme recours ultime.

2.3.2. Influence de la notion d'expérience sur le cycle du raisonnement à base de cas

Le cycle du raisonnement à base de cas

La façon dont le raisonnement à base de cas se décompose a fait l'objet de nombreuses modélisations. On citera [Aamodt & Plaza 1994], [Bichindaritz 1994] ainsi que [Robba 1992] pour une décomposition du raisonnement par analogie, proche parent du raisonnement à base de cas. Globalement néanmoins, les différentes modélisations sont convergentes et se différencient principalement par une décomposition plus ou moins fine de telle ou telle étape. La figure 1.5 présente les grandes étapes que l'on retrouve généralement dans tous ces modèles.

À un premier niveau, le raisonnement à base de cas comporte trois phases. La phase de recherche consiste à retrouver au sein de la base de cas au moins un cas suffisamment proche du problème posé. La phase d'adaptation recouvre tout ce qui concerne l'utilisation des cas trouvés en mémoire afin de résoudre le problème considéré. Ces deux premières phases forment le cœur véritable du raisonnement à base de cas. La troisième et dernière phase est présente de façon plus ou moins développée en fonction des systèmes. Il s'agit de la phase d'apprentissage au cours de laquelle on cherche à faire évoluer la base de cas en fonction du nouveau cas résultant de la résolution du problème.

Rechercher un cas en mémoire conduit à se poser trois questions :

- que doit-on chercher? Compte tenu des données du problème à résoudre et du mode de structuration de la base de cas, il convient de déterminer les critères qui permettront de retrouver des cas intéressants.
- où doit-on chercher? Munis des critères de recherche, il faut être capable de cerner la partie de la base de cas où les cas compatibles avec ces critères peuvent être trouvés.
- comment savoir si ce que l'on a trouvé sera utile? Il est nécessaire de disposer, dans une certaine mesure, d'un moyen de savoir si ce que l'on a trouvé est pertinent pour la résolution du problème posé, ou tout du moins, de donner un ordre de préférence sur les cas sélectionnés.

Le schéma le plus classique concernant la phase de recherche est le suivant. On dispose d'une base de cas possédant un index de type hiérarchique, celui-ci pouvant être soit une hiérarchie de traits ou un réseau discriminant. La description du problème fait apparaître un certain nombre de traits qui sont alors utilisés afin d'accéder à l'ensemble

des cas indexés par cette conjonction de traits. On applique alors une mesure de similarité entre les cas trouvés et la représentation du problème pour éliminer les cas non pertinents et ordonner les autres. Bien entendu, il existe de multiples autres solutions possibles. En supposant par exemple une organisation de la base de cas “à plat”, on peut très bien être amené à parcourir (en parallèle ou non) toute la base et à appliquer une mesure de similarité entre la représentation du problème et chaque cas de la base. Les trois étapes mises en évidence se trouvent alors condensées en une seule.

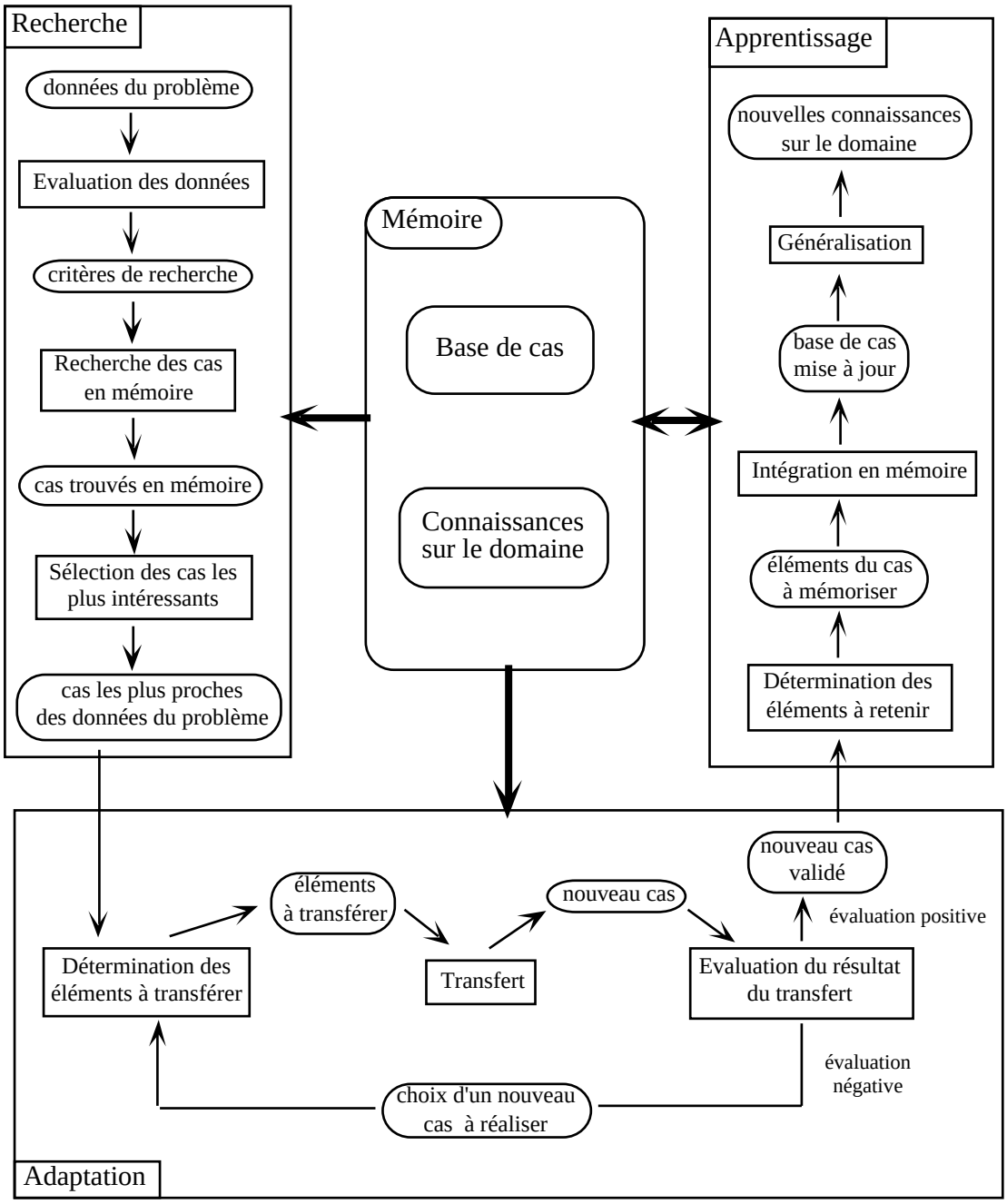


Fig. 1.5 - Cycle du raisonnement à base de cas

Comme dans le cas de la phase de recherche, la phase d'adaptation se décompose en une étape d'évaluation des données et de détermination du travail spécifique à réaliser, une étape de réalisation de ce travail et enfin, une étape d'évaluation des résultats de ce même travail. Ici, la tâche à mener consiste à transférer les éléments de solution présents au sein d'un cas vers le cas incomplet que constitue la représentation du problème posé. Bien entendu, ce transfert n'est généralement pas une simple copie mais s'accompagne d'une adaptation de ces éléments aux données du problème. Il convient donc tout d'abord de déterminer ce qui va être transférer, ensuite de réaliser le transfert et enfin, d'évaluer si le résultat obtenu est véritablement une solution correcte au problème abordé.

La phase d'apprentissage possède quant à elle un statut un peu particulier dans la mesure où elle n'intervient pas dans la résolution d'un problème lorsque celui-ci se pose mais influence plutôt la résolution des problèmes qui seront rencontrés dans le futur. Elle a pour objectif de déterminer ce que le nouveau cas obtenu à la suite de la phase d'adaptation peut apporter à la base de cas afin que celle-ci contribue à l'amélioration de l'ensemble de la chaîne.

La première étape de ce processus consiste à apprécier ce qui, dans ce nouveau cas, est susceptible d'engendrer un tel changement. Il est tout à fait possible que le cas soit suffisamment proche de certains déjà présents dans la base pour que sa mémorisation ne soit pas utile. On peut également n'en retenir que la partie spécifiquement nouvelle. L'étape suivante est la mémorisation proprement dite de ce que l'on a retenu du nouveau cas. Elle va au delà d'un simple stockage puisqu'elle recouvre également l'examen de l'impact de l'intégration de ce nouveau cas au sein de la base. Cela peut se traduire ainsi par la modification de la structuration de la base de cas au travers d'une réorganisation de ses index.

Comme le montre la figure 1.5, le raisonnement à base de cas n'exclut pas l'utilisation de connaissances sur le domaine. Celles-ci sont le plus souvent données a priori mais elles peuvent également être le produit d'une construction automatique. C'est l'objet de la dernière étape de la phase d'apprentissage. Elle a en effet pour objectif d'essayer de généraliser un ensemble de cas dans le but de produire de nouvelles connaissances abstraites sur le domaine. On tente habituellement de déterminer si une telle généralisation est possible entre le nouveau cas et ceux avec lesquels il se retrouve stocké, ceux-ci étant a priori les cas de la base qui lui sont le plus similaires.

Le cycle du raisonnement à base d'expériences

Le raisonnement à base d'expériences ne se différencie pas véritablement du raisonnement à base de cas quant aux grandes phases de son cycle. Les différences interviennent plutôt au niveau de la nature des étapes composant ces phases. Ces

différences sont motivées par une caractéristique générale de l'approche à base d'expériences : on ne fait pas l'hypothèse de l'existence a priori de connaissances abstraites sur le domaine. Cette contrainte implique que les étapes d'évaluation de toutes les phases se trouveront forcément limitées dans la mesure où elles s'appuient par nature assez fortement sur les connaissances du domaine.

Cette limitation est néanmoins partiellement compensée par une autre caractéristique des expériences. Celles-ci se présentent en effet sous la forme d'agrégats au sein desquels il est possible de retrouver chaque expérience mais qui sont également organisés de façon à ce que toutes leurs parties communes soient fusionnées. Ce mécanisme permet de pondérer les éléments composant un agrégat en fonction de leur degré de récurrence. On obtient ainsi une représentation à deux niveaux : on a d'une part le niveau des expériences individuelles, un peu comparable aux cas d'une "classique" base de cas; d'autre part, on a le niveau des agrégats, au sein duquel on perd l'information que tel élément est véritablement apparu en même temps que tel autre, mais qui contient en revanche une information sur l'importance relative des différents éléments qui le composent. Cette information, qui peut être vue comme une première forme de généralisation, offre également un moyen à la fois pour guider la sélection des éléments à considérer au début de chaque phase, que ce soit avant la recherche en mémoire ou avant le transfert par exemple, et pour évaluer les résultats obtenus en fin de phase.

Phase de recherche

Nous avons vu précédemment (cf. paragraphe traitant de l'apprentissage à partir d'expériences et du raisonnement à base de cas au sein de la partie 2.2.1) que la mémoire des expériences possède une structure "à plat". Cette caractéristique, conjuguée à l'existence d'une pondération des éléments constitutifs des agrégats, pèse fortement en faveur d'un processus de recherche travaillant sur la base d'une indexation de surface. Compte tenu de cette faiblesse des critères de recherche, il est nécessaire que les cas ainsi sélectionnés soient ensuite examinés plus attentivement. Une mesure de similarité doit donc être définie afin de vérifier que ces cas peuvent être effectivement utiles pour traiter le problème courant. Du fait de l'hypothèse de l'absence de connaissances a priori sur le domaine, cette mesure ne devra s'appuyer que sur des éléments déjà capitalisés au sein de la mémoire des expériences pour approfondir la similarité supposée par la première phase de recherche. En l'occurrence, les meilleurs indicateurs de fiabilité dont on dispose dans un tel contexte sont indiscutablement les pondérations dont sont affublés les constituants des agrégats. La similarité entre un cas de la mémoire et la représentation d'un nouveau problème peut ainsi non seulement être évaluée en fonction de la similarité de leurs traits mais également tenir compte de l'importance relative de ceux-ci.

Phase d'adaptation

Les trois phases du raisonnement à base de cas peuvent profiter de l'existence de connaissances définies a priori sur le domaine. Néanmoins, c'est certainement la phase d'adaptation qui en est la plus dépendante. Son niveau de sophistication est en fait assez directement en rapport avec le degré d'élaboration des connaissances disponibles sur le problème traité. Ces connaissances ne sont d'ailleurs pas tant utiles dans la réalisation même de l'adaptation que pour son contrôle. [Kolodner 1993] montre bien que de nombreuses techniques d'adaptation existent mais qu'elles ne peuvent être appliquées en aveugle, sans direction d'ensemble. Il est donc essentiel de déterminer quel est, ou quels sont, les outils d'adaptation à utiliser dans une situation particulière. C'est l'objet du contrôle, et pour que celui-ci puisse être efficace, il a besoin de savoir ce qui caractérise le domaine considéré, de faire la part entre ce qui est important et ce qui relève du détail.

Compte tenu de sa volonté de ne pas dépendre d'une connaissance a priori sur le domaine, il est assez évident que le raisonnement à base d'expériences se caractérise par une phase d'adaptation peu élaborée. Ce fait est renforcé par un autre trait du raisonnement à base d'expériences.

Dans la réutilisation des cas, Aamodt et Plaza [Aamodt & Plaza 1994] séparent l'opération de copie de celle d'adaptation proprement dite en précisant que les tâches de classement se contentent le plus souvent d'une forme élémentaire de réutilisation ne comprenant que la première opération. Or, dans l'approche à base d'expériences, la dimension raisonnement est largement dominée par la notion de classement. Il s'agit de rattacher les nouvelles expériences à celles que l'on a déjà rencontrées et les opérations effectuées sur ces nouvelles expériences, en améliorant la compréhension que l'on peut en avoir, visent à mettre en évidence les relations qu'elles entretiennent avec celles déjà présentes en mémoire. L'analyse thématique, exemple qui nous intéresse tout particulièrement ici, consiste ainsi essentiellement à reconnaître qu'un ensemble de propositions traitent du même sujet, chose qui est réalisée par comparaison avec des descriptions de situations déjà mémorisées. On peut donc concevoir que dans une grande partie des cas, la phase d'adaptation dans le raisonnement à base d'expériences s'assimile à une copie. Cette copie n'est pas pour autant aveugle et les informations contenues dans les agrégats sur l'importance relative des constituants du domaine sont une fois encore utilisables pour orienter le processus.

Phase d'apprentissage

De par la liaison assez étroite que le raisonnement à base de cas établit entre résolution de problèmes ou interprétation et apprentissage, nous avons déjà eu l'occasion

d'examiner dans le paragraphe 2.2.1 les différentes caractéristiques de sa phase d'apprentissage. Aussi nous nous contenterons ici d'en souligner deux aspects majeurs, en liaison avec les principes de l'approche à base d'expériences.

Tout d'abord, la phase d'apprentissage du raisonnement à base d'expériences se caractérise par une étape de généralisation en deux temps. Le premier temps est très étroitement fusionné avec l'étape d'intégration en mémoire. Au sein de la mémoire, nous avons vu que les expériences sont regroupées en agrégats au sein desquels des informations sur la récurrence de leurs traits, donc sur le degré de généralité de ceux-ci, sont déjà présentes. Ces pondérations représentent ainsi une première forme de généralisation, même si elle reste implicite dans la mesure où un choix des éléments à retenir n'est pas réalisé. Ce choix, ainsi que la construction des généralisations proprement dites, font l'objet du second temps. Alors que le premier temps de la généralisation intervient de façon continue et systématique, celui-ci opère ponctuellement et sur des éléments bien ciblés présentant des caractéristiques particulières. Il s'agit alors de faire émerger les connaissances stables contenues dans la mémoire des expériences et de les transformer afin de produire de nouvelles connaissances abstraites. Ce second temps marque donc une rupture, un changement de représentation profond qui s'approche d'une conception plus "classique" de la notion de généralisation.

Le second aspect à souligner concerne l'étape de détermination des éléments à retenir. L'une des spécificités de l'approche à base d'expériences est la volonté de faire émerger progressivement les éléments fondamentaux des expériences au fur et à mesure de l'accumulation de ces dernières et de leur confrontation les unes avec les autres. Ce processus doit intervenir en s'appuyant sur le plus petit ensemble possible de données a priori. L'étape de choix des éléments à retenir est donc par nature à l'encontre de la philosophie adoptée si elle vient en préalable de l'intégration en mémoire. En revanche, elle conserve une justification entière et remplit d'ailleurs un rôle essentiel lorsqu'elle intervient en préambule de l'étape de généralisation visant à construire de nouvelles connaissances abstraites, c'est-à-dire lors du second temps de la généralisation.

Récapitulatif

Dans une première partie, nous avons cherché à cerner le problème que nous posons ici, celui de l'apprentissage automatique de connaissances pragmatiques à partir de textes, et à fixer les grandes lignes de la solution que nous lui apportons. Nous avons ainsi précisé le type des connaissances visées, les connaissances sur les situations prototypiques du monde, et leur intérêt majeur pour la compréhension de textes. Nous avons également exposé en quoi leur grande variété et leur étendue conduisent

naturellement dans la voie de l'apprentissage automatique et en quoi les textes constituent pour le moment un support privilégié pour cet apprentissage.

Nous avons ensuite mis en évidence la difficulté, inhérente au problème abordé, que constitue la dépendance étroite entre apprentissage et compréhension. Afin de briser cet apparent cercle vicieux, nous proposons de mettre en œuvre un principe d'amorçage : les connaissances apprises doivent permettre d'améliorer la compréhension qui elle-même contribue à son tour au raffinement des connaissances disponibles. Ce principe premier est soutenu par trois autres : l'apprentissage doit être incrémental; il doit s'appuyer de façon privilégiée sur la notion de similarité et la compréhension doit être capable de réutiliser les connaissances produites par cet apprentissage. Nous montrons par la suite l'importance de la notion de mémoire, notamment du fait du caractère très progressif du processus d'apprentissage. Une référence dans ce domaine est la théorie de la mémoire dynamique de Schank que nous analysons rapidement afin de montrer pourquoi sa structuration a priori ne s'accorde pas avec nos hypothèses de base.

Dans la seconde partie de ce chapitre, nous avons essayé de donner une vue globale de l'approche retenue. Celle-ci est fondée sur la notion d'expérience que nous définissons comme le résultat, en termes de représentations internes, de l'interaction d'un système avec son environnement. À l'occasion de cette définition, nous montrons les proximités avec la notion de cas ainsi que la filiation avec les travaux de Vygotski sur la formation des concepts.

Nous examinons ensuite l'impact de la notion d'expérience sur l'apprentissage. Les liens avec les formes d'apprentissage reconnues les plus proches sont explorés. Nous nous situons ainsi par rapport au regroupement conceptuel et à l'apprentissageintervenant dans le cadre du raisonnement à base de cas. Nous présentons à la suite MoHA, un modèle d'apprentissage s'appuyant spécifiquement sur la notion d'expérience et qui constitue le cadre global dans lequel nous nous inscrivons pour réaliser l'apprentissage automatique de connaissances pragmatiques.

Enfin, nous analysons ce que la notion d'expérience induit au niveau de la compréhension. Nous montrons que la similarité entre la notion de cas et celle d'expérience conduit à situer le raisonnement à base d'expériences comme très voisin de certaines formes de raisonnement à base de cas, en l'occurrence lorsque celui-ci est employé dans des tâches interprétatives et comme recours en l'absence de connaissances sur le domaine. Nous étudions pour finir plus précisément quelles modifications du cycle du raisonnement à base de cas sont apportées par l'approche à base d'expériences. En particulier, nous établissons que le cycle reste globalement le même mais que la contrainte de limitation des connaissances a priori sur le domaine influence la plupart des étapes, en particulier les étapes d'évaluation.

Chapitre 2

Les systèmes apprenant des connaissances pragmatiques à partir de textes

Cette partie vise à présenter les travaux existants relatifs à l'apprentissage automatique de connaissances pragmatiques à partir de textes. Après un panorama général, nous détaillons plus spécifiquement quatre systèmes, caractéristiques chacun d'une approche particulière. Nous finissons par une discussion insistant plus particulièrement sur les limites de ces systèmes et mettant en relation celles-ci avec les grandes orientations de notre approche.

1. Vue d'ensemble

Les travaux dans le domaine de l'apprentissage automatique de connaissances pragmatiques à partir de textes ont globalement suivi les tendances plus générales de l'apprentissage symbolique. Cela s'explique par le fait que ces travaux sont issus très majoritairement de la communauté de l'apprentissage automatique, plutôt dans sa composante modélisation cognitive, que de la communauté du traitement automatique des langues.

Dans un premier temps, les approches proprement inductives (Similarity-Based Learning) ont prévalu. Un système tel qu'IPP [Lebowitz 1983] en atteste. Les textes, en l'occurrence des articles de journaux traitant du terrorisme international, y étaient surtout considérés comme des collections de faits et l'objectif poursuivi était de généraliser ceux-ci sans chercher à extraire la structure causale profonde qui sous-tendait leur présence.

De ce fait, les généralisations construites ne s'avéraient pas toujours pertinentes, la simple détection de coïncidences et l'utilisation de biais syntaxiques se révélant parfois insuffisantes face au problème de la discrimination des traits importants.

Le système GENESIS [Mooney & DeJong 1985] illustre le second paradigme important apparu dans la sphère de l'apprentissage symbolique : l'Explanation-Based Learning (EBL). L'application de cette approche au problème considéré ici conduit dans un premier temps à analyser un texte en ayant recours à des capacités explicatives avancées et à généraliser ensuite l'explication ainsi obtenue pour créer de nouvelles

connaissances pragmatiques. La généralisation de chaque nouvelle explication conserve la validité de la chaîne causale trouvée à l'aide de connaissances générales alors que les détails sont supprimés. Dans les faits, on peut considérer que les textes permettent essentiellement de spécialiser la connaissance que le système possède déjà. Ce dernier peut être ainsi plus efficace et plus sensible lorsqu'il rencontre des textes s'inscrivant dans le même contexte.

Mais cette méthode impose une contrainte majeure : construire une explication profonde demande en effet une grande quantité de connaissances sur le domaine considéré.

Les recherches qui ont suivi ont essayé de surmonter les faiblesses de ces deux techniques en les combinant. On a ainsi obtenu des systèmes d'apprentissage multi-stratégies caractérisés chacun par une façon particulière de faire coopérer ces deux techniques. Avec OCCAM [Pazzani 1988], Pazzani complétait une théorie du domaine en faisant appel à l'induction, de façon à pouvoir mettre en œuvre un apprentissage orienté EBL. Dans une optique légèrement différente, Danyluk [Danyluk 1987] guidait un processus inductif en utilisant les explications fournies par un processus de type EBL.

Cependant, ces deux systèmes présentent une limitation commune : la connaissance provenant des textes ne peut pas être utilisée pour améliorer les capacités du module de compréhension tant qu'une étape de généralisation n'est pas intervenue. Face à un nombre d'exemples trop faible dans un domaine où il n'existe pas encore de connaissances, on a alors le choix entre ne pas produire d'explication du tout, produire une explication très générale, c'est-à-dire peu informative, ou bien choisir de généraliser à partir du peu d'exemples disponibles et risquer de s'exposer aux mêmes problèmes que les approches purement inductives.

Le raisonnement à partir de cas (CBR) donne en revanche la possibilité de travailler directement à partir des exemples quand les abstractions font défaut. Les systèmes SWALE [Schank & Leake 1989] et AQUA [Ram 1993] ont exploré cette voie. Ils cherchent à appliquer des schémas d'explication déjà connus en les adaptant de façon plus ou moins importante à la nouvelle situation. Si ce processus réussit, ils généralisent les deux structures explicatives en présence.

Toutefois, les schémas d'explication que ces systèmes conservent en mémoire pour servir de support à la compréhension des textes ne peuvent pas être véritablement vus comme des cas étant donné qu'ils correspondent à des représentations de textes généralisées. Il s'agit là plutôt d'une sorte de mariage entre CBR et EBL.

Dans ce qui suit, nous examinons plus en détail pour chacune des quatre tendances esquissées ci-dessus un système particulièrement représentatif.

2. IPP

Présentation

Le système IPP (Integrated Partial Parser) est l'un des tout premiers systèmes d'apprentissage automatique de connaissances pragmatiques à partir de textes. C'est également, avec CYRUS [Kolodner 1983] [Kolodner 1983], l'un des premiers systèmes à avoir implémenté les idées exprimées par Schank dans la théorie de la mémoire dynamique. IPP est donc caractérisé par l'importance qu'il accorde à la notion de mémoire ainsi qu'à la force du lien entre compréhension et apprentissage.

IPP travaille à partir de très courts textes appartenant à un domaine ciblé, en l'occurrence des dépêches d'agence de presse se résumant le plus souvent à une phrase et concernant des actes de terrorisme. Il en construit une représentation sous forme de schéma sur la base de ceux qu'il possède déjà en mémoire. La représentation d'un texte ainsi construite est ensuite mémorisée en relation avec les schémas ayant servi à la construire. Si le nouveau schéma est suffisamment proche d'un de ceux-ci, les deux schémas sont généralisés, ce qui conduit à l'élaboration d'un nouveau schéma, factorisant les traits communs aux deux schémas originels. Ceux-ci restent néanmoins présents en mémoire mais sont simplement stockés de façon différente.

La représentation des connaissances

La mémoire d'IPP se compose d'un ensemble de hiérarchies de schémas. Chacune de ces hiérarchies possède comme racine un **S-MOP** ("Simple-Memory Organization Packet"). Un S-MOP est la représentation d'une situation générale du domaine considéré, comme une extorsion ou une attaque par exemple dans le cas traité par Lebowitz. L'ensemble des S-MOPs constituent la connaissance fournie a priori au système afin que celui-ci dispose d'un niveau minimal de compréhension des textes. Les S-MOPs forment également le cadre au sein duquel la généralisation est réalisée. Cette connaissance est spécifique de chaque domaine abordé. Pour celui des actes de terrorisme, Lebowitz s'appuie sur trois S-MOPs : SEXTORT, SATTACK-PERSON, SDESTRUCTIVE-ATTACK.

Les schémas composant le corps des hiérarchies sont des **spec-MOPs** ("specialized-Memory Organization Packet"). Sur le plan de leur structure et de leur contenu, ils sont identiques aux S-MOPs mais à la différences de ces derniers, ils sont construits automatiquement par IPP à partir des textes qui lui sont soumis. Lorsque ce sont des nœuds terminaux d'une hiérarchie, il s'agit directement de représentations de textes. Les

nœuds intermédiaires entre la racine et les feuilles correspondent quant à eux à des généralisations construites à partir des représentations de texte.

Chaque schéma est constitué d'un ensemble d'attributs. Chacun d'entre eux peut avoir plusieurs facettes, chacune dotée d'une valeur. Ces valeurs sont soit des propriétés élémentaires des entités du domaine, soit des "Actions Units" (**AUs**), c'est-à-dire des structures décrivant des événements concrets. On retrouve la structure des MOPs de Schank. Les S-MOPs sont équivalents aux MOPs tandis que les AUs sont assimilables aux scènes. La figure 2.1 donne un exemple de schéma construit comme représentation d'un texte.

Three gunmen kidnapped a 67 year-old retired industrialist yesterday outside his house near this north Italian town, police said.

EV13 (S-EXTORT)

Hostages	Age	old
	Occupation-Type	retired
	Role	businessman
	Status	estab
	Gender	male
Actor	Number	3
Methods	Act-Unit	$kidnap
Area	Location	western-europe
Nation	Location	Italy

Fig. 2.1 - Un texte et le schéma le représentant (d'après [Lebowitz 1983])

La hiérarchie de schémas prend la forme d'un réseau discriminant. Pour un schéma d'un niveau donné, on ne fait ainsi figurer que les attributs, les facettes et les valeurs qui sont différents de ceux figurant dans le schéma de niveau supérieur dont il est une spécialisation.

La dimension compréhension

À partir du texte, la première tâche du processus de compréhension consiste à identifier le S-MOP qui en est le plus proche. Celui-ci fournit en effet le cadre, autrement dit l'ensemble des traits, qui doit permettre d'analyser le texte. L'analyse proprement dite consiste à identifier la valeur des facettes des attributs formant le schéma retenu comme filtre d'analyse. Celui-ci peut être un spec-MOP dans le cas où une spécialisation du S-MOP initialement choisi a été jugée suffisamment similaire à la situation dont le texte

rend compte. Le choix d'un tel spec-MOP est réalisé en parcourant le réseau discriminant sur lequel s'appuie la hiérarchie des spécialisations du S-MOP initial. Que le schéma retenu pour servir de cadre d'analyse soit un S-MOP ou un spec-MOP, la représentation du texte est construite comme une spécialisation de celui-ci.

Cette analyse est d'abord de nature explicative car elle met en évidence le rôle des éléments constituant le texte par la détermination de l'attribut et de la facette dont ils dépendent. Mais elle est également de nature prédictive dans la mesure où certains traits présents au niveau du schéma d'analyse ne sont pas explicités dans le texte. Ils sont alors ajoutés à la représentation du texte, sachant que la justification de leur présence ou son infirmation doit venir sur le plus long terme des mécanismes d'apprentissage.

La dimension apprentissage

Du point de vue de l'apprentissage, IPP se situe résolument dans la sphère des systèmes de type SBL. Son originalité, dans ce cadre, réside dans la place qu'il accorde à la notion de mémoire. Il en résulte un mécanisme d'apprentissage spécifique reposant sur les trois principes suivants :

- la généralisation se fonde uniquement sur la détection de traits communs;
- la généralisation est rapide, c'est-à-dire qu'elle se fait à partir de peu d'exemples;
- les généralisations obtenues ne sont pas figées mais continuent à évoluer en mémoire en fonction de l'utilisation qui en est faite.

Le fonctionnement de base de l'apprentissage est le suivant. À l'issue du processus de compréhension, on dispose d'un schéma représentant le texte. Celui-ci est une spécialisation d'un S-MOP ou d'un spec-MOP. On examine alors s'il n'existe pas une autre spécialisation de ce x-MOP partageant suffisamment de traits communs avec la nouvelle représentation de texte. Dans l'affirmative, un nouveau spec-MOP rassemblant tous ces traits communs est créé. Les deux spécialisations précédentes sont directement rattachées à cette généralisation.

Ce descriptif couvre les deux premiers points mis en évidence ci-dessus. Le troisième s'appuie quant à lui sur deux notions : la prédictivité et l'assurance. La première est attachée aux traits composant les schémas. Elle caractérise la spécificité d'un trait vis-à-vis d'un schéma. Par là même, elle rend compte également de la pertinence à appliquer un certain schéma lorsque l'on détecte une configuration donnée de traits dans un texte. En pratique, la prédictivité d'un trait est évaluée par le nombre de fois où ce trait apparaît dans les spec-MOPs spécialisant un S-MOP. Cette notion est utilisée directement pour faire évoluer les réseaux discriminants permettant d'accéder aux spec-MOPs lors de la

compréhension. Les spécialisations d'un x-MOP ne sont en effet pas indexés par rapport à lui par tous les traits qui l'en distinguent mais seulement par ceux qui possèdent un degré de prédictivité suffisant.

L'assurance, quant à elle, est directement attachée aux spec-MOPs. Elle tente de cerner le degré de validité que l'on doit accorder à une généralisation créée. L'assurance d'un spec-MOP évolue en fonction de l'utilisation qui en est faite pour la compréhension. Lorsque l'on essaie de spécialiser un spec-MOP pour construire une représentation de texte, l'assurance de ce spec-MOP est augmentée chaque fois qu'un de ses traits considérés comme prédictifs peut être appliqué et elle est diminuée chaque fois qu'un de ces mêmes traits est contredit. Sur le plus long terme, si l'assurance d'un spec-MOP devient suffisamment grande, celui-ci devient permanent et prend le même statut qu'un S-MOP tandis que si elle tombe en dessous d'un certain seuil, le spec-MOP disparaît de la mémoire.

Discussion

En dépit de son ancienneté, IPP est un système particulièrement intéressant en comparaison de notre approche. Dans le domaine qui nous occupe, il est en effet l'un des rares systèmes à mettre véritablement en avant un apprentissage inductif minimisant l'importance des connaissances a priori sur le domaine et intégrant la prise en compte d'une évolution sur le long terme des structures généralisées dans le cadre d'une mémoire. Il généralise de fait des représentations de texte en se fondant uniquement sur la détection de similarités de surface et il est capable à la fois de faire émerger les traits les plus caractéristiques des généralisations au travers de la notion de prédictivité et de faire évoluer plus globalement le degré de confiance qu'il accorde à ces mêmes généralisations en fonction de leur utilisation réelle.

Ces points d'accord vis-à-vis de certains des principes de notre approche s'accompagnent également de quelques points de divergence et de quelques critiques. Tout d'abord, on peut s'interroger sur la capacité réelle d'IPP à traiter des textes dans la mesure où la très grande majorité des exemples donnés se résument à une seule phrase. En dehors de tout problème de définition de la notion de texte, cela signifie qu'IPP n'est capable d'appréhender que des textes thématiquement homogènes, n'abordant chacun qu'une seule situation.

Par ailleurs, le mécanisme de compréhension est visiblement très dépendant des connaissances décrivant le domaine, autrement dit des S-MOPs fournis a priori. La compréhension consiste essentiellement à trouver dans les textes les valeurs des traits des S-MOPs ou de leurs spécialisations. Il apparaît donc évident qu'en l'absence de ces connaissances, la construction des représentations de texte ne peut avoir lieu. Il est

également évident qu'en leur fournissant leur cadre de description, elles contraignent fortement la forme et le contenu de ces représentations. En dépit de sa dominante inductive, IPP effectue donc plus une spécialisation de la description d'un domaine qu'il ne contribue faire émerger celle-ci.

Du point de vue de l'apprentissage, l'ordre dans lequel les opérations sont réalisées nous semble également un point de discussion à soulever. IPP construit d'abord des généralisations très rapidement (à partir de deux exemples seulement), généralisations dont il confirme ou infirme ensuite la validité de manière plus avancée en fonction de l'utilisation qui en est faite lors de la compréhension d'autres textes. Le processus d'apprentissage est ainsi capable de produire des connaissances directement utilisables par le processus de compréhension. Cette façon de faire rend compte en fait du couplage étroit existant entre compréhension et apprentissage au sein de la théorie de la mémoire dynamique. Néanmoins, elle se heurte au problème de la remise en cause des structures construites lorsque celles-ci ont elles-mêmes servies de support à la construction d'autres structures. Dans le cas d'IPP, si un spec-MOP se révèle non pertinent et qu'il possède des spécialisations, que fait-on de ces dernières si on fait disparaître le spec-MOP en question? Cela revient en quelque sorte à construire tout un édifice et à s'apercevoir, alors qu'il est déjà fort avancé, qu'un certain nombre de ses constituants sont défectueux.

Ainsi que nous l'avons esquissé au chapitre 1, notre approche est résolument inverse. Nous cherchons à constituer progressivement les connaissances que nous voulons apprendre et c'est seulement lorsque celles-ci sont stables que nous procédons à une généralisation. Cela suppose bien entendu que le processus de compréhension soit capable de faire usage de connaissances dans un état encore relativement instable.

3. GENESIS

Présentation

Le système GENESIS est la réalisation concrète des idées avancées par DeJong au début des années 80 [DeJong 1981] [DeJong 1982] [DeJong 1983] sur la possibilité d'apprendre de nouvelles connaissances à partir d'un faible nombre d'exemples à condition de disposer de capacités explicatives, nécessairement sous-tendues par des connaissances importantes sur le domaine considéré. Ce courant de recherche s'est ensuite structuré [Mitchell et alii 1986] [DeJong & Mooney 1986] pour donner naissance à l'Explanation-Based Learning.

Le principe général de GENESIS consiste donc à produire d'abord une explication assez profonde du texte[1] qui lui est soumis pour essayer ensuite de la généraliser afin de produire un nouveau schéma, caractérisant comme dans notre cas une situation et prenant la forme d'une configuration d'actions et d'états. La généralisation est opérée en conservant les relations causales mises en évidence par le processus de compréhension, ce contrôle étant sous la dépendance des connaissances du domaine.

La représentation des connaissances

L'ensemble des connaissances sémantiques et pragmatiques de GENESIS sont représentées par des schémas, schéma désignant ici une structure rassemblant un ensemble d'attributs pouvant avoir chacun une ou plusieurs valeurs. Ces schémas sont de trois grands types. On distingue les schémas de type *Action*, les schémas de type *État* et ceux de type *Objet*. Tous les schémas sont inclus dans une hiérarchie au sein de laquelle prévaut un principe d'héritage. Les schémas Objet représentent des entités du monde de référence et rassemblent l'ensemble de leurs propriétés intrinsèques (par exemple leur taille, leur poids, leur forme si ce sont des objets physiques). Leurs attributs varient donc en fonction de ces propriétés.

Les schémas Action ou État comportent en revanche un ensemble fixe d'attributs. Pour les premiers, on distingue ainsi[2] :

- les *rôles* et leurs valeurs par défaut, c'est-à-dire les types d'entités impliqués dans l'action;
- les *conditions* sous lesquelles l'action peut intervenir. On fait la part entre les *pré-conditions*, qui sont les états devant être vrais pour que l'action se produise et les *motivations*, qui sont les états (croyances ou buts) qui peuvent motiver un acteur pour réaliser l'action;
- la *description* de l'action en elle-même. Elle est constituée d'un ensemble d'états et d'actions de plus bas niveau.
- les *résultats* liés à la réalisation de l'action. On différencie ici les *effets*, l'ensemble des états qui vrais à la fin de l'action, et les *terminaisons d'états*, autrement dit l'ensemble des états qui ne sont plus vrais du fait de la réalisation de l'action;
- les schémas intervenant dans la gestion du processus de compréhension fondé sur l'activation et l'instanciation de schémas. On référence à ce niveau les *schémas*

[1] Les textes traités par GENESIS sont de style narratif, assez proches dans l'esprit de ceux dont s'occupe IPP. Ils sont cependant plus longs, bien que n'abordant eux aussi qu'une seule situation.

[2] Les valeurs de chacun des attributs sont des schémas.

suggérés, en l'occurrence ceux qui possèdent l'action dans leur attribut description, et les *schémas déterminants*, qui, s'ils sont tous présents, sont le signe de la survenue de l'action.

En ce qui concerne les schémas État, les attributs possibles sont :

- les *rôles* et leurs valeurs par défaut;
- les *inférences*. Ce sont les états qui peuvent être directement déduits de celui considéré. Par exemple, si quelqu'un possède 1 million de francs, on peut en déduire qu'il possède 100000 francs;
- les *antécédents possibles*. Il s'agit des actions ayant cet état dans leurs effets.

Il est à noter que les schémas appris par GENESIS ne sont que du type Action.

La dimension compréhension

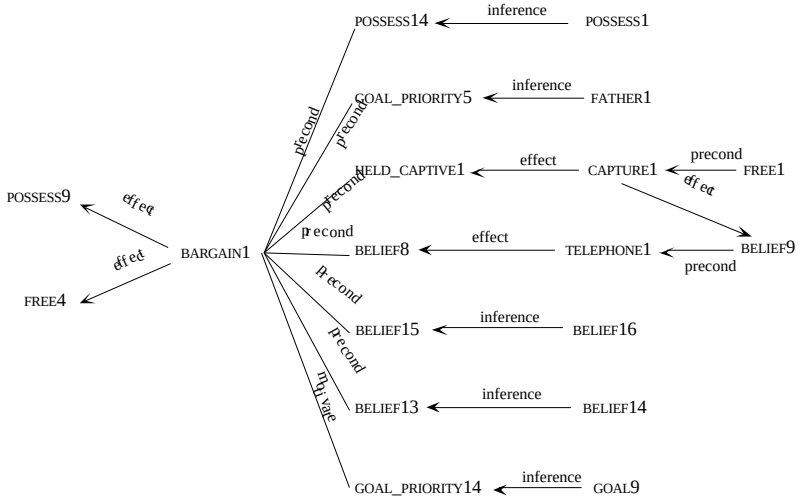
L'objectif de la compréhension dans GENESIS est la construction d'une chaîne causale unissant les actions et les états, soit directement explicités dans le texte considéré, soit que l'on peut déduire de ceux-ci à partir des connaissances dont on dispose sur le domaine. La figure 2.2 donne un exemple du type de texte traité et de la chaîne causale que GENESIS établit comme résultat de sa compréhension. Ne sont représentés sur cette figure que les actions et les états de plus haut niveau. Les relations de ces chaînes causales sont de quatre types : *pré-condition*, *effet*, *motivation* et *inférence*. Ces types de relations reprennent une partie de ceux existant entre schémas Action et/ou État du fait de leurs attributs.

Le processus de compréhension de GENESIS agit selon deux modes. S'il trouve dans sa mémoire au moins un schéma reprenant une partie ou la totalité de l'enchaînement d'actions et d'états présents dans le texte, il adopte une stratégie principalement dirigée par les attentes engendrées par ce ou ces schémas telle qu'elle est appliquée dans SAM [Cullingford 1978] par exemple. Dans le cas contraire, il essaie de combiner les actions et les états qu'il peut trouver parmi ceux dont il dispose à la manière d'un planifieur du type de PAM [Wilensky 1983].

La première stratégie s'appuie sur les attributs *schémas suggérés* et *schémas déterminants* des schémas Action. Les textes sont traités proposition par proposition. Pour chaque nouvelle proposition, on active les schémas suggérés du schéma Action associé à l'action explicitée dans le texte, si la proposition en comporte une. Dans le même temps, on examine si l'action ou l'état de la proposition correspond à l'un des schémas déterminants d'un schéma précédemment activés lors du traitement de propositions antérieures. Lorsque tous les schémas déterminants d'un schéma actif ont été

trouvés dans le texte, ce schéma est confirmé de façon définitive. On peut alors ajouter à la chaîne causale en cours de construction tous les états et les actions faisant partie de sa description.

Fred is the father of Mary and is a millionnaire. John approached Mary. She was wearing blue jeans. John pointed a gun at her and told her he wanted her to get into his car. He drove her to his hotel and locked her in his room. John called Fred and told him John was holding Mary captive. John told Fred if Fred gave him $250 000 at Trenos then John would release Mary. Fred gave him the money and John released Mary.



POSSESS9	John has $250,000.	BELIEF8	Fred believes John is holding Mary captive.
FREE4	Mary is free.	TELEPHONE1	John calls Fred and tells him that he is holding Mary captive.
BARGAIN1	John makes a bargain with Fred in which John releases Mary and Fred gives $250,000 to John.	BELIEF9	John believes he is holding Mary captive.
POSSESS14	Fred has $250,000.	BELIEF15	John believes Fred has $250,000.
POSSESS1	Fred has millions of dollars.	BELIEF16	John believes Fred has millions of dollars.
GOAL_PRIORITY5	Fred wants Mary free more than he wants to have $250,000.	BELIEF13	John believes Fred wants Mary to be free more than he wants to have $250,000.
FATHER1	Fred is Mary's father.	BELIEF14	John believes Fred is Mary's father.
HELD_CAPTIVE1	John is holding Mary captive.	GOAL_PRIORITY4	John wants to have $250,000 more than he wants to hold Mary captive.
CAPTURE1	John captures Mary.	GOAL9	John wants to have $250,000.
FREE1	Mary is free.		

Fig. 2.2 - Un texte et la chaîne causale que GENESIS en extrait (d'après [Mooney & DeJong 1985])

La première stratégie de compréhension essaie donc de trouver des schémas de plus haut niveau qui puissent expliciter les relations causales entre les actions et les états dans le texte. La seconde essaie d'établir ces relations en exploitant différents mécanismes inférentiels tout en restant au même niveau de connaissances. On commence par ajouter à la représentation du texte en cours de construction, appelée modèle, les inférences immédiates possibles à partir de chacun des schémas directement issus du texte : les effets d'une action ou les inférences répertoriées au niveau d'un état. Ce fonctionnement en chaînage avant est complété par un mécanisme en chaînage arrière. Celui-ci entre en action lorsque l'on souhaite par exemple retrouver au niveau du modèle les pré-conditions d'une action déjà avérée. La référence existant dans les schémas État aux actions pouvant les produire (les antécédents possibles) est également exploitée dans ce cadre. Enfin, on peut examiner si des buts très généraux, appelés également buts thématiques (tels que se

nourrir, devenir riche, etc.), peuvent expliquer la présence de certaines actions au niveau du modèle.

La dimension apprentissage

L'apprentissage dans GENESIS prend globalement la forme d'une généralisation guidée par les connaissances du domaine. Il comprend trois étapes : la détermination de l'opportunité de généraliser, la généralisation proprement dite de la chaîne causale construite par le processus de compréhension et le découpage de la chaîne généralisée afin de produire un nouveau schéma de type Action.

Afin d'éviter une explosion du nombre de schémas, GENESIS applique des critères plus ou moins liés au type de texte considéré[1] pour savoir si une représentation de texte doit mener à la création d'un nouveau schéma. La chaîne causale doit d'abord conduire à l'accomplissement d'un but d'un des acteurs. Ce but doit ensuite être suffisamment général pour être rencontré à nouveau ultérieurement. Autrement dit ce doit être un but thématique. Enfin, le schéma formé ne doit pas déjà exister au sein la mémoire.

La généralisation proprement dite s'effectue selon la procédure suivante. Les schémas constituant la chaîne causale du texte sont d'abord remplacés par le sommet de la hiérarchie des schémas, ce qui correspond à une généralisation maximale. On cherche ensuite à spécialiser ces éléments de façon à ce que les relations causales qui les unissent soient respectées. Le point de départ de cette spécialisation est le schéma exprimant le but thématique qu'explique la chaîne causale. Dans l'exemple de la figure 2.2 par exemple, on part ainsi de l'état POSSESS9 et l'on cherche à spécialiser son prédécesseur dans la chaîne de façon à ce que le schéma obtenu ait comme effet cet état. En remontant ainsi la chaîne causale de proche en proche, on spécialise progressivement tous les constituants de celle-ci.

La dernière étape est menée en se fondant sur un ensemble de règles simples. Les pré-conditions du nouveau schéma sont ainsi formées des états de la chaîne causale qui n'ont pas de prédécesseur et qui ne sont pas liés à leur successeur par un lien de type motivation. Au contraire, ses effets sont les états de cette même chaîne qui n'ont pas de successeur.

[1] Dans les textes narratifs, on s'intéresse essentiellement au devenir des différents personnages et aux raisons qui les font agir.

Discussion

Ainsi que leurs auteurs le mettent en avant, GENESIS n'est pas un système capable d'apprendre sans qu'il n'existe auparavant une théorie du domaine déjà constituée. Son apprentissage consiste soit à spécialiser des schémas existants, soit à figer des chaînes de raisonnement que le système sait élaborer pour en faire de nouveaux schémas. L'apprentissage réalisé permet donc de gagner en efficacité, particulièrement lorsqu'il évite un raisonnement de type planification, et en sensibilité, dans le cas où une spécialisation d'un schéma peut être déclenchée alors que celui-ci n'aurait pu l'être, parce que trop général. Si l'on se place dans un contexte de modélisation cognitive, on peut rapprocher cet apprentissage de ce qui se passe chez un adulte lisant des textes lui apportant des informations sur des situations spécifiques nouvelles mais relevant de sujets qui ne lui sont pas étrangers.

Ainsi que nous l'avons montré au chapitre 1, notre perspective est assez différente et se rapproche davantage de celle de l'enfant abordant un nouveau domaine. Une de nos motivations pour aborder le problème selon cet angle réside dans la difficulté qu'il y a à cerner précisément quelles peuvent être l'étendue et la forme des connaissances à un stade avancé du développement cognitif. Un certain nombre d'hypothèses (cf. travaux sur les réseaux sémantiques [Collins & Quillian 1969], [Rosch 1977] et les schémas [Bartlett 1932], [Minsky 1975], [Bobrow & Norman 1975] par exemple) prévalent quant à leur forme mais la façon dont un vaste ensemble de connaissances pourraient être représentées en se reposant sur ces hypothèses semble encore inaccessible en supposant que l'on souhaite se rapprocher des caractéristiques humaines. Dès lors, il nous semble moins spéculatif de centrer notre attention sur les mécanismes qui conduisent à la formation de ce niveau de connaissances.

En dehors de la différence d'approche qui nous sépare de ces travaux, il nous semble que l'absence de progressivité de l'apprentissage opéré par GENESIS en limite la portée. Dans [Mooney & DeJong 1985], l'exemple donné illustre comment le système est capable, après le traitement d'un texte et l'apprentissage qui en résulte, de comprendre un autre texte à propos du même sujet et pour lequel, avant considération du premier texte, il lui était impossible de construire la chaîne causale reliant ses différents événements. Néanmoins, on constate que le texte donné pour réaliser l'apprentissage détaille la situation en question suffisamment fortement pour que la généralisation puisse se faire à partir d'un seul texte. Or, il apparaît que de tels textes relèvent davantage de l'exception que de la règle. Le plus souvent, un texte ne donne des précisions que sur un aspect d'une situation et ce n'est que par le recoupement de plusieurs textes que l'on peut espérer avoir une image à peu près complète de l'intégralité de la situation. Il faut donc qu'un système d'apprentissage à partir de textes soit au moins à même de différer son processus

de généralisation jusqu'à ce qu'il possède suffisamment d'éléments pour que ce processus puisse agir. Cette aptitude doit en outre s'accompagner de la possibilité de détecter la similitude des situations abordées par les textes.

4. OCCAM

Présentation

Parmi les systèmes étudiés ici, OCCAM se distingue par la façon dont il réalise son apprentissage, et cela sur deux points. Tout d'abord, il met en œuvre une forme d'apprentissage originale, à mi-chemin entre EBL et SBL, appelée *Theory-Driven Learning* (TDL). Cette technique peut être vue globalement comme une forme d'induction guidée par des connaissances trop générales pour assurer la validation définitive des connaissances apprises mais permettant tout de même de jouer le rôle d'un biais efficace.

Ensuite, il tente de dépasser les problèmes posés par chacune des grandes stratégies d'apprentissage existantes, en l'occurrence l'EBL et le SBL, auxquels il ajoute le TDL, en les combinant dans un système multi-stratégies dans l'optique de compenser leurs insuffisances par leurs forces respectives.

Dans le cas d'OCCAM, la façon dont les différentes stratégies sont agencées est fixe. On privilégie toujours les formes d'apprentissage reposant sur des connaissances existantes et parmi celles-ci, on donne la préférence aux connaissances les plus spécifiques. Cette attitude se justifie par le fait que la validité des connaissances apprises est d'autant plus élevée qu'elle est établie par des connaissances, mêmes générales, plutôt que par l'observation de simple cooccurrences. Dans cette optique, les méthodes de type SBL sont utilisées en dernier recours avec comme objectif de produire les connaissances qui sont nécessaires au fonctionnement des méthodes du type EBL.

La représentation des connaissances

Dans OCCAM, les connaissances pragmatiques que l'on utilise et que l'on apprend apparaissent sous trois formes

- des relations causales élémentaires. Elles permettent d'exprimer les relations directes existant entre les actions et les états comme le fait par exemple que le résultat d'une action est un certain état;

- des patrons de causalité ("causal patterns"). Ces patrons explicitent dans quelles conditions il est possible de reconnaître des relations causales dans un enchaînement d'actions et d'états. Ces connaissances servent de support au TDL;

- des schémas. Ceux-ci sont comparables aux schémas utilisés au sein de GENESIS. Comme eux, ils permettent de figer des configurations caractéristiques de relations causales.

Ces trois types de connaissances pragmatiques s'appuient tous sur une représentation conceptuelle des actions, des états et des objets sous forme de dépendances conceptuelles [Schank 1972].

Les relations causales

Les relations causales sont les briques élémentaires dont sont constitués les deux autres types de connaissances. Elles sont également à la base des représentations de texte. L'objectif du processus de compréhension est en effet de mettre en évidence les relations causales qui unissent les actions et les états présents dans les textes. La figure 2.3 donne un exemple d'une telle relation. On y exprime que le fait de toucher un ballon de baudruche avec un objet pointu provoque l'éclatement de ce même ballon.

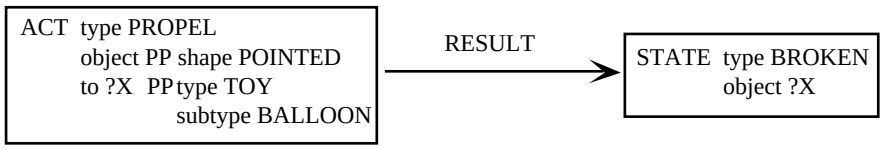


Fig. 2.3 - Une relation causale élémentaire (d'après [Pazzani 1991])

Au niveau sémantique, chaque action (ACT), état (STATE), objet (PP) ou relation (RELATION) est représenté par une dépendance conceptuelle. Celle-ci possède une tête donnant son type général et un ensemble, éventuellement vide, de couples rôle/valeur exprimant chacun une contrainte. La valeur d'un rôle (par exemple ici, type, shape et object sont des rôles) est elle-même une dépendance conceptuelle. Les valeurs terminales sont des dépendances sans rôle (comme PROPEL ou TOY). Des variables (?X ici) permettent de spécifier des contraintes d'égalité entre valeurs.

L'expression des relations causales s'effectue au travers de relations entre dépendances conceptuelles. Pour l'expression de la causalité physique, on en distingue deux. La relation RESULT (cf. figure 2.3) traduit le fait qu'un état est le résultat d'une action tandis que la relation ENABLE caractérise le fait qu'un état permet le déclenchement d'une action.

Les patrons de causalité

Les patrons de causalité expriment une forme d'inférence de nature inductive que l'on peut grossièrement résumer de la façon suivante : si un enchaînement donné d'actions et d'états intervient, certaines conditions étant posées sur ceux-ci, sur leurs relations ainsi

que sur leurs composants, alors cet enchaînement est explicable par une causalité sous-jacente qui est explicitée par le patron qu'il a permis de déclencher. Le patron de la figure 2.4 établit ainsi que si une action se produit et se trouve suivie d'un changement d'état ayant pour objet l'entité destination de cette action (antécédent du patron) alors ce changement d'état est le résultat de l'action en question (conséquent du patron). L'application de ces patrons via le mécanisme de TDL permet de produire de nouvelles règles causales élémentaires.

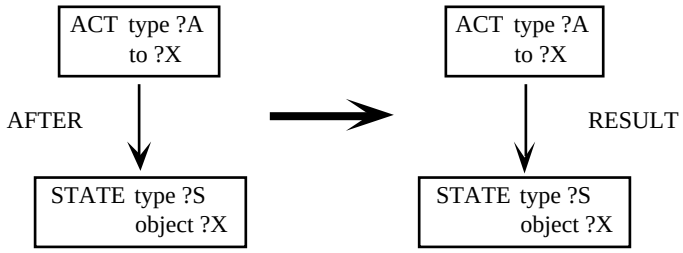


Fig. 2.4 - Un patron de causalité (d'après [Pazzani 1991])

OCCAM dispose de 30 de ces patrons, répartis en trois grands types :

- les "exceptionless causal patterns". Ce sont des patrons semblables à celui de la figure 2.4. Ils sont applicables lorsque l'on constate qu'une même action produit toujours le même résultat et qu'il existe des régularités au niveau des rôles composant cette action et cet état.
  Exemple : on observe un ensemble de situations dans lesquelles à chaque fois, le fait de toucher un ballon avec un objet pointu est suivi de l'éclatement du ballon. En appliquant un patron de ce type, on pourra alors construire la règle de la figure 2.3;

- les "dispositional causal patterns". Ces patrons s'intéressent au contraire aux cas dans lesquels une action ne produit pas toujours le même résultat mais que ces différences se justifient par une différence de valeur intervenant au niveau des rôles composant l'action et l'état en question.
  Exemple : on dispose de plusieurs observations. Dans une partie de celles-ci, le fait de toucher le ballon avec un objet est suivi de l'éclatement de ce dernier et dans l'autre, il ne se passe rien de tel. Si l'on constate que dans les observations avec éclatement, l'objet est pointu et que dans les observations sans éclatement, il ne l'est pas, alors on pourra appliquer un "dispositional causal pattern" afin de retrouver la règle de la figure 2.3;

- les "historical causal patterns". Dans l'esprit, ils sont identiques aux "dispositional causal patterns" mais ne se focalisent pas sur les mêmes différences. Ils se concentrent en effet sur celles touchant les actions et les états ayant précédé

l'enchaînement action-état considéré. Ils rendent compte du fait qu'une action peut avoir des résultats différents en fonction des actions qui l'ont précédée.

Exemple : comme dans le cas le cas précédent, on dispose d'un ensemble d'observations que l'on peut scinder en deux groupes : les cas où un ballon éclate quand on le touche avec un objet pointu et les cas où il ne se passe rien. Si l'on remarque que les observations du premier groupe sont toutes précédées de la même action, en l'occurrence le fait de souffler dans le ballon, alors que ceci n'est pas vrai pour les observations du second groupe, on pourra alors appliquer un "historical causal pattern" qui mettra en évidence que souffler dans le ballon le place dans un état (le ballon est gonflé) tel que l'action de le toucher avec un objet pointu le fait éclater.

Sur un plan général, on notera que ces différents patrons forment une théorie générale de la causalité tandis que les règles qu'ils permettent de créer, qui en sont une sorte de spécialisation, représentent une théorie de la causalité liée au domaine considéré.

Les schémas

Dans OCCAM, la théorie du domaine est représentée à la base par un ensemble de relations causales élémentaires. Néanmoins, il est intéressant, aussi bien pour des raisons d'efficacité et de sensibilité que nous avons vues à propos de GENESIS, de reconnaître des structures de niveau supérieur exprimant des régularités dans l'agencement de ces relations causales. Les schémas correspondent à ces structures. Plus généralement, ces schémas ont ici la même fonction, le même usage et approximativement la même structure que ceux de GENESIS. Comme eux, ils sont organisés de façon hiérarchique.

Ainsi que le montre la figure 2.5, un schéma est composé de deux parties principales :

- un événement généralisé. Il s'agit d'une dépendance conceptuelle décrivant à un haut niveau quelles sont les circonstances, la nature et les résultats de l'événement considéré. Cette description sert de déclencheur à l'application du schéma.

  La figure 2.5 donne un exemple d'événement généralisé pour un schéma représentant une situation de contournement d'embargo : un pays *?A* refuse de vendre à un état *?T* une ressource (rôle *threat*)[1]; mais un autre pays accepte de vendre cette ressource à *?T* moyennant un prix plus élevé (rôle *response*). En final, *?T* obtient donc tout de même la ressource (rôle *result*).

  On remarquera que cette description, au travers de certains de ses rôles (ici *threat, response, result*), se décompose en différents attributs, autrement dit des rôles

[1] A priori, l'état *?A* veut ainsi faire pression sur l'état *?T* car ce dernier a besoin de cette ressource. Toutefois, tout cela n'est pas explicité dans le schéma.

faisant référence à des états et des actions. D'ailleurs, les valeurs de ces rôles ne se limitent pas à des dépendances conceptuelles mais prennent également la forme d'autres schémas dans le cas où il s'agit de situations plus développées et non d'actions ou d'états simples. Contrairement à GENESIS toutefois, ces attributs ne sont pas normalisés – ils peuvent donc changer d'un schéma à un autre – et n'incarnent pas la structure causale du schéma.

- une chaîne causale. Celle-ci rend compte des différentes relations existant entre les actions et les états composant le schéma. Ces relations sont de nature causale (*RESULT*, *ENABLES*), temporelle (*BEFORE*) ou explicitent les motivations des acteurs du schéma. Lors de la compréhension, alors que l'événement généralisé assure la sélection du schéma à partir des éléments présents dans le texte, la chaîne causale est la dimension du schéma qui apporte l'explication de leur présence, c'est-à-dire, si l'on assimile un schéma à une relation causale, qui met à jour comment on passe de l'action à l'état qui traduit son résultat.

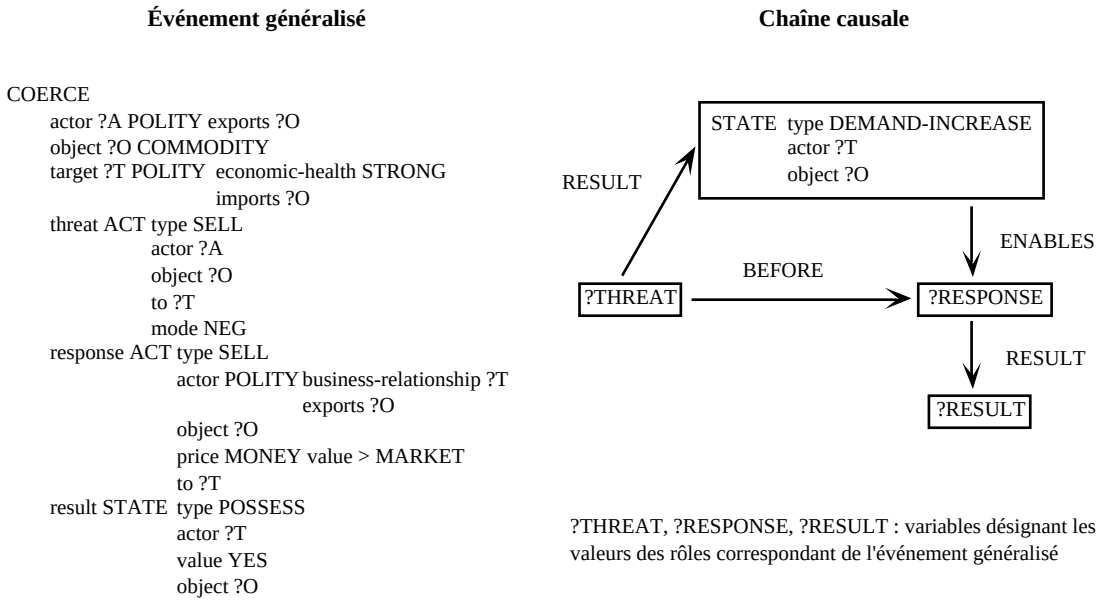


Fig. 2.5 - Un schéma (événement généralisé et chaîne causale) (d'après [Pazzani 1988])

On précisera enfin qu'à l'instar des relations causales élémentaires qui sont apprises par OCCAM, les schémas possèdent des informations à propos de la confiance que l'on peut leur accorder. En particulier, on enregistre le nombre de fois où ils ont été utilisés avec succès et au contraire, le nombre de fois où cette utilisation a conduit à un échec. Ils conservent en outre une trace des relations causales qui sont intervenues dans leur construction.

La dimension compréhension

Le processus de compréhension d'OCCAM est très proche de celui de GENESIS. Comme lui, son but est de produire une explication des enchaînements d'actions et d'états que l'on observe dans les textes. Les explications obtenues sont des chaînes causales similaires à celle illustrée par la figure 2.2.

Comme lui, il s'appuie sur les deux mêmes stratégies pour réaliser cet objectif. La première consiste à instancier un ou plusieurs schémas existants que l'on a sélectionnés à partir des éléments présents dans le texte. L'analyse s'apparente alors à un mécanisme progressif de confrontation des constituants des schémas, chacun de ces constituants représentant une hypothèse, avec les éléments provenant du texte, lesquels sont assimilés à autant de faits susceptibles de confirmer ou d'infirmer les hypothèses avancées. Si l'application d'un schéma se trouve confirmée par ce moyen, on peut alors transposer les relations existant entre les constituants de ce schéma aux éléments similaires du texte.

La seconde stratégie est quant à elle plus constructive dans la mesure où elle consiste à chaîner les uns aux autres les actions et les états explicités par le texte en utilisant les relations causales élémentaires présentes au niveau de la théorie du domaine mais en guidant ce processus au moyen de connaissances très générales sur la façon dont s'effectuent ces enchaînements. Ces connaissances s'incarnent dans des principes très largement applicables tels que ceux-ci par exemple : un changement d'état peut être le résultat d'une action; un acteur peut réaliser une certaine action à la suite d'une autre si cette dernière est la source d'un but de l'acteur que la première action satisfait. Globalement, cette démarche d'analyse s'assimile au travail réalisé par les planificateurs.

La dimension apprentissage

La dimension apprentissage d'OCCAM est véritablement ce qui constitue son originalité. Celle-ci repose sur sa volonté de concilier les approches analytiques et synthétiques de l'apprentissage, à la fois par la mise en œuvre d'une technique médiane, le TDL, et par l'agencement complémentaire de modules spécifiques de chacune de ces deux tendances.

Cet apprentissage permet l'acquisition des trois types de connaissances que nous avons décrits précédemment : les relations causales élémentaires, les patrons de causalité et les schémas. Le point de départ est l'apprentissage des relations causales élémentaires puisque celles-ci servent à la fois à l'apprentissage par EBL des schémas et à l'apprentissage de nouveaux patrons de causalité. L'apprentissage des schémas consiste à

enregistrer la façon dont ces relations sont prototypiquement agencées alors que la création de nouveaux patrons de causalité résulte de leur abstraction. L'apprentissage des relations causales s'effectue quant à lui par SBL, ou par TDL lorsqu'un patron de causalité est applicable. En termes de connaissances initialement nécessaires pour l'amorçage de cette chaîne d'apprentissage, les patrons de causalité viennent donc en tête mais ils ne sont pas seuls dans la mesure où certaines connaissances de haut niveau restent nécessaires (cf. processus d'analyse du type planification et EBL).

L'apprentissage de schémas par EBL

La façon dont l'EBL est mis en œuvre ici s'apparente à l'EBR (Explanation-Based Refinement) utilisé par AQUA (cf. description ci-après au §2.5). L'analyse d'un texte met en évidence un certain nombre de relations causales, lesquelles suggèrent à leur tour un schéma d'explication abstrait. Ce schéma est ensuite raffiné en fonction des informations présentes dans le texte. Une généralisation de chacun des constituants du schéma instancié obtenu est enfin opérée en prenant soin de conserver la validité de l'explication représentée, et donc la validité des relations causales regroupées.

Les schémas ainsi construits s'appuient sur des relations causales qui ont été apprises au préalable en ayant recours dans le meilleur des cas au TDL ou moins favorablement au SBL. Quoi qu'il en soit, ces relations n'ont pas un caractère de certitude et peuvent être remises en cause. Il en est donc de même pour les schémas qu'elles ont contribués à bâtir. Ainsi, lorsqu'un schéma appris conduit à réaliser une mauvaise prédiction, une des actions déclenchées consiste à vérifier si les relations causales qui le constituent sont toujours considérées comme valides. Si l'une d'elles a été éliminée, le schéma est lui aussi supprimé.

L'apprentissage de relations causales par SBL

L'apprentissage de type SBL opéré par OCCAM s'apparente à ce qui est réalisé par le système UNIMEM que nous avons décrit au chapitre 1. De manière comparable aux patrons de causalité, il s'appuie sur la dimension temporelle afin de faire émerger la dimension causale. Les différents événements relatés par un texte sont ainsi répartis en différentes tranches de temps et les règles causales apprises sont dérivées d'enchaînements action-état dans lesquels l'action et l'état se trouvent dans la même tranche de temps.

L'apprentissage s'opère en deux temps. Une première phase consiste à regrouper sur des critères de similarité assez syntaxiques des enchaînements action-état apparus dans différents textes. On formera par exemple l'ensemble de tous les enchaînements dans

lesquels le fait de toucher un ballon est suivi, au cours de la même unité temporelle, de son éclatement. La seconde phase vise à créer une description généralisée pour chacune des classes d'enchaînements constituées. Si une telle description peut être construite, elle devient une nouvelle règle causale. Cette construction est opérée en supprimant les rôles n'ayant pas de valeurs homogènes entre les différents exemples et en généralisant de façon minimale les valeurs des autres.

Après avoir été élaborées, les nouvelles règles poursuivent leur évolution de la même façon que les Gen-Nodes dans UNIMEM. Lorsque de nouveaux textes sont analysés, ces règles sont appliquées afin de prédire les états devant apparaître et l'on confronte le résultat de cette application avec ce qui est effectivement présent dans les textes. On peut de cette manière augmenter ou diminuer la confiance que l'on accorde à une règle ou bien encore créer de nouvelles règles plus spécifiques

L'apprentissage de relations causales par TDL

Le TDL est la procédure qui permet d'appliquer les patrons de causalité afin de créer de nouvelles règles causales propres au domaine considéré. À partir d'un même enchaînement action-état, il est en effet possible d'appliquer un grand nombre de ces patrons et il est donc nécessaire de disposer d'une méthode de choix.

Le TDL est déclenché dès que l'on observe un changement d'état inexpliqué. Il se déroule en quatre étapes. On commence par extraire de la mémoire les observations similaires, c'est-à-dire des changements d'état ayant la même action de départ. On partitionne ensuite l'ensemble obtenu en deux sous-ensembles suivant que le changement d'état possède (exemple positif) ou non (exemple négatif) le même état final que celui que l'on cherche à expliquer.

L'étape suivante est l'application proprement dite des patrons de causalité. Le principe général sous-tendant cette opération consiste à opter toujours en faveur du patron le plus simple lorsque plusieurs sont en concurrence. On cherchera donc à appliquer en premier les patrons de type "exceptionless", sauf en cas de présence d'exemples négatifs, avec lesquels ils sont incompatibles, puis ceux de type "dispositional" et enfin ceux de type "historical". Lorsque la concurrence se situe au sein de l'une de ces classes, le choix est réalisé aléatoirement. Avant de réaliser l'application proprement dite d'un patron, on effectue une première généralisation en ne retenant que les rôles communs à tous les changements d'état observés. C'est cette observation généralisée qui est ensuite appariée aux antécédents des patrons de causalité.

Enfin, la dernière étape conduit à la création des nouvelles règles causales suivant une procédure spécifique de chaque classe de patrons. Pour ceux de type "exceptionless", il

suffit d'instancier le conséquent du patron en fonction du résultat de l'appariement de l'observation généralisée avec l'antécédent. Pour les deux autres classes, il faut au préalable déterminer quel est l'élément de l'observation influant sur le changement d'état. On utilise à cet effet la dichotomie entre exemples positifs et exemples négatifs. Les patrons de type "dispositional" spécifient le rôle de l'action concerné mais il reste à préciser le trait de la valeur de ce rôle qui est important du point de vue causal (par exemple le fait que l'objet soit pointu dans le cas de la règle de la figure 2.3 sur l'éclatement d'un ballon). Pour cela, on rassemble l'ensemble des traits des valeurs de ce rôle au sein des exemples positifs et l'on ne retient parmi eux que ceux n'intervenant pas dans ce même rôle au sein des exemples négatifs. En final, on choisit au hasard un des traits restants pour figurer dans la règle causale. La procédure est identique pour les patrons de type "historical" en remplaçant les rôles par les actions précédant le changement d'état.

La validation des règles construites par TDL, nécessaire du fait de l'absence de contrôle par une théorie du domaine, est du même type que celle prévalant pour celles construites par SBL : on incrémente ou on décrémente un compteur associée à chaque règle en fonction du succès ou de l'échec des prédictions qu'elle réalise. Si la valeur de ce compteur tombe en dessous d'un seuil donné, la règle est supprimée.

L'apprentissage de patrons de causalité

L'objectif est de produire de nouveaux patrons de causalité à partir des relations causales créées par SBL. La méthode d'apprentissage retenue à cet effet se fonde également sur la notion de similarité entre exemples. Néanmoins, cette dernière est exploitée ici différemment de l'usage qui en est fait dans l'apprentissage des relations causales. Plutôt que d'attendre la présence de suffisamment d'exemples pour généraliser, il s'agit en effet de créer un nouveau patron à partir d'une seule relation causale et d'affiner ensuite sa définition en fonction des relations causales similaires rencontrées par la suite.

En pratique, on crée ainsi un patron de type "exceptionless" et un patron de type "dispositional" pour chaque changement d'état constaté dans un texte. Dans le premier cas, on définit les contraintes d'égalité portant sur les rôles en fonction de la présence d'une même valeur entre un rôle de l'action et un rôle de l'état[1]. Dans le second, on retient en plus chacun des traits de ces valeurs communes comme acteur possible du changement d'état. Le patron ainsi créé est comparé à ceux déjà présents en mémoire et si

[1] Pour qu'un patron puisse être créer de cette manière, il faut au moins qu'un objet soit commun à l'action et à l'état.

l'un d'entre eux est suffisamment proche, ils sont fusionnés. On ne retient alors que les contraintes d'égalité communes ainsi que, pour les patrons de type "dispositional", les traits communs des valeurs associées jugés causalement significatifs. Chaque fusion de deux patrons opérée de la sorte permet de généraliser progressivement les patrons de causalité et les rend ainsi plus largement applicables.

Discussion

Sur un plan général, le système OCCAM est particulièrement intéressant dans la mesure où il fait cohabiter de façon complémentaire des mécanismes relevant de deux stratégies d'apprentissage différentes. De notre point de vue, cet intérêt est renforcé par le fait que cette cohabitation est conçue dans une perspective d'amorçage. Les techniques fondées sur la similarité[1] sont utilisées afin d'amasser les connaissances (i.e. les relations causales élémentaires) nécessaires au fonctionnement des techniques à base de connaissances. Il est ainsi possible, sur le principe, de faire appel à un noyau restreint de connaissances d'un certain type (i.e. les patrons de causalité) afin d'acquérir un vaste ensemble de connaissances d'un autre type (i.e. les schémas).

Cette dimension doit d'ailleurs être soulignée étant donné que ce n'est pas la seule façon de faire cohabiter les méthodes fondée sur la similarité et celles à base de connaissances. [Lebowitz 1990] montre ainsi comment utiliser l'EBL à la suite du SBL dans UNIMEM afin d'expliquer et de valider les généralisations réalisées. Dans une autre optique encore, [Danyluk 1987] applique d'abord l'EBL afin de produire une explication simple des événements relatés dans des textes puis utilise cette explication pour déterminer sur quels éléments la similarité doit porter lors du SBL.

Notre intérêt pour OCCAM est néanmoins teinté d'une certaine réserve quant à la liaison entre apprentissage et compréhension. Il nous semble en effet que les possibilités mises en avant par l'apprentissage, notamment en ce qui concerne l'amorçage, se trouvent en quelque sorte bridées par la dimension compréhension. Ce phénomène se manifeste aussi bien pour l'apprentissage des schémas que pour celui des relations causales.

La figure 2.6 montre le début d'un des textes sur lesquels le TDL a été appliqué. Elle met nettement en évidence la nécessité d'une part, de faire apparaître de façon explicite les conséquences des actions et d'autre part, de découper les textes en intervalles de temps afin de délimiter les changements d'état. Des tests ont certes été menés pour montrer la résistance du TDL vis-à-vis du bruit touchant ces deux paramètres mais ce bruit doit rester néanmoins limité. Or, transposées à des textes moins contrôlés, ces deux exigences

[1] On inclut le TDL dans ces techniques dans la mesure où les patrons de causalité ne servent qu'à guider et non à valider. Par ailleurs, ils ne sont pas spécifiques d'un domaine, ce qui rend envisageable leur énumération.

impliquent la présence de connaissances qui ne sont pas placées dans le champ d'apprentissage d'OCCAM et qui ne sont pas non plus véritablement explicitées en tant que besoin initial.

(0) Karen was thirsty. (1) She pushed the door away from the cupboard. The cupboard was open. (2) She took a small red plastic cup from the cupboard. Mike pushed the light switch. She had the cup. The cup was not in the cupboard. The light was on. (3) She pushed the door to the cupboard. The cupboard wasn't open. ...

Fig. 2.6 - Un extrait de texte servant de support au TDL (d'après [Pazzani 1991])

Le problème est globalement similaire pour les schémas puisque le processus de compréhension qui construit la représentation du texte servant à leur apprentissage s'appuie soit sur les schémas déjà existants, soit sur des connaissances générales à propos de la causalité et des motivations des personnages. Or, ces dernières ne font pas partie non plus des connaissances apprises par OCCAM et s'assimilent en revanche aux connaissances devant être fournies a priori au système GENESIS. Il ne semble donc pas possible dans OCCAM d'envisager l'apprentissage de schémas en partant uniquement des relations causales apprises soit par SBL, soit par TDL.

5. AQUA

Présentation

AQUA est un système de compréhension d'histoires, en l'occurrence des textes de type dépêches d'agence traitant principalement d'actes de terrorisme, reposantentièrement sur la notion de question. Une question, que Ram nomme également but de connaissance, traduit une attente que possède le compreneur. Mais contrairement aux systèmes, tels que GENESIS par exemple, fondant leur compréhension sur l'activation de schémas, cette attente n'est pas assimilable à une connaissance par défaut portant sur la situation considérée et que l'on cherche à reconnaître dans les textes. Il s'agit véritablement de questions auxquelles le système cherche à répondre, c'est-à-dire de connaissances qu'il se fixe comme but de trouver au sein des textes qu'il traite. Toute la compréhension dans AQUA s'effectue donc au travers d'un jeu de questions-réponses évoluant continuellement. Un nouveau texte permet de répondre à certaines questions restées en suspens précédemment mais celui-ci, en offrant les moyens de pousser la compréhension plus loin, est également la source de nouvelles interrogations. Grâce à ce mouvement de va-et-vient, on raffine progressivement la connaissance que l'on possède sur le domaine.

AQUA met ainsi en œuvre de façon naturelle un apprentissage incrémental et intègre la capacité de travailler à partir de connaissances incomplètes. La notion de question, qui fonde donc aussi l'apprentissage au sein d'AQUA, définit même l'architecture du système dans la mesure où tout le système s'organise autour d'un agenda de questions. Lorsqu'une question reste sans réponse lors du processus de compréhension, celui-ci est mis en attente et la question correspondante est enregistrée comme pendante. Lorsqu'une réponse peut y être apportée par le traitement de nouvelles entrées, ce processus est réactivé et poursuit son cours. Ce mécanisme de mise en attente/réactivation peut intervenir aussi bien à différents moments de l'analyse d'un même texte qu'entre analyses de textes différents.

AQUA applique cette approche à base de questions au domaine des explications motivationnelles. L'analyse qu'il opère vise en effet à mettre en évidence les buts et les croyances qui sous-tendent les actions des différents personnages d'un récit. Elle est réalisée par un raisonnement à base de cas reprenant pour une part importante celui développé dans le cadre du système SWALE [Kass et alii 1986]. L'utilisation de ce type de raisonnement provient ici à la fois d'un souci d'efficacité, des connaissances générales étant par ailleurs toujours disponibles pour produire les explications, et du souci de produire l'explication la plus juste et la plus précise possible par rapport à la situation abordée. Ce sont là des caractéristiques proches de celles de GENESIS et l'on peut donc affirmer que globalement, AQUA s'inscrit dans la mouvance de l'EBL.

La représentation des connaissances

Comme tous les systèmes de compréhension de textes opérant en profondeur, AQUA dispose d'un large éventail de connaissances. Il possède ainsi un ensemble de concepts lui permettant de représenter les objets, les actions et les états du monde de référence. Ces concepts sont organisés au sein d'une structure hiérarchique autorisant l'héritage multiple. Par ailleurs, AQUA utilise les MOPs pour caractériser les situations, vues ici comme des regroupements d'actions et donc assimilables à des actions complexes.

Mais le type de connaissances véritablement central dans AQUA est la notion, reprise de SWALE, de *schémad'explication* ("explanation pattern"), appelé aussi *XP* [Schank 1986]. Un schéma d'explication représente un agencement prototypique d'actions et d'états unis par des relations causales, cette configuration explicitant les raisons pour lesquelles un agent exécute une action. Il s'agit donc de connaissances proches des TOPs proposés par Schank. Les relations causales présentes dans les XPs font partie des

connaissances de base d'AQUA au même titre que les concepts et les MOPs. Elles peuvent être vues comme des XPs primitifs, sans structure interne.

On distingue deux types d'XPs : les schémas d'explication abstraits et les cas explicatifs ("explanatory cases"). Les premiers font partie des connaissances initiales d'AQUA et rendent compte des principes généraux qui lient les buts, les croyances, les états émotionnels et sociaux des personnages à leurs actions. Un principe tel que "un acteur réalise une action parce que le résultat de cette action satisfait un de ses buts" en est un exemple typique. Ils forment une sorte de théorie générale de la motivation des personnages. Les cas explicatifs sont au contraire acquis par AQUA au fur et à mesure de son fonctionnement et représentent des explications motivationnelles liées à des situations spécifiques. Ils sont le résultat de la spécialisation des XPs abstraits intervenue à la suite du traitement d'un ou de plusieurs textes relatifs à ces situations. Une affirmation telle que "les shiites libanais pratiquent des attentats suicides à la bombe parce que ce sont des fanatiques religieux" est un exemple du type de connaissances que ces XPs incarnent.

XP Sacrifice_d'un_but

EXPLAINS (PRE-XP-NODES)

8 - l'agent A a exécuté l'action M aboutissant à satisfaire son but G1 mais allant à l'encontre de son but G2

XP-ASSERTED-NODES

1 - l'agent A donne une priorité supérieure à G1 par rapport à G2

INTERNAL-NODES

2 - A croît que le fait d'accomplir l'action M permet de satisfaire G1
3 - A croît que l'accomplissement de l'action M va à l'encontre de G2
4 - A met en balance G1 et G2
5 - A décide d'accomplir M
6 - A réalise M
7 - l'accomplissement de M produit un effet positif du point de vue de G1 et négatif du point de vue de G2

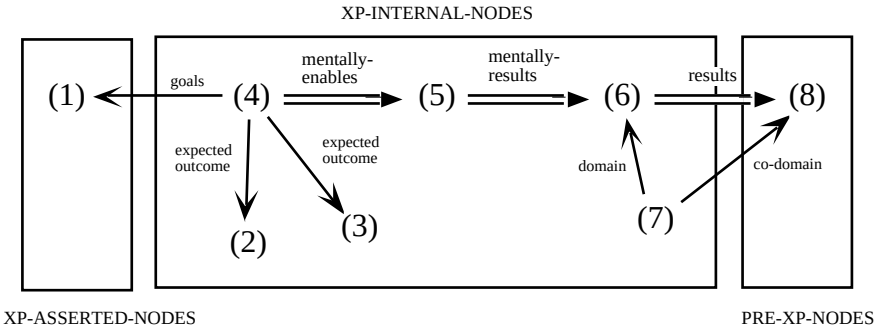


Fig. 2.7 - Un XP abstrait

La structuration de ces deux types d'XPs est toutefois identique. Un XP est un graphe acyclique orienté formé de trois types de nœuds :

- les XP-ASSERTED-NODES forment les prémisses de l'explication représentée par le XP. Ils constituent les conditions nécessaires à la dérivation des conclusions;
- les PRE-XP-NODES sont les conclusions de l'explication. Parmi eux, figure un nœud particulier, le nœud EXPLAINS, qui est plus spécifiquement l'élément expliqué par le XP. Il doit représenter une relation motivationnelle;
- les INTERNAL-XP-NODES permettent de faire le lien entre les XP-ASSERTED-NODES et les PRE-XP-NODES. Ce sont en quelque sorte les chaînes du raisonnement sous-tendant l'explication. Ces nœuds ont liés entre eux de même qu'aux nœuds des autres types par les relations causales élémentaires que nous avons évoquées précédemment.

La figure 2.7 donne un exemple d'XP abstrait représentant la notion de sacrifice : un agent A sacrifie un de ses buts, G2, pour accomplir un but plus important, G1.

Sur le plan de l'organisation générale, les différents XPs abstraits sont organisés hiérarchiquement. Les cas explicatifs sont quant à eux rattachés à l'XP abstrait dont ils sont la spécialisation. Ils sont par ailleurs indexés à la fois par la situation (action ou MOP M) ainsi que par le stéréotype de l'acteur de cette situation (agent A). L'indexation par ces trois dimensions permet, dans l'optique du raisonnement à base de cas utilisé par la compréhension, d'accéder rapidement aux cas explicatifs les plus pertinents vis-à-vis de la relation motivationnelle que l'on cherche à expliquer dans un texte.

La dimension compréhension

La compréhension dans AQUA s'articule autour de deux processus. On distingue ainsi le processus d'Explication d'une part, et le processus de Lecture d'autre part. Les deux opèrent à partir de la représentation sémantique des propositions constituant les textes. Dans le dialogue question-réponse que nous avions évoqué dans la présentation générale, le premier est celui qui pose les questions tandis que le second est celui chargé d'apporter les réponses.

Le processus d'Explication

Ce processus a pour vocation d'expliquer toute action intentionnelle apparaissant dans un texte. La base de son fonctionnement est la détection d'anomalies dans les actions qui lui sont soumises. Une anomalie, en l'occurrence, est un comportement allant à l'encontre des buts de l'agent considéré. Par exemple, le fait de commettre un attentat-suicide est une anomalie dans la mesure où cette action vient s'opposer à l'un des buts premiers des individus qui est de préserver leur vie. La recherche de ces anomalies est guidée dans

AQUA par les XPs abstraits qui lui sont donnés initialement. Le nœud EXPLAINS de chacun d'entre eux représente en effet une question à laquelle on peut soumettre l'entrée courante.

Dans l'exemple de la figure 2.7, le XP considéré est source de la question "l'action exécutée par l'agent a-t-elle conduit à la satisfaction d'un but de cet agent au détriment d'un autre de ses buts?". Une réponse positive à cette question conduit à sélectionner l'XP *Sacrifice_d'un_but* comme explication possible de l'action. La hiérarchie des XPs abstraits traduit l'ordre imposé sur les questions à poser, le tout formant une sorte de réseau discriminant. La question "l'agent veut-il le résultat de l'action" est ainsi suivie, en cas de réponse négative, de la question "l'agent connaît-il le résultat de l'action?".

Si une anomalie est détectée, le processus d'explication est mis en action. Celui-ci repose sur un raisonnement à base de cas cherchant à appliquer des XPs déjà en mémoire pour résoudre l'anomalie décelée. Ces XPs sont préférentiellement des cas explicatifs, supposés particulièrement adaptés à la situation abordée, ou des XPs abstraits, qui apportent une explication générale dans le cas où aucun XP spécifique ne peut être appliqué. La recherche des cas explicatifs s'opère par un processus également à base de questions, portant sur les trois dimensions servant à indexer les XPs : l'anomalie qu'ils expliquent (cf. XPs sélectionnés par la détection des anomalies), l'action ou la situation qui leur sert de cadre et le personnage qui est l'agent de cette action ou de cette situation.

À l'issue de cette phase de recherche, on obtient donc un ensemble d'XPs dont on réalise ensuite l'instanciation en fonction des données propres au texte traité. Cette instanciation commence par le nœud EXPLAINS de chaque XP et se poursuit, à la manière d'une abduction, en "remontant" les relations causales de l'XP et en instanciant ses INTERNAL-XP-NODES. Le processus s'arrête au niveau des XP-ASSERTED-NODES où trois cas se présentent pour chaque nœud : soit le XP-ASSERTED-NODE est vérifié par les données déjà disponibles; soit il est réfuté, auquel cas c'est l'XP dans son entier qui est réfuté car tous ses XP-ASSERTED-NODEs doivent être confirmés pour que l'XP s'applique; soit il reste en tant qu'hypothèse à vérifier.

Chaque anomalie détectée donne ainsi lieu à la construction d'un arbre d'hypothèses. Sa racine est constituée par l'anomalie elle-même. Ses nœuds de premier niveau sont formés par les différents XPs ayant été sélectionnés et instanciés pour expliquer l'anomalie. Chacun de ces XPs est appelé une hypothèse. Il donne lieu à autant de feuilles dans l'arbre qu'il possède de XP-ASSERTED-NODES non vérifiés. Ceux-ci sont des questions de vérification d'hypothèse ("Hypothesis Verification Questions" (HVQs)). Elles forment les attentes ultérieures du système.

Lorsque toutes les HVQs d'une hypothèse sont satisfaites, celle-ci est considérée comme vérifiée et ses concurrentes sont éliminées. On obtient alors l'explication de l'anomalie au travers de l'XP instancié. Si aucune hypothèse n'est validée, les questions restant en suspens sont indexées en fonction des concepts qui les composent et le processus d'explication de cette anomalie est mis en attente en attendant de nouveaux éléments.

Le processus de Lecture

Lorsque le processus de compréhension traite une nouvelle proposition, il commence par examiner si celle-ci ne permet pas de répondre à une question restée en suspens en mémoire. Les concepts constituant la proposition forment un index qu'AQUA utilise pour accéder aux questions auxquelles cette proposition est susceptible de répondre. Si une telle réponse est fournie, le processus d'explication attachée à la question est réactivé. On détermine alors si cette réponse réfute l'hypothèse concernée, si elle la valide ou si elle ne fait que supprimer une HVQ parmi d'autres demeurant pour le moment sans réponse.

Il est à noter que cet usage des nouvelles entrées n'empêche pas l'application du processus d'explication à leur égard si cela s'avère nécessaire. C'est d'ailleurs ainsi que l'on peut raffiner des explications déjà élaborées et pousser plus loin la compréhension.

La dimension apprentissage

Dans AQUA, l'apprentissage s'articule autour de trois axes. Le premier correspond à l'acquisition de nouvelles connaissances. En pratique, il s'agit de spécialiser des XPs abstraits pour former des cas explicatifs. Chacun d'entre eux caractérise la rencontre d'un schéma de décision général (l'XP abstrait) dans un contexte particulier. L'hypothèse sous-tendant cet apprentissage est celle d'une partie du raisonnement à base de cas : bien que la théorie générale sur le domaine contienne un grand nombre de potentialités, seules quelques unes sont en pratique réalisées. Il est donc efficace de les "noter" lorsqu'on les rencontre et de les utiliser prioritairement car il est fort probable de les rencontrer à nouveau par la suite.

Le deuxième axe est celui que nous avons appréhendé lors de la présentation générale. Il s'agit en effet du raffinement d'XPs déjà en mémoire via le mécanisme de question-réponse caractérisant la compréhension. Le but est donc de rendre plus achevées des connaissances incomplètes.

Le dernier axe concerne l'indexation des XPs au sein de la mémoire. Compte tenu de l'utilisation du raisonnement à base de cas, ce problème est de fait particulièrement important dans AQUA. Il s'agit à la fois d'améliorer l'indexation même des XPs, notamment en répertoriant tous leurs contextes d'utilisation, mais également de construire les index les plus pertinents possibles.

La création de nouveaux XPs

Lorsqu'un XP abstrait a été utilisé pour bâtir une explication dans une situation dans laquelle il n'avait jamais été employé auparavant, AQUA déclenche la procédure d'apprentissage visant à construire un nouveau cas explicatif à partir de cette explication. La méthode pour ce faire est appelée "Explanation-Based Refinement" (EBR). C'est une variante de l'EBL dans laquelle on spécialise un schéma d'explication général sur la base des relations causales mises en évidence dans une explication particulière.

La création d'un nouveau cas explicatif s'effectue en trois étapes. La première est une copie pure et simple de l'XP abstrait à partir duquel l'explication a été élaborée. La deuxième consiste à spécialiser les nœuds formant cet XP en fonction de ceux présents au niveau de l'explication et jouant le même rôle. C'est dans cette tâche qu'intervient plus spécifiquement l'EBR. Chaque nœud est en effet spécialisé de façon à retenir, dans la hiérarchie des concepts, les concepts les plus abstraits respectant les contraintes imposées par les relations causales présentes dans l'XP. On notera que l'on retrouve ici un mécanisme de même nature que celui présidant à la généralisation dans GENESIS.

La troisième étape vise à compléter l'XP abstrait par les informations nouvelles présentes dans de l'explication. Il est en effet possible que dans le cadre de la situation spécifique traitée, les prémisses de l'XP abstrait puissent être expliquées par des relations causales. Cela signifie d'une part, que les XP-ASSERTED-NODES concernés doivent être transformés en INTERNAL-XP-NODES et d'autre part, que les nœuds expliquant les anciennes prémisses doivent être ajoutés en tant que nouvelles prémisses. Ils sont de fait ajoutés comme autant de XP-ASSERTED-NODES et généralisés. Dans leur cas, on retient les concepts situés au dessus d'eux dans la hiérarchie des concepts et respectant les relations causales dans lesquelles ils sont impliqués. Au contraire des nœuds ayant un équivalent dans l'XP abstrait initial, il n'existe en effet pas de référence supérieure que l'on puisse leur appliquer. Une dernière dimension de cette troisième étape consisterait à expanser certains nœuds ou certaines relations pour lesquels on aurait une description plus détaillée (des XPs de plus bas niveau par exemple). Elle n'a cependant pas été implémentée.

Le raffinement des XPs existants

La clé d'un apprentissage progressif est la capacité à gérer des connaissances incomplètes. Le raffinement progressif des XPs participe à cette aptitude. Il satisfait deux besoins : d'une part, répondre aux questions restées en suspens dans le cas où une hypothèse a pu être vérifiée par d'autres moyens que la vérification de toutes ses prémisses, comme cela se produit lors de la compréhension; d'autre part, pousser plus

avant les explications en cherchant à expliquer les prémisses des XPs. Ce dernier besoin conduit à la production de nouvelles questions qui viennent entretenir ce cycle de questions-réponses. Lorsqu'une réponse est trouvée dans un texte, elle est bien entendu généralisée de façon à s'ajuster au niveau initial de la question.

L'indexation des XPs

Lorsqu'un nouvel XP est intégré au sein de la mémoire, il est indexé par trois de ses caractéristiques : le type d'anomalies qu'il explique, en l'occurrence l'XP abstrait dont il est la spécialisation; l'action ou la situation dont l'XP est la justification; enfin, le stéréotype de l'acteur de la situation. Le choix de ces index s'explique aisément par le fait qu'un XP est une explication apportée à la présence d'un couple situation-acteur.

En ce qui concerne le type d'anomalies, la détermination de l'index relève du processus de compréhension. Dans le cas de la situation et de l'acteur, le principe général du choix de l'index est identique : on retient la généralisation de l'élément initialement présent dans l'XP (en remontant dans la hiérarchie correspondante) qui respecte les relations causales imposées par l'XP. L'objectif est encore une fois de conserver la plus grande spécificité possible afin d'éviter les généralisations abusives qui rendraient le raisonnement à bas de cas moins efficace.

L'indexation suivant l'acteur de la situation présente une particularité supplémentaire. Les différents stéréotypes ne sont pas donnés a priori comme c'est le cas des situations mais sont construits par apprentissage. Un stéréotype est composé d'un ensemble de traits, tels que les buts et les croyances du personnage, impliqués directement dans le type d'explication construit ici. L'exemple suivant donne le stéréotype construit pour un jeune libanais forcé à commettre un attentat suicide à la bombe suite à un chantage exercé sur lui :

Typical goals	Preserve_life (t) Destroy(Object) (f) Avoidance_goal(State) (q)
Typical goal-orderings	Avoidance_goal(State) over Preserve_life (q)
Typical beliefs	Religious_zeal = not fanatic (t)

Typical features Age = teenage age (h)
Religion = Shiite Moslem (h)
Gender = Male (h)
Nationality = Lebanese (h)

avec

t : trait qui doit être vrai pour le personnage

f : trait qui doit être faux pour le personnage

q : trait incomplet donnant lieu à une question pouvant être complétée ultérieurement

h : trait présent au niveau des exemples ayant servi à bâtir le stéréotype mais sans relation causale explicitée avec les explications construites. Ces traits sont considérés comme des hypothèses et ont une valeur d'indice.

La détermination des traits composant un stéréotype est réalisée suivant le principe de généralisation des index exposé ci-dessus : chaque trait est donc le résultat d'une généralisation d'un élément d'XP respectant les contraintes imposées par celui-ci. Par ailleurs, les stéréotypes sont eux-mêmes la source d'un apprentissage incrémental. Les traits qui sont marqués en tant que question entrent en effet dans le mécanisme de question-réponse du processus de compréhension. Enfin, les traits hypothèses peuvent se voir justifiés par des explications provenant de nouveaux textes et prendre alors le statut de trait à part entière.

Discussion

Parmi les trois systèmes présentés précédemment, AQUA est le plus proche de GENESIS de par sa stratégie générale. Dans les deux cas, il s'agit en effet de spécialiser des connaissances abstraites et largement applicables afin de les rendre plus opérationnelles. Les critiques faites à GENESIS sur ce point peuvent donc être reprises pour AQUA et même amplifiées puisque celui-ci gère un ensemble de connaissances plus étendu encore. En particulier, on ne manquera pas de remarquer la taxinomie réalisée des anomalies possibles, qui n'est pas sans lien avec la rage classificatrice qui caractérise beaucoup des travaux s'inscrivant dans la lignée schankienne. Or le statut de ces connaissances est globalement flou. Elles apparaissent comme générales, c'est-à-dire modélisant les connaissances d'un être humain, mais la façon dont elles ont été mises en évidence ne garantit rien quant à leur portée réelle de ce point de vue et les rapproche plutôt de l'ingénierie des connaissances. Il est de même difficile de juger de leur étendue exacte par rapport à un ensemble qui n'est pas mieux cerné.

AQUA se distingue cependant de GENESIS sur une des caractéristiques importantes de ce dernier, à propos de laquelle nous avions émis une réserve. AQUA est en effet capable de réaliser un apprentissage incrémental alors que celui de GENESIS est résolument du type tout ou rien. AQUA peut ainsi cumuler l'analyse de plusieurs textes pour former un schéma et s'avère de ce fait moins tributaire de la forme des textes. Cette capacité trouve ses racines dans le couplage véritablement très étroit qu'il instaure entre compréhension et apprentissage au travers du mécanisme de question-réponse. La compréhension d'AQUA est en effet fortement dirigée par des buts de connaissance, c'est-à-dire des buts visant l'acquisition de nouvelles connaissances spécifiques. Mais elle repose également, au travers de la détection des anomalies, sur un questionnement général et systématique de la justification des nouvelles entrées. Ce questionnement entretient le cycle des questions-réponses en étant la source, par les tentatives d'application des connaissances existantes, de nouveaux buts de connaissance.

La progressivité de l'apprentissage dans AQUA le rapproche des principes que nous avons adoptés. Il en est de même de l'utilisation du raisonnement à base de cas. Toutefois, celle-ci est pour une part importante assez différente de notre conception. Dans AQUA, l'appellation de cas pour désigner les XPs spécifiques est un peu abusive dans la mesure où ce sont des structures résultant déjà d'un processus de généralisation. Dès lors, on ne situe pas très bien la différence entre le raisonnement à base de cas d'AQUA et le raisonnement à base de schémas utilisé dans GENESIS, les méthodes semblant en pratique assez similaires.

En revanche, il apparaît que le fait de spécialiser des schémas généraux en fonction de ce que l'on trouve dans les textes va au delà, dans AQUA, du simple accroissement d'efficacité ou de précision. Le mécanisme de raffinement des cas explicatifs, en relançant le processus d'explication sur les prémisses de ces XPs, permet en effet de poursuivre un apprentissage qui dépasse la simple spécialisation et qui correspond à un véritable approfondissement.

6. Discussion générale

Si les quatre systèmes que nous avons étudiés précédemment ont bien la capacité commune d'apprendre des connaissances pragmatiques à partir de textes, ils ont également la propriété commune de le faire en utilisant un ensemble de connaissances pragmatiques fournies a priori au système. C'est particulièrement vrai pour des systèmes comme GENESIS ou AQUA, qui revendiquent même l'utilisation de ces connaissances; mais c'est également ce que l'on peut constater pour un système plutôt orienté SBL,

comme IPP, ou pour OCCAM, dont l'utilisation de plusieurs stratégies d'apprentissage est pourtant réalisée dans la perspective d'un possible amorçage.

En final, on obtient donc un ensemble de systèmes illustrant chacun un certain nombre de principes intéressants mais limitant leur application à un micro-monde du fait des connaissances pragmatiques qu'ils requièrent. Celles-ci sont en effet toujours spécifiques d'un domaine. Même s'il est possible de mettre en évidence des principes généraux à propos de la causalité physique ou de la motivations des individus, il est difficile en pratique d'en dresser un inventaire exhaustif. Comme le montre OCCAM d'autre part, ils ne dispensent pas de la nécessité de posséder une théorie du domaine car ils sont trop généraux pour être appliqués systématiquement de façon directe. Leur spécialisation est donc impérative.

Les systèmes étudiés sont donc dépendants du domaine dans lequel a été démontré leur fonctionnement. Leurs auteurs soulignent toujours le fait que cette dépendance est comparable à celle existant entre le résultat d'un programme et les données qui lui sont fournies en entrée et qu'elle ne remet pas en cause le corps du système qui met en œuvre les idées avancées. Toutefois, cette affirmation reste globalement à démontrer dans la mesure où aucun de ces systèmes n'a été appliqué à un grand nombre de domaines différents. En particulier, il ne faut pas négliger des limitations en termes de changement d'échelle qui risquent de se manifester si l'on cumule un grand nombre de connaissances de différents domaines.

Par ailleurs, cette démarche pose le problème de la nécessaire modélisation manuelle de chaque domaine abordé à opérer préalablement à tout apprentissage automatique. Comme nous l'avons vu au chapitre 1, il s'agit d'un travail intrinsèquement difficile et qui n'est pas, de plus, facile à mener sur une large échelle. En fait, la démarche inverse semble plus adaptée : à partir d'un ensemble important de données, en l'occurrence des textes, un système automatique est capable de construire des connaissances qui pourraient être par la suite validées, voire corrigées, par un utilisateur.

Cette différence d'approche prend ses racines dans deux partis pris. D'abord, les systèmes considérés n'ont pas véritablement pour objet d'apprendre des connaissances pragmatiques à partir de textes. Il s'agit en fait pour eux plus d'un domaine d'application plus que d'un but en soi. Leur objectif véritable est en pratique de démontrer la validité d'une méthode d'apprentissage ou celle de principes cognitifs plus généraux. Cela explique que la nécessité d'un lourd travail de modélisation préalable ne soit pas dissuasive en regard des résultats obtenus par apprentissage.

Le second parti pris concerne le type de compréhension que l'on fait intervenir et les connaissances que l'on veut apprendre. Tous les systèmes que nous avons étudiés ont choisi de centrer leur processus de compréhension sur l'explication des raisons pour lesquelles différents événements s'enchaînent d'une manière spécifique au sein d'un texte. Ils cherchent ainsi à mettre à jour la causalité qui sous-tend cet enchaînement. La même logique est suivie pour le type des connaissances apprises. L'objectif est de faire émerger des schémas représentant le déterminisme causal sous-jacent à une situation donnée.

Or, chercher à déterminer le pourquoi d'une situation représente déjà un niveau de compréhension très élevé, qui explique ce besoin impératif de connaissances initiales. Cela explique également que les travaux mentionnés ne soient pas issus de la communauté de l'informatique linguistique. Celle-ci cherche en effet à valider de plus en plus ses méthodes sur une large échelle, ce qui a pour effet de limiter la profondeur des analyses que ces méthodes permettent de mener. Les travaux sur l'apprentissage de connaissances qui s'inscrivent dans ce cadre s'axent donc sur des problèmes comme la construction d'ontologies [Zweigenbaum et alii 1997] ou l'extraction de relations sémantiques [Béguin et alii 1997] pour lesquels on n'a pas nécessairement besoin d'un niveau d'analyse fin. L'existence de celui-ci à grande échelle enlèverait d'ailleurs une bonne part de justification à ces travaux.

Cette mise en perspective avec l'informatique linguistique souligne en outre un point commun aux quatre systèmes présentés. Ils s'appuient tous sur le résultat d'une analyse sémantique des textes sous forme de dépendances conceptuelles. De leur point de vue, les textes se présentent donc comme des collections de faits formalisés et la dimension analyse linguistique est clairement placée en dehors du champ de leur objet. Or, considérer comme acquis la possibilité de mener une analyse sémantique de façon automatique est clairement une simplification méritant une certaine attention.

Une position extrême à ce propos pourrait être de remettre complètement en cause la validité de ces travaux du fait de l'incertitude planant sur leurs pré-requis. Non seulement nous ne savons pas s'il sera possible un jour d'obtenir une analyse sémantique dans les conditions d'existence qui lui sont fixées par ces travaux mais plus gênant encore, en l'absence d'une procédure systématique d'analyse, il règne un flou, qui n'a rien d'artistique, sur la forme précise du résultat de cette analyse. En particulier, tout le problème de l'équivalence entre différentes formes d'expression et donc, du choix d'un niveau de représentation homogène, se retrouve volontairement ou involontairement laissé dans l'ombre. En résumé, cette position revient à nier l'intérêt de toute recherche sur le type d'apprentissage considéré ici avant d'avoir résolu le problème de l'analyse sémantique si l'on admet que cette dernière lui est indispensable.

Même si elle semble supportée par des arguments convaincants, cette conception se heurte elle aussi à une difficulté méthodologique. Chaque niveau de traitement (morphologie, syntaxe, sémantique, pragmatique) est la source d'ambiguïtés qu'il est nécessaire de lever afin de construire le sens des énoncés qui sont analysés. Il est néanmoins reconnu qu'il est d'autant plus aisé de résoudre les ambiguïtés apparaissant à un niveau qu'il est possible de s'appuyer sur des indications provenant du niveau supérieur. C'est ainsi par exemple que les meilleurs étiqueteurs sont globalement ceux qui procèdent à une analyse syntaxique, même partielle. L'intérêt d'une telle approche a été également démontré par un analyseur déterministe comme ANDI [Rady 1983] au sein duquel des informations de nature sémantique, voire pragmatique [Séligman 1985] permettent de lever des ambiguïtés syntaxiques. Cet intérêt est illustré à un degré supérieur encore au travers d'un système comme CARAMEL [Sabah & Briffault 1993] dont l'architecture multi-agents a justement pour objet de mettre en œuvre la flexibilité nécessaire à ce type de rétro-actions. Il n'est donc pas du tout évident qu'une approche totalement du type bottom-up soit viable pour aborder le problème qui est le nôtre.

Cette constatation justifie notre volonté de mettre l'accent sur l'importance de l'amorçage. L'hypothèse sous-jacente est la suivante : des capacités d'analyse, même faibles, offrent la possibilité d'accumuler un certain nombre de ressources si elles sont couplées avec un mécanisme d'apprentissage. Ces ressources permettent de construire un processus d'analyse plus puissant qui, à son tour, pourra servir de support à l'apprentissage de nouvelles ressources, elles-mêmes plus élaborées. De proche en proche, on peut ainsi espérer faire évoluer à la fois les connaissances et les capacités de compréhension d'un système de façon à parvenir au niveau auquel se placent les travaux que nous avons étudiés. On notera d'ailleurs qu'OCCAM met déjà en pratique ces idées mais limite leur application à la dimension apprentissage, ce qui restreint leur portée.

Par ailleurs, cette démarche ne remet en cause ni la nécessité d'une flexibilité des processus de compréhension à un stade donné de "développement", ni l'intérêt, pour lever les ambiguïtés à un niveau, de s'appuyer sur des connaissances des niveaux supérieurs. Elle se contente en fait d'affirmer que ces principes peuvent être appliqués aux différents stades d'évolution des connaissances et des processus de compréhension, sans chercher à modéliser directement un état terminal qu'il est difficile de cerner ex nihilo.

Ramenée au problème que nous nous posons ici, cette idée suggère que le niveau de compréhension visé par les systèmes étudiés est trop élevé pour être abordé de but en blanc. Avant de mettre à jour les ressorts d'une situation, une première tâche consiste tout simplement à déterminer ce qui la constitue. Autrement dit, avant de s'intéresser au

pourquoi, on peut déjà essayer de préciser le quoi, c'est-à-dire les actions, les états et les acteurs composant la situation.

Dans ce cadre, la dimension compréhension s'identifie au problème de la reconnaissance des situations dans les textes, donc à une forme d'analyse thématique. On peut remarquer d'ailleurs que dans tous les systèmes présentés ici, les textes traités n'abordent qu'une seule situation, qu'ils développent de façon très linéaire. Le caractère parfois un peu artificiel de ces textes explique qu'un découpage en situations ne soit pas nécessaire. Toutefois, la forme des textes est généralement beaucoup moins favorable et il est fréquent que les exposés de situations différentes soit assez intriqués les uns dans les autres, en particulier lorsqu'il s'agit de textes journalistiques. L'analyse thématique devient alors indispensable et se révèle déjà en elle-même une tâche ardue. Se concentrer d'abord sur cette question avant d'aborder le problème de l'explication des faits n'est donc pas uniquement une simplification mais une étape préalable nécessaire pour que le processus d'explication intervienne dans un cadre cohérent.

Du point de vue de l'apprentissage, faire émerger les situations et leurs éléments caractéristiques sans chercher dans un premier temps à établir la causalité qui les sous-tend présente l'avantage de se prêter plus facilement à un apprentissage de type SBL et donc, de ne pas requérir les connaissances mêmes que l'on veut acquérir.

Par ailleurs, l'existence d'une telle représentation des situations offre la possibilité d'apprendre dans des conditions beaucoup plus favorables les connaissances de nature causale qui servent à construire des explications. Les mécanismes développés par exemple au sein d'OCCAM pour l'apprentissage de relations causales (par SBL ou TDL) pourraient de fait être plus précis en disposant de l'information que deux changements d'état qu'ils se proposent de regrouper sont intervenus dans des situations proches, voire identiques, ou au contraire dans des situations radicalement différentes. Les connaissances que l'on construit ainsi ne sont donc pas vues comme une fin en soi mais s'inscrivent dans une évolution vers des traitements plus élaborés.

Récapitulatif

Ce chapitre nous a permis de présenter quatre systèmes réalisant un apprentissage automatique de connaissances pragmatiques à partir de textes. Le plus ancien d'entre eux, IPP, est proche de notre point de vue par l'utilisation qu'il fait de la similarité et de la récurrence des traits des situations pour généraliser.

GENESIS incarne quant à lui une approche assez radicalement opposée puisqu'il met en avant la possibilité de construire une représentation d'une situation à partir d'une seule évocation de celle-ci. Cette aptitude est sous-tendue par un processus explicatif produisant

une représentation causale des textes ainsi que par la présence des connaissances sur le domaine servant tout à la fois à alimenter ce processus et à guider la généralisation. Cette démarche se retrouve dans AQUA qui y ajoute la capacité, importante à nos yeux, de construire les représentations des situations de façon progressive, en allant chercher les informations qui lui sont nécessaires dans différents textes.

Enfin, le système OCCAM illustre la possibilité de faire cohabiter ces deux approches de façon cohérente et d'obtenir ainsi un système multi-stratégies de plus large application que les systèmes précédents. Cette combinaison est en effet mise au service de l'amorçage : les techniques à base de similarité sont utilisées afin de constituer les connaissances qui sont nécessaires au fonctionnement des techniques à base de connaissances.

En dépit de leurs spécificités, ces quatre systèmes présentent un ensemble de caractéristiques communes qui en limitent la portée de la même façon. Ils s'appuient tous sur une analyse sémantique, cherchent à construire une représentation causale des textes et supposent pour cela l'existence a priori d'une quantité plus ou moins importante de connaissances sur le domaine où ils sont appliqués.

Notre objectif étant véritablement l'acquisition de nouvelles connaissances et pas seulement la démonstration de certains principes, nous plaidons quant à nous en faveur d'une démarche plus progressive dans laquelle le niveau d'analyse de ces systèmes est vu, non pas comme un pré-requis, mais comme un but à plus long terme que l'on peut atteindre par paliers. Dans ce processus, l'activité se déroulant à chacun des stades contribue à l'élaboration des connaissances qui seront utiles au mécanisme d'analyse du stade suivant.

Plutôt que de nous intéresser directement à la mise en évidence de la causalité motivant l'enchaînement des actions, nous avons donc choisi de nous concentrer sur le problème de la reconnaissance des situations dans les textes, plus directement abordable sans l'hypothèse de la nécessaire présence de connaissances sur le domaine. En outre, cette analyse thématique apparaît comme un pré-requis pour un processus d'explication dans la mesure où il lui permet de délimiter son champ d'intervention en définissant ce qui doit être expliqué.

Le pendant de ce choix en ce qui concerne la dimension apprentissage est la volonté de ne pas chercher d'emblée à acquérir des schémas causaux mais plutôt de faire émerger en premier lieu les actions, les états et les acteurs qui forment le corps des situations.

Chapitre 3

Principes et vue d'ensemble du système ANTHAPSI

Le système d'analyse thématique et d'apprentissage de situations que nous présentons, ANTHAPSI est formé de deux grandes composantes développant des principes similaires mais à des niveaux différents. Notre objectif ici est de présenter les principes qui leur sont communs et de donner une vue d'ensemble de l'architecture du système, chaque chapitre venant à la suite étant chargé d'en détailler une partie spécifique.

1. Introduction

Le système ANTHAPSI (ANalyse THématique et APprentissage de SItuations), qui est la concrétisation du travail que nous avons mené, est une tentative pour marier l'analyse thématique et l'apprentissage de connaissances pragmatiques en s'appuyant sur la notion d'amorçage. Il comprend deux grandes composantes : MLK et ROSA. Celles-ci réalisent les mêmes tâches, en l'occurrence la mise en évidence des situations dans les textes et l'apprentissage d'une représentation de ces situations, mais à des niveaux différents.

MLK travaille à partir d'une représentation sémantique des propositions des textes et produit des structures organisant ces propositions en fonction des situations qu'elles évoquent. ROSA, quant à lui, opère à un niveau beaucoup moins élaboré puisqu'ilprend comme entrée le résultat d'un étiqueteur morpho-syntaxique et produit une représentation des situations sous forme de vecteurs de mots.

L'une des hypothèses que nous cherchons à valider au travers du système ANTHAPSI est qu'un module comme ROSA, certes plus frustre que MLK mais néanmoins fonctionnellement similaire, peut servir à amorcer un module tel que MLK, capable de mener des analyses plus élaborées et de construire des représentations beaucoup plus structurées, mais imposant également des exigences plus importantes concernant ses pré-requis.

Nous cherchons également à montrer, par le fait que ROSA est lui-même amorcé par un niveau plus élémentaire encore, que cette démarche peut être appliquée de façon récursive.

2. Architecture du système ANTHAPSI

Conformément à ce que nous avions indiqué dans l'introduction, le système ANTHAPSI se décompose en deux sous-systèmes, MLK et ROSA. Cette dichotomie, ainsi que la structuration interne de ces deux sous-systèmes, sont illustrées par la figure 3.1.

Cette figure laisse également apparaître que ROSA comprend elle-même deux composantes, assez fortement intriquées. La première, appelée SEGCOHLEX (SEGmentation thématique par utilisation de la COHésion LEXicale), recouvre un mécanisme de segmentation thématique autonome, c'est-à-dire ne dépendant que d'une source de connaissances construite une fois pour toutes. SEGCOHLEX constitue ainsi une amorce pour la seconde composante de ROSA, SEGAPSITH (SEGmentation et APprentissage de SIgnatures THématiques), et plus indirectement, pour l'ensemble du système aussi puisque SEGAPSITH sert à l'amorçage de MLK.

On remarquera d'ailleurs la similitude de l'interaction, d'une part entre SEGCOHLEX et SEGAPSITH, et d'autre part entre SEGAPSITH et MLK. Dans les deux cas, un processus de segmentation thématique vient apporter son concours au démarrage d'un autre processus de segmentation thématique, a priori plus élaboré, ou tout du moins reposant sur des connaissances plus structurées. La principale différence entre ces deux interactions concerne la représentation pré-thématique[1] des textes en entrée. SEGCOHLEX et SEGAPSITH s'appuient en effet sur la même représentation, en l'occurrence le résultat d'un étiqueteur morpho-syntaxique, tandis que MLK repose sur une représentation pré-thématique de beaucoup plus haut niveau que celle de SEGAPSITH.

2.1. ROSA

ROSA a été conçu en ayant à l'esprit le souci de la robustesse, même si celle-ci devait s'exprimer au détriment de la précision. L'objectif de ROSA est en effet de fournir une analyse thématique véritablement opérationnelle sur un vaste ensemble de textes, sans faire l'hypothèse de processus amont non directement utilisables.

[1] Il s'agit de la représentation des textes produite par les processus venant en amont de l'analyse thématique.

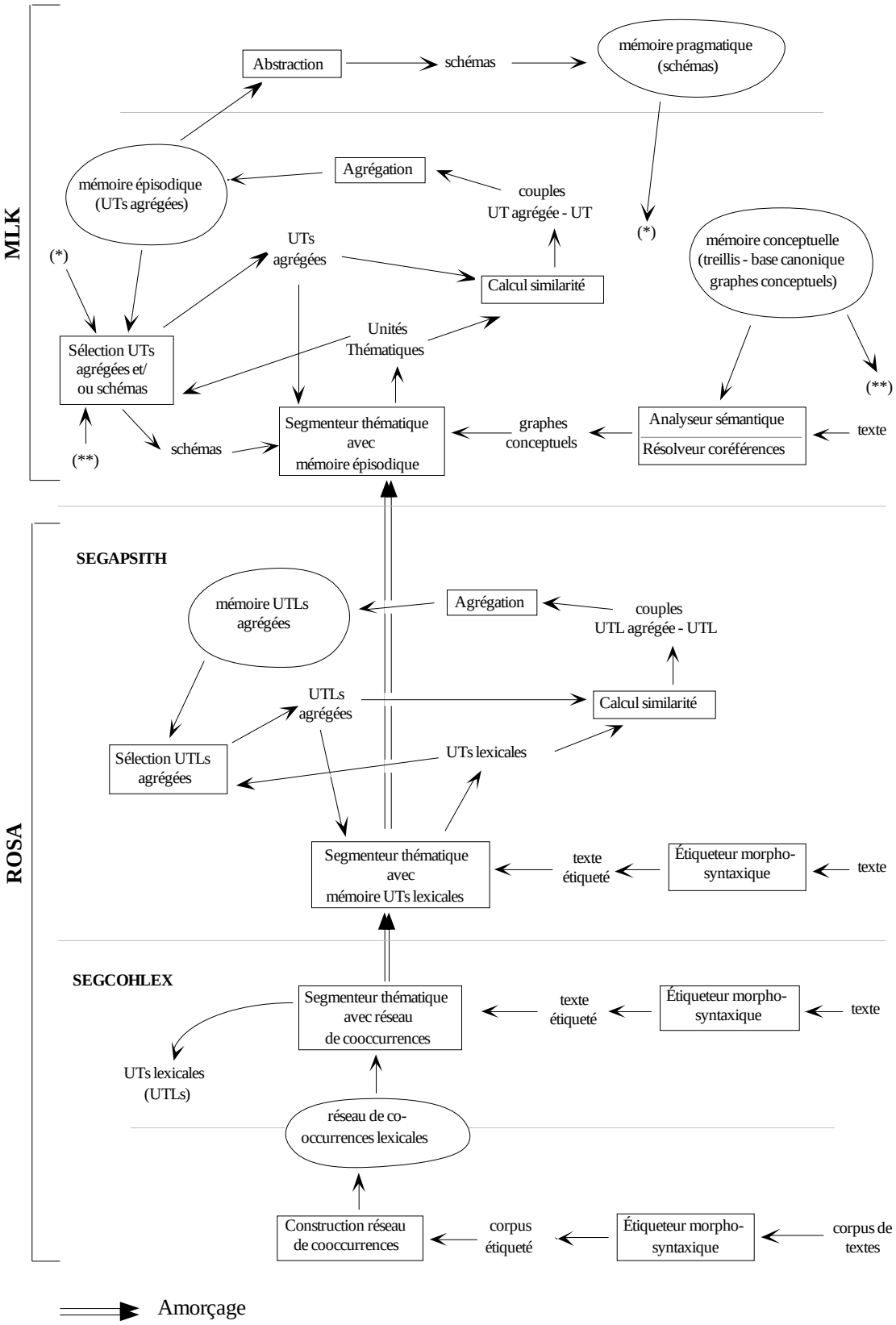


Fig. 3.1 - Architecture générale du système ANTHAPSI

Cette analyse thématique recouvre deux volets : d'une part la segmentation thématique, c'est-à-dire le découpage des textes en fonction des situations, ou plus

généralement des thèmes qu'ils abordent; d'autre part l'identification des thèmes, c'est-à-dire la capacité d'associer un segment de texte à la représentation du thème dont il est une expression. Cette tâche d'identification thématique s'accompagne ici d'une tâche de construction de la représentation des thèmes dans la mesure où celle-ci n'est pas supposée donnée a priori.

2.1.1. SEGCOHLEX

SEGCOHLEX constitue le socle de ROSA en fournissant un premier mécanisme de segmentation thématique reposant entièrement sur la notion de cohésion lexicale. La figure 3.2 montre que SEGCOHLEX comporte elle-même deux composantes. La première est celle réalisant la construction de la source de connaissances capturant la notion de cohésion lexicale, en l'occurrence un réseau de cooccurrences lexicales. La seconde est le processus de segmentation thématique proprement dit exploitant cette cohésion lexicale. Toutes deux se fondent sur les résultats d'un étiqueteur morpho-syntaxique permettant à la fois d'accéder à la forme canonique désambiguïsée des mots et de supprimer les mots dits outils, c'est-à-dire sans valeur du point de vue thématique.

Il est à noter que la relation entre ces deux composantes est purement séquentielle. La construction du réseau de cooccurrences s'effectue une fois pour toutes à partir d'un très vaste ensemble de textes. Lorsque ce réseau est utilisé par la seconde composante, le première n'est plus active. Ce choix est ici davantage un choix d'ordre pratique qu'il n'est un choix de principe. Rien n'empêcherait véritablement d'intégrer les cooccurrences des textes traités par ANTHAPSI au réseau initial. Une telle intégration irait même dans le sens du principe en vertu duquel différents textes peuvent être traités avec des connaissances de niveaux différents suivant qu'un niveau possède ou ne possède pas une représentation des domaines abordés par ces textes. Néanmoins, il existe également une différence d'échelle suffisamment importante entre la masse de textes nécessaire pour constituer un réseau de cooccurrences (plusieurs millions de mots) et le nombre de textes assimilable par un système comme ANTHAPSI pour que cette liaison ne soit pas intéressante.

2.1.2. SEGAPSITH

SEGAPSITH est la partie de ROSA représentant l'analogue de MLK. Elle réalise à la fois une analyse thématique complète au sens où nous l'avons définie précédemment, avec segmentation thématique et identification des thèmes, et elle construit par apprentissage une représentation de ces thèmes.

SEGAPSITH reste néanmoins au niveau des mots. Les segments de texte distingués, appelés Unités Thématiques Lexicales (UTLs), sont des listes de mots. Les thèmes sont également représentés par des ensembles de mots, dotés chacun d'un poids caractérisant son degré d'importance par rapport aux autres mots. Ils sont formés par l'agrégation de plusieurs UTLs jugées similaires, en vertu des mécanismes d'apprentissage exposé précédemment. Un thème s'identifie donc à une UTL agrégée. Celle-ci est aussi appelée *signature thématique*, en référence à la notion de "concept signature" développée dans [Hovy & Lin 1997].

Précisons qu'à l'instar de MLK, le but est ici de former une représentation des situations prototypiques du monde de référence. La granularité thématique retenue est donc assez faible. Il ne s'agit pas de caractériser des thèmes larges tels que la bourse, le football ou le chômage. On essaie plutôt de cerner le vocabulaire caractérisant des événements plus circonscrits tel qu'un krach boursier, une grève des transports ou la sortie d'un nouveau film.

La capacité pleinement opérationnelle de SEGAPSITH, due au niveau où il intervient, conjuguée à sa proximité avec MLK sur le plan des principes lui confère par ailleurs un rôle spécifique de terrain de validation et d'expérimentation des principes sous-tendant chacun des grands niveaux d'ANTHAPSI. Son association avec SEGCOHLEX lui donne un rôle similaire vis-à-vis de l'amorçage entre ces niveaux.

2.2. MLK

MLK est le niveau haut du système ANTHAPSI. Il réalise globalement les mêmes tâches que le module ROSA mais, au contraire de celui-ci, met l'accent sur la précision de l'analyse thématique et la structuration des représentations des situations. Ces exigences imposent naturellement une analyse pré-thématique plus sophistiquée. Cette dernière produit en l'occurrence une représentation sémantique des propositions des textes sous forme de graphes conceptuels [Sowa 1984].

La brique élémentaire de représentation des situations passe donc des simples mots aux assemblages de concepts, équivalents à des assertions de la logique des prédicats. Les UTs de MLK sont ainsi des ensembles de graphes conceptuels et les UTs agrégées qui sont contenues dans sa *mémoire épisodique* sont assimilables à des ensembles de graphes conceptuels pondérés suivant les mêmes principes que les mots dans les UTLs agrégées.

En dehors de ce changement des constituants élémentaires de représentation, l'analyse thématique de MLK se veut plus précise que celle de ROSA dans la mesure où les UTs qu'elle forme ne sont plus nécessairement des séquences de phrases extraites directement du texte mais peuvent être le rassemblement de propositions plus ou moins dispersées dans celui-ci. Notre volonté est même sur ce point de nous affranchir le plus possible d'un modèle a priori de la structuration du discours afin de pouvoir traiter les textes au style très entrelacé.

La dernière dimension de MLK à aborder au cours de ce survol lui est proprement spécifique puisque n'existant pas du tout au niveau de ROSA. Le stade d'évolution caractérisant ce module est en effet suffisant pour envisager l'abstraction des UTs agrégées formées au sein de la mémoire épisodique. Cette opération a pour but de construire des schémas. Ceux-ci sont proches de la notion de MOP que nous avons présentée au chapitre 1 et incarnent la connaissance pragmatique stable du système concernant le contenu des situations prototypiques du monde de référence. Ces schémas peuvent être ponctuellement utilisés par le processus d'analyse thématique de MLK au même titre que les UTs agrégées ou bien, s'ils couvrent entièrement les domaines abordés par un texte, être le support d'un mécanisme spécifique tel que celui présenté dans [Grau 1983].

3. Principes

3.1. Principes généraux de l'amorçage

3.1.1. Deux types d'amorçage

L'amorçage apparaît ici sous deux formes. Au sein même de chaque niveau de développement, incarnés dans le cas présent par ROSA et MLK, la conjugaison d'un processus d'analyse de textes et d'un processus d'apprentissage permet de construire des connaissances qui sont ensuite exploitées par ce même processus d'analyse afin de traiter de nouveaux textes de manière plus avancée. On part ainsi d'un noyau de connaissances peu élaborées, représentant l'amorce du système et lui conférant des capacités d'analyse minimales, puis on affine et on étend progressivement ces connaissances par le biais des informations apportées par les textes. Le point central de cet amorçage est donc l'aptitude du processus d'analyse des textes à prendre en compte les connaissances qui sont apprises afin d'améliorer sa capacité à mettre en évidence les informations présentes dans les textes.

Ce mécanisme se rapproche en fait du mode de fonctionnement des systèmes de raisonnement à base de cas intégrant au moins la forme minimale d'apprentissage que

constitue l'intégration automatique des nouveaux cas créés. Ces systèmes disposent d'une bibliothèque initiale de cas qu'ils utilisent afin de résoudre des problèmes qui leur sont soumis. Ils enregistrent ensuite une représentation de ces problèmes et de la solution qu'ils leur ont apporté et forment ainsi de nouveaux cas. Ceux-ci peuvent à leur tour être exploités ultérieurement afin de résoudre de nouveaux problèmes. Précisons néanmoins que cette démarche débouche plus fréquemment sur une spécialisation que sur un amorçage véritable.

Le second type d'amorçage intervient entre un niveau de développement et celui qui le suit. Il est directement en relation avec la première forme d'amorçage puisqu'il vise à résoudre le problème important de la constitution du noyau initial de connaissances qui est à l'origine de l'amorçage intra-niveau. L'idée qui prévaut ici est qu'il est possible de discrétiser en différents niveaux la spirale de progression du développement d'un système de façon à ce que l'activité de chacun de ces niveaux contribue à la formation des premières connaissances du niveau qui le suit.

Dans le cas qui nous occupe, ROSA, en étant capable de découper des textes en segments thématiquement cohérents, fournit à MLK une première capacité d'analyse lui permettant de constituer un ensemble de représentations de situations. Celles-ci sont ensuite exploitées par MLK afin de mener une analyse thématique plus fine, notamment parce que sachant mettre à jour des segments discontinus. Le traitement de nouveaux textes avec ces capacités accrues ouvre alors la voie au premier type d'amorçage.

3.1.2. Un amorçage en trois phases

La différenciation de deux formes d'amorçage représente en fait la discrétisation d'un processus plus continu de passage de relais d'un niveau N au niveau N+1 qui le suit, comme entre SEGCOHLEX et SEGAPSITH ou entre SEGAPSITH et MLK. Trois phases peuvent y être distinguées :

- 1ère phase : le niveau N+1 puise une part significative de ses capacités d'analyse au sein du niveau N dans la mesure où il ne dispose pas encore des connaissances à son niveau pour alimenter son propre mécanisme d'analyse. En revanche, le processus d'apprentissage du niveau N+1 est déjà à l'œuvre sur les premières représentations de texte construites afin de produire ces connaissances initialement inexistantes. C'est lors de cette phase que l'amorçage inter-niveau s'exprime de la façon la plus vive;
- 2ème phase : c'est une phase de fonctionnement mixte dans laquelle des connaissances du niveau N+1 sont utilisables mais ne suffisent pas nécessairement pour mener l'analyse complète d'un texte, soit parce que des domaines n'y sont pas

encore représentés, soit parce que la représentation de certaines situations est encore incomplète. Il y a donc coopération entre l'analyse du niveau N et celle du niveau N+1 afin de produire la représentation thématique de nouveaux textes. À mesure que le processus d'apprentissage du niveau N+1 exploite ces représentations pour compléter des connaissances partielles ou bien construire celles encore inexistantes, un transfert progressif s'opère depuis l'analyse du niveau N vers celui du niveau N+1. Ce transfert trouve son aboutissement dans la troisième phase;

- 3ème phase : dans cette phase, les ressources du niveau N+1 sont suffisantes pour lui assurer une autonomie de fonctionnement vis-à-vis du niveau N. On se trouve alors pleinement dans le cas d'un système qui raffine et étend progressivement ses propres connaissances au contact des informations tirées des textes.

Il faut ajouter que la transition d'une phase à l'autre ne se fait pas de manière homogène pour l'ensemble du système. Différents domaines de connaissances peuvent donc se trouver à des stades différents dans le passage d'un niveau à un autre. Dans le cas d'un système distinguant plus de deux niveaux, on peut même imaginer que celui-ci fonctionne suivant les méthodes d'un niveau N pour un domaine de connaissances et suivant les méthodes par exemple d'un niveau N+2 pour un autre domaine. Cela dépend en pratique de l'histoire du système en question, donc des textes qui lui ont été soumis.

Ce dernier aspect souligne le fait que l'utilisation d'un apprentissage progressif et d'un amorçage induit une dépendance forte vis-à-vis de l'ordre dans lequel les textes sont traités. Il met en relief à cette occasion l'importance de la façon dont le système est éduqué : lui présente-t-on successivement des textes appartenant à des domaines différents ou au contraire tout un ensemble de textes relatifs au même domaine?. Commence-t-on par de petits textes uniformes sur le plan thématique ou des textes plus longs et aussi plus diversifiés quant aux sujets abordés? Les stratégies retenues concernant ce point ont à l'évidence une influence notable sur la nature des connaissances disponibles et leur évolution, ce qui affecte par voie de conséquence les capacités de l'intégralité du système.

Le décomposition de l'amorçage effectuée ci-dessus fait ressortir par ailleurs une propriété supplémentaire que le processus d'analyse des textes doit posséder pour opérer dans le cadre défini ici. La première phase et la deuxième phase impliquent en effet que ce processus puisse intégrer les résultats d'un autre processus d'analyse, ayant globalement des objectifs similaires mais agissant à un niveau plus élémentaire, afin de travailler dans un mode que l'on peut qualifier de dégradé. C'est ce qui se passe par exemple lorsque le

segmenteur thématique de SEGAPSITH fait appel à celui de SEGCOHLEX ou que l'analyse thématique de MLK utilise les services du segmenteur de SEGAPSITH.

Cette méthode de collaboration est à séparer du cas le plus général dans lequel un processus exploite le résultat d'un second mais traitant d'un autre problème. Que celui-ci se situe plutôt en aval, en amont ou sur le même plan que le premier, il intervient généralement à un niveau similaire, avec des représentations compatibles. Lorsque l'on utilise par exemple une représentation sémantique sous forme de graphes conceptuels, on ne fait pas appel aux connaissances pragmatiques qui sont capturées par un réseau de cooccurrences lexicales. L'amorçage oblige en revanche à se pencher sur ce type d'interactions.

3.2. Principes de l'amorçage intra-niveau

Dans ce chapitre, nous présentons globalement les principes communs à ROSA et MLK régissant l'amorçage intra-niveau, lequel représente la part la plus importante de notre travail. L'amorçage inter-niveau repose quant à lui sur des mécanismes plus spécifiques aux processus d'analyse de texte de ROSA et MLK et seront exposés après la description de ceux-ci.

3.2.1. Vue d'ensemble et analyse des textes

La figure 3.2 donne une vue d'ensemble de la façon dont l'analyse de texte et l'apprentissage s'organisent au sein d'un niveau. Ainsi que nous l'avions esquissé au chapitre 1, le point central de cet édifice est une mémoire. Celle-ci fournit les connaissances nécessaires à l'analyse des textes en même temps qu'elle s'alimente de ses résultats, après passage de ceux-ci par une phase d'apprentissage.

Pour être plus précis, l'analyse des textes comporte une première étape, qualifiée ici d'*analyse pré-thématique*, visant à passer des textes sous forme de chaînes de caractères à une première représentation au sein de laquelle les informations nécessaires à l'analyse thématique ont été mis à jour. Bien entendu, ces informations dépendent du niveau auquel on se situe. Pour ROSA, la détermination de la catégorie grammaticale des mots et de leur forme canonique suffit. Pour MLK en revanche, les pré-requis sont beaucoup plus forts puisque cette analyse pré-thématique correspond à une analyse sémantique des textes.

Le premier processus véritablement spécifique de notre approche est celui de *segmentation thématique* des textes. Il a pour objectif de construire des Unités Thématiques (UTs). Celles-ci rassemblent tous les éléments du texte, parmi ceux produits par l'analyse pré-thématique, relatifs à une même situation évoquée par le texte. Une UT

correspond donc à la représentation qui est donnée d'une situation au travers d'un texte. La segmentation thématique produit en finale autant d'UTs qu'il y a de situations abordées de façon significative par un texte.

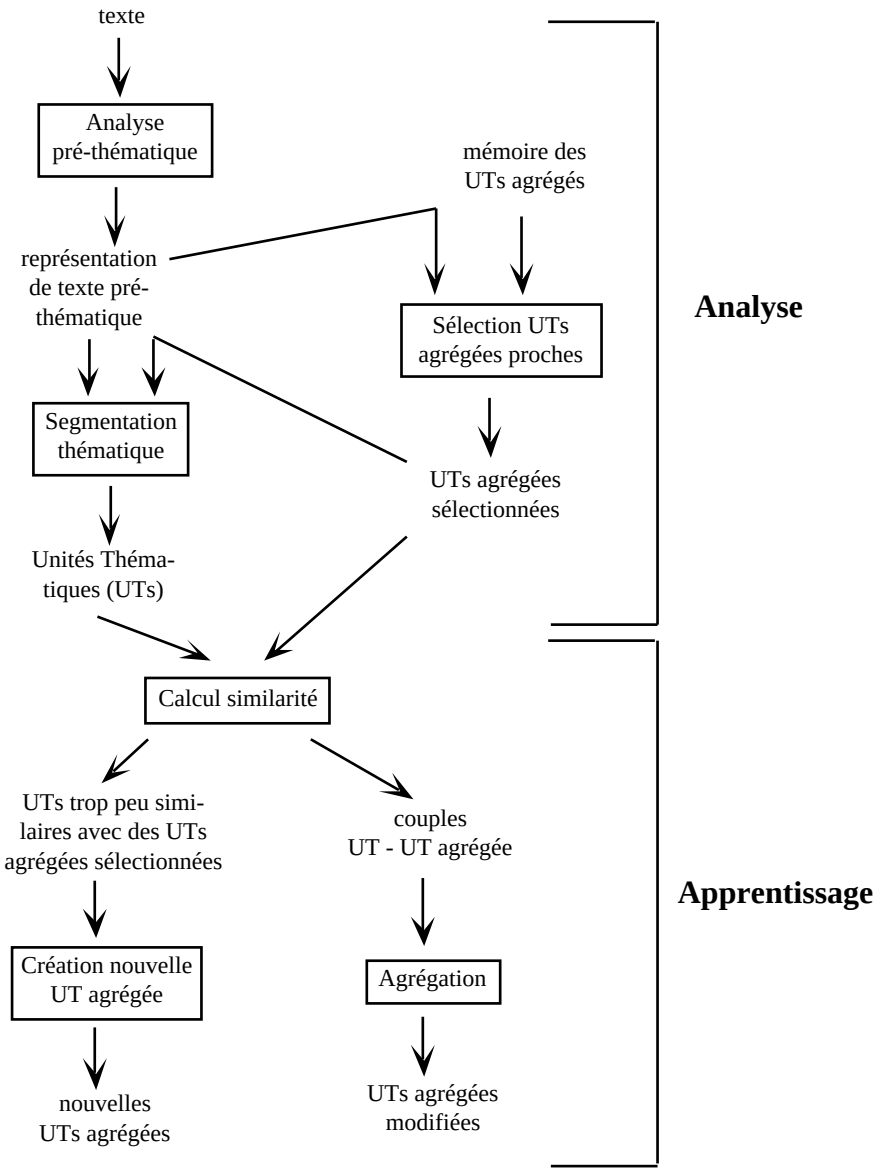


Fig. 3.2 - Décomposition fonctionnelle de l'organisation intra-niveau

La figure 3.2 montre que cette analyse des textes s'appuie sur le contenu de la mémoire. Celle-ci est de ce fait dotée d'un mécanisme de *sélection de connaissances* permettant de récupérer en son sein les représentations des situations apprises en fonction du contexte courant d'analyse. L'importance que revêt ici la notion d'apprentissage incrémental et l'absence d'un cadre fixé a priori imposent à ce mécanisme de s'adapter continûment à l'évolution de la mémoire, ce qui le distingue des réseaux d'index statiques généralement utilisés pour cette tâche.

À l'issue de la segmentation thématique, on dispose donc à la fois des UTs mises en évidence dans les textes et des connaissances de la mémoire qui sont présumées les plus proches de ces UTs.

3.2.2. L'apprentissage des situations

Les principes

Les UTs construites à partir des textes constituent le matériau de base qui sert à l'élaboration d'une représentation générale des situations prototypiques du monde de référence. Bien entendu, l'image d'une situation donnée par un seul texte ne suffit pas à produire une telle représentation. Elle est à la fois trop spécifique et incomplète. La présence en son sein de certains éléments n'a pas de lien nécessaire avec la situation tandis qu'au contraire, d'autres ne sont pas évoqués explicitement car le rédacteur se repose sur la capacité du lecteur à les retrouver à partir de ses connaissances et des indices qu'il lui fournit. D'autres éléments enfin sont présents sous une forme trop particulière et doivent être généralisés.

L'idée consiste donc, pour construire une représentation générale des situations, à confronter celles qui sont construites à partir de différents textes. On s'appuie pour cela sur deux hypothèses. La première, relative au problème de la spécificité, énonce que les éléments non nécessairement liés à une situation ne se répètent pas de façon systématique au travers de différents textes. En vertu de cette hypothèse, on considère que les éléments caractéristiques d'une situation sont à l'inverse ceux qui apparaissent de façon récurrente dans l'évocation de cette situation.

Ces éléments se manifestent en pratique sous des formes diverses, plus ou moins générales. Si l'on s'attache ainsi à la représentation de la situation *Emménager*, le lieu de l'emménagement peut être un appartement, plus précisément un studio, un loft ou un duplex par exemple, une maison, un château ou bien encore un lieu d'habitation sans type particulier, comme c'est le cas lorsque le texte n'indique que la ville. Cette variété dans la manifestation des éléments composant une situation permet de saisir quel niveau de généralité chacun d'entre eux ne doit pas dépasser pour avoir encore une signification du point de vue de la situation. Elle permet également de s'affranchir de la trop grande spécificité qui résulterait de la considération d'un seul texte.

La seconde hypothèse est quant à elle un outil face au problème de l'incomplétude de la représentation des situations. Elle affirme que différents textes évoquant une même situation ne le font pas de la même manière. Il existe certes des points communs entre ces évocations, ce qui permet de les rapprocher, mais elles présentent également une

variabilité assez importante. Telle dimension de la situation va être détaillée par un rédacteur tandis qu'elle ne sera qu'effleurer par un autre. Ces différences ont des sources multiples parmi lesquelles on peut citer le type des textes, une dépêche d'agence de presse n'obéit ni aux mêmes contraintes, ni aux mêmes objectifs qu'un passage de roman, l'importance de la situation vis-à-vis du propos général ou bien encore le style propre du rédacteur. Toutes ces différences sont néanmoins instructives dans la mesure où leur cumul permet de construire une représentation assez complète d'une situation.

Les deux hypothèses présentées ci-dessus sous-tendent le mécanisme choisi pour réaliser la confrontation et en définitive, la fusion des différentes évocations d'une situation. Ainsi que nous l'avons esquissé lors du chapitre 1, nous pensons que l'agrégation de ces différentes représentations, dans la mesure où elles ont été reconnues similaires, offre la possibilité de faire émerger une description générale des situations. Pour être plus précis, l'évocation d'une situation dans un texte prend ici la forme d'une UT et le résultat de leur agrégation est appelé UT agrégée.

Si l'on transfère les idées de Vygotski du domaine conceptuel au domaine pragmatique, on peut assimiler la description émergente des situations au stade des concepts, caractérisés par une définition en intention. Ces concepts sont le produit de l'abstraction des complexes, fonctionnellement similaires aux concepts, mais, contrairement à eux, définis en extension comme le regroupement d'un ensemble d'expériences jugées similaires. Les UTs agrégées constituent l'équivalent de ces complexes au niveau pragmatique. Pour nous situer par rapport à la terminologie employée au chapitre 1, une UT est vue comme une expérience et une UT agrégée représente un agrégat d'expériences.

Une UT agrégée est donc construite en suivant les principes d'un agrégat d'expériences : on rassemble les éléments communs des UTs dans des structures uniques et l'on pondère ces structures en fonction de la récurrence des éléments qui les composent. En s'appuyant sur ces poids, on obtient une forme de généralisation implicite par la capacité qu'ils confèrent à différencier l'importance relative des éléments des situations.

Les mécanismes de l'apprentissage

La mémoire qui sert de support à l'analyse des textes est donc formée d'un ensemble d'UTs agrégées. À l'issue de la segmentation thématique d'un texte, on dispose de ce fait d'une ou plusieurs UTs, selon le nombre de situations évoquées par le texte, et d'un ensemble d'UTs agrégées, obtenues par le mécanisme de sélection de connaissances évoqué plus haut et supposées proches des UTs du texte. Il faut préciser que l'on

conserve le lien entre une UT et les UTs agrégées sélectionnées qui ont contribué à sa création.

La première phase de l'intégration en mémoire des nouvelles UTs consiste à déterminer pour chacune d'entre elles si elle peut être agrégée à l'une des UTs agrégées qui lui sont liées. Pour cela, on s'appuie sur une mesure de similarité entre une UT et une UT agrégée. Sur le principe, cette mesure cherche à évaluer le rapport entre le nombre, combiné à leur importance, des éléments communs à l'UT et à l'UT agrégée sur le nombre et l'importance de l'ensemble des éléments de chacune de ces deux entités. On obtient par conséquent deux rapports qui sont en finale combinés afin de fournir une évaluation globale de la similarité entre l'UT et l'UT agrégée.

On remarquera que cette mesure, contrairement à la forme générale des mesures inspirées de [Tversky 1977], ne fait pas intervenir les constituants différents, pas plus ceux de l'UT que ceux de l'UT agrégée. Nous reviendrons plus précisément sur ce point lors de l'exposé détaillé de chacun des niveaux. Toutefois, la justification globale que l'on peut dès à présent avancer est que du fait de notre approche, la présence d'éléments différents n'est pas forcément significative d'une dissimilarité des entités comparées dans la mesure où une UT, a fortiori une UT agrégée, est une entité assez bruitée. Les éléments qui forment l'essence de la situation représentée par une UT se retrouvent en effet fréquemment noyés dans un ensemble d'éléments beaucoup plus contingents exprimant les circonstances particulières d'une occurrence de la situation.

Une mesure de similarité présentant les caractéristiques pré-citées est donc appliquée entre chacune des UTs construites pour un texte et les UTs agrégées qui lui sont associées. Si la valeur de cette mesure dépasse un seuil fixé pour l'une au moins de ces UTs agrégées, l'UT issue du texte est agrégée avec l'UT de la mémoire pour laquelle la similarité est la plus forte. Dans le cas contraire, elle est mémorisée comme une nouvelle UT agrégée. L'opération d'agrégation consiste simplement à augmenter la valeur de récurrence des éléments communs et à ajouter les éléments nouveaux apportés par l'UT venant du texte. Le poids des éléments d'une UT agrégée étant donné par le rapport entre leur valeur de récurrence et le nombre d'UTs dont elle est constituée, le poids des éléments communs se trouve naturellement renforcé tandis que, par le simple effet de l'accroissement du nombre d'UTs regroupées, le poids des autres éléments diminue.

4. Les limites a priori du système ANTHAPSI

Ainsi que nous l'avons développé au §3 de ce chapitre, un des points que nous souhaitions illustrer au travers du système ANTHAPSI est la possibilité pour un

processus manipulant des connaissances caractérisées par un degré d'élaboration donné de contribuer au démarrage d'un processus ayant globalement le même objectif mais travaillant à partir de connaissances plus élaborées.

Cette démarche est mise en œuvre d'une part entre les composantes SEGCOHLEX et SEGAPSITH de ROSA, et d'autre part entre SEGAPSITH et MLK. La situation est néanmoins assez différente d'un cas à l'autre. En passant de SEGCOHLEX à SEGAPSITH, on ne change pas en effet la nature des unités de représentation élémentaires. Ce sont toujours des mots et d'ailleurs, la représentation pré-thématique des textes reste la même. La différence provient uniquement d'une évolution dans la structuration des connaissances manipulées.

Dans SEGCOHLEX, on utilise un réseau de cooccurrences lexicales, connaissance que l'on peut qualifier de très peu structurée du point de vue de la représentation des situations. Les relations entre mots n'y sont en effet pas uniquement le produit de la description d'une même situation mais rendent compte également d'aspects syntaxiques ou sémantiques. En revanche, le mécanisme d'analyse de SEGAPSITH s'appuie sur des UTLs agrégées, qui, même si elles restent des configurations de mots, sont spécifiques de la représentation des situations. Du fait de la relation entre SEGCOHLEX et SEGAPSITH, on peut considérer en fait que les UTLs agrégées structurent le réseau de cooccurrences sur le plan thématique.

Entre SEGAPSITH et MLK au contraire, le changement des représentations manipulées est beaucoup plus radical puisqu'il ne touche pas seulement la façon dont leurs constituants sont structurés mais également la nature de ceux-ci. On saute ainsi des mots aux concepts, et même plus précisément de mots isolés à des graphes conceptuels, qui sont des groupements structurés de concepts. Cette différence au niveau des entités manipulées par l'analyse thématique se traduit de fait par une différence quant à la nature de l'analyse pré-thématique requise. Alors que SEGAPSITH se contente d'un simple étiqueteur morpho-syntaxique, MLK nécessite le recours à une analyse syntaxico-sémantique, couplée à une résolution des co-références.

Ce saut quant à la nature des représentations employées ainsi que des processus chargés de les produire est une des limites du système ANTHAPSI dans son état actuel. Un amorçage est effectif entre SEGAPSITH et MLK sur le plan de l'analyse thématique mais il n'existe pas en revanche en ce qui concerne les pré-requis de cette analyse. On peut donc affirmer que l'amorçage entre SEGAPSITH et MLK est pour le moment un raccourci davantage destiné à illustrer une démarche générale qu'à constituer une référence sur laquelle il n'y aura pas lieu de revenir.

ROSA est représentatif d'un niveau opérationnel pouvant servir de point de départ à un amorçage tel que nous le concevons. MLK spécifie en revanche la forme que pourrait revêtir un certain point d'aboutissement de cet amorçage. Entre ces deux bornes, il reste encore à définir toute une gradation permettant de passer progressivement de l'un à l'autre. Cette définition passe sans aucun doute par la prise en compte de dimensions supplémentaires de l'analyse de textes. C'est toutefois un point que nous avons volontairement laissé de côté pour le moment devant la nécessité de fixer une base de travail suffisamment stable, même au prix de certaines hypothèses simplificatrices.

Récapitulatif

Notre première tâche dans ce chapitre a été de présenter plus précisément l'architecture générale d'ANTHAPSI, le système que nous proposons ici. Celle-ci laisse apparaître deux composantes principales : ROSA et MLK. ROSA est dirigé par le souci d'être opérationnel sur une large échelle. Il s'appuie donc sur des pré-requis peu exigeants, en l'occurrence un étiqueteur morpho-syntaxique, mais produit des représentations peu sophistiquées. L'analyse thématique ne fait que délimiter des blocs de texte contigus et l'apprentissage produit une représentation des situations sous forme d'ensembles de mots pondérés. ROSA est lui-même formé de deux modules. SEGCOHLEX réalise une analyse thématique fondée sur un réseau de cooccurrences lexicales servant à l'amorçage de SEGAPSITH, qui assume à proprement parler les fonctions de ROSA.

MLK, quant à elle, manipule des connaissances beaucoup plus élaborées puisqu'il s'agit de graphes conceptuels. Elle peut ainsi mettre en œuvre une analyse thématique plus fine, s'abstrayant de la linéarité des textes, et construire des représentations des situations plus structurées sous forme d'ensembles de graphes conceptuels pondérés.

Le second volet de ce chapitre s'est attaché pour sa part à présenter les principes sur lesquels repose ANTHAPSI. Le plus important d'entre eux soutient qu'un processus d'analyse thématique et d'apprentissage de connaissances sur les situations peut être mis en œuvre en étant amorcé par un processus ayant les mêmes objectifs mais s'appuyant sur des connaissances moins élaborées. Un examen plus détaillé de ce principe nous a conduit à différencier deux types d'amorçage étroitement intriqués : l'amorçage prenant place au sein d'un même niveau de connaissances et celui intervenant entre un niveau de connaissances et un niveau de connaissances plus élaborées.

La totalité du déroulement d'un amorçage a quant à elle été divisée en trois phases. Dans la première phase, le processus d'analyse du niveau N s'appuie totalement sur celui du niveau N-1. Dans la seconde, son fonctionnement est mixte. Il ne peut en effet agir de

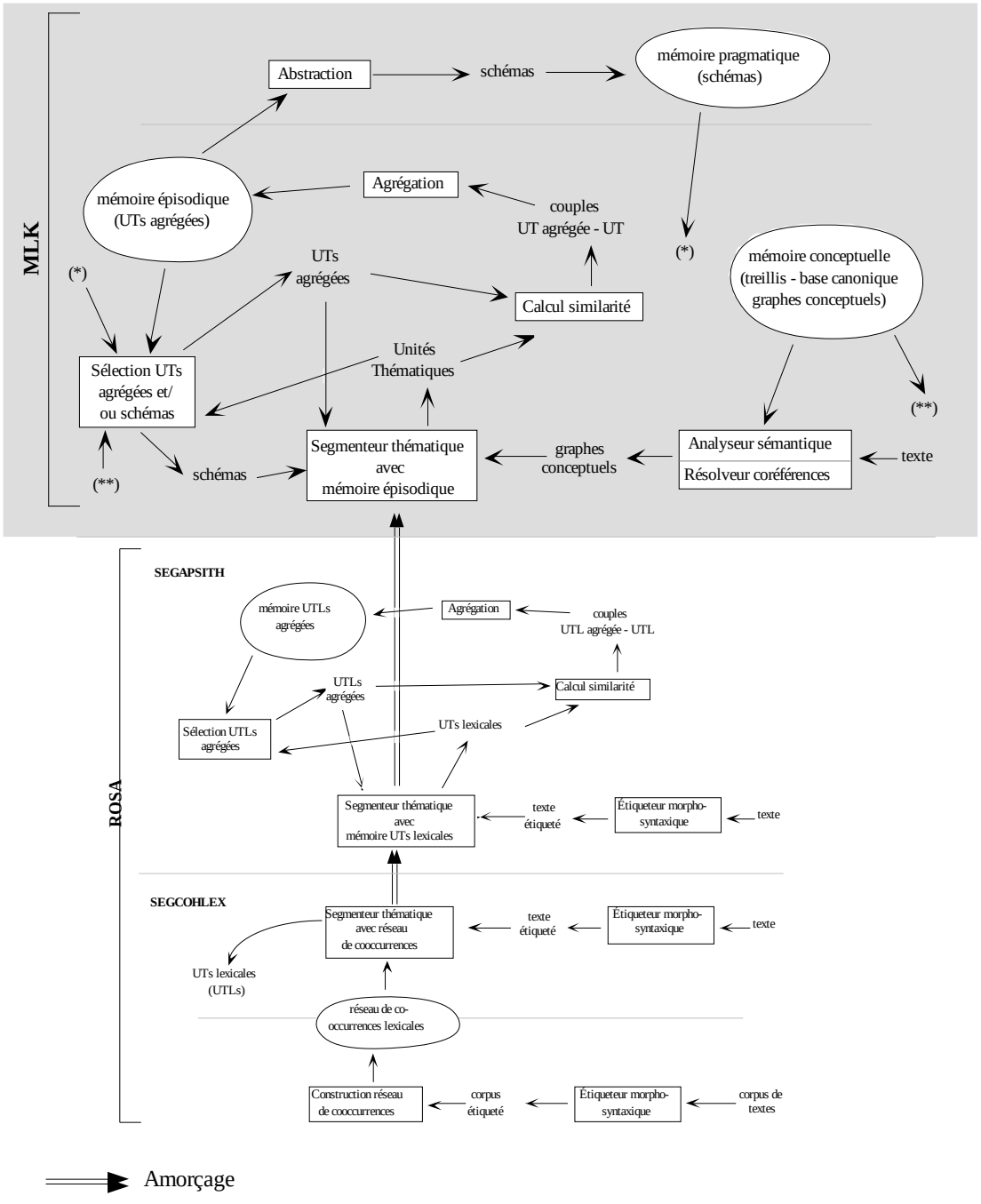
façon autonome que lorsque les connaissances de son niveau existe pour la situation évoquée. Dans la dernière enfin, l'intégralité d'un texte peut être traitée par le processus d'analyse du niveau N.

L'amorçage le plus étudié ici est l'amorçage intra-niveau. Il repose sur la notion de mémoire. L'analyse thématique produit des représentations des situations, appelées Unités Thématiques (UTs), qui sont ensuite agrégées en fonction de leur similarité pour former des agrégats (UTs agrégées), stockés en mémoire. Ces agrégats sont à leur tour utilisés par le processus d'analyse thématique afin de traiter de nouveaux textes.

Nous avons souligné enfin que l'amorçage de MLK par ROSA n'est que partiel dans la mesure où il ne touche que la dimension thématique. Le passage d'un étiquetage morpho-syntaxique, exploité par ROSA, à une analyse sémantique des textes, pré-requis indispensable pour fournir à MLK le niveau de connaissances qu'il sait traiter, a volontairement été laissé de côté ici du fait de l'impossibilité pratique à modéliser tous les niveaux intermédiaires.

Partie II

MLK


MLK
Abstraction
schémas
mémoire pragmatique
(schémas)
mémoire épisodique
(UTs agrégées)
Agrégation
couples
UT agrégée - UT
(*)
UTs
agrégées
Calcul similarité
mémoire conceptuelle
(treillis - base canonique
graphes conceptuels)
Unités
Thématiques
Sélection UTs
agrégées et/
ou schémas
(**)
schémas
Segmenteur thématique
avec
mémoire épisodique
graphes
conceptuels
Analyseur sémantique
Résolveur coréférences
texte
ROSA
SEGAPSITH
mémoire UTLs
agrégées
Agrégation
couples
UTL agrégée - UTL
UTLs
agrégées
Calcul similarité
Sélection UTLs
agrégées
UTs lexicales
Segmenteur thématique
avec
mémoire UTs lexicales
texte
étiqueté
Étiqueteur morpho-
syntaxique
texte
SEGCOHLEX
Segmenteur thématique
avec réseau
de cooccurrences
texte
étiqueté
Étiqueteur morpho-
syntaxique
texte
UTs lexicales
(UTLs)
réseau de co-
occurrences lexicales
Construction réseau
de cooccurrences
corpus
étiqueté
Étiqueteur morpho-
syntaxique
corpus de
textes
Amorçage

MLK

Nous commençons la présentation détaillée du système ANTHAPSI par celle de MLK, à la fois sa composante de plus haut niveau (cf. figure ci-contre) et sa composante de référence. MLK travaille à partir d'une représentation sémantique des textes et produit en final des schémas représentant des connaissances pragmatiques stables. Il met en œuvre les principes présentés dans les chapitres 1 et 3 en les appliquant dans un contexte de représentations très structurées.

Nous présenterons d'abord les connaissances sémantiques sur lesquelles MLK s'appuie ainsi que les connaissances pragmatiques sous forme de schémas qu'il produit en final (chapitre 4). Cet exposé sera suivi de la description de la représentation des textes manipulée à ce niveau (chapitre 5). La dimension apprentissage de MLK sera ensuite abordée en détail : en premier lieu au travers de la mémoire épisodique assurant la mémorisation des représentations de texte et l'émergence de connaissances pragmatiques sous la forme d'agrégats (chapitre 6); puis au travers de l'abstraction de ces agrégats en schémas stables (chapitre 7). La présentation de MLK se terminera par celle de son analyse thématique, productrice des représentations de texte et utilisatrice des connaissances construites à partir de ces représentations (chapitre 8).

Chapitre 4

Mémoires conceptuelle et pragmatique

Nous commençons ici la description de MLK par celle des deux formes de connaissances qui l'encadrent : les connaissances sémantiques en amont, qui constituent le pré-requis à son activité, et les connaissances pragmatiques en aval, qui en sont le produit final. La mémoire conceptuelle, qui rassemble les connaissances sémantiques déclaratives d'ANTHAPSI, s'appuie sur le formalisme des graphes conceptuels dont nous donnons une description générale tout en insistant sur les points plus particulièrement importants vis-à-vis de notre travail. La mémoire pragmatique, chargée d'abriter la représentation déclarative des situations prototypiques du monde de référence, utilise un formalisme de schémas spécifique que nous présentons en détail.

1. Mémoire conceptuelle

1.1. Rôle de la mémoire conceptuelle

Fondamentalement, le rôle de la mémoire conceptuelle est d'abriter les connaissances sémantiques déclaratives de MLK, qui sont plus globalement celles d'ANTHAPSI également. Ces connaissances sémantiques ont une double vocation. Tout d'abord, elles définissent les unités de représentation conceptuelle disponibles, les concepts, ainsi que les relations conceptuelles qui sont utilisables pour les assembler afin de former des représentations plus complexes.

Nous ne chercherons pas ici à définir la notion de concept autrement qu'en faisant référence à la conception qui en est développée dans la plupart des travaux utilisant les réseaux sémantiques [Brachman 1979] [Kayser 1987]. Nous nous contenterons de remarquer que dans le domaine du traitement automatique des langues, la notion de concept s'identifie bien souvent à celle de signifié de mot. Cela n'est pas étranger au fait que les concepts et les relations conceptuelles y sont notamment utilisées dans le but de construire une représentation sémantique des propositions des textes ou des dialogues.

La seconde vocation des connaissances sémantiques est de définir la façon dont chaque concept peut être utilisé dans le cadre d'une assertion. À ce titre, elles forment un fond commun de connaissances permettant de faire la part entre des assertions qui dérivent de

ce fond commun et qui font donc sens par rapport à lui, et celles qui n'en sont pas dérivables et qui sont donc considérées sans signification. Ainsi la proposition

la brique regarde la paresse

ne peut être représentée sémantiquement de façon directe si les connaissances sémantiques existantes spécifient que l'agent du concept *Regarder* doit être un *Être_vivant* et son objet, un *Objet_physique*, sachant que le concept *Brique* ne se définit pas comme une sorte de *Être_vivant*, pas plus que le concept *Paresse* ne se définit comme une sorte de *Objet_physique*. Bien entendu, il est toujours possible de trouver un contexte dans lequel cette proposition peut être interprétée mais cela nécessite l'intervention préalable de processus d'interprétation extérieurs à une analyse sémantique par filtrage. Ces processus sont en l'occurrence capables de détecter que les mots brique et paresse ne font pas référence aux concepts *Brique*[1] et *Paresse*[2] ainsi que de retrouver les concepts qui sont réellement désignés par ces mots dans le contexte considéré.

Une analyse sémantique "classique" se contente en revanche d'essayer de construire, pour une proposition donnée, une assertion qui dérive strictement des connaissances sémantiques existantes et qui en représente le sens par rapport à ces connaissances. Pour cela, elle doit désambiguïser les mots de la proposition, c'est-à-dire choisir dans la liste des concepts possibles pour caractériser les différents sens de chacun de ces mots celui qui s'applique dans le contexte de la proposition, ainsi que mettre à jour les relations conceptuelles qui existent entre ces concepts dans le cadre de la proposition et qui font la cohérence de cette dernière. Une telle analyse permet de mettre à jour ce que l'on appelle le sens littéral des propositions. Ce sont les résultats de cette analyse sémantique que MLK prend comme entrée.

En dehors de ces deux rôles généraux, les connaissances sémantiques ont un statut particulier de pivot au sein de MLK. Elles interviennent en effet de façon constitutive à la fois dans le cadre de la mémoire pragmatique, pour former les schémas qui la composent, dans les représentations de texte, qui s'appuient largement sur le résultat de l'analyse sémantique, et enfin dans le cadre de la mémoire épisodique, puisque celle-ci est formée d'agrégats de représentations de texte.

[1] Correspondant au morceau de terre cuite, moulée, qui sert à ériger des murs.

[2] Défini comme le comportement de celui qui évite l'effort et aime l'oisiveté.

1.2. Forme de la mémoire conceptuelle : les graphes conceptuels

Le formalisme que nous avons adopté pour la représentation des connaissances sémantiques est celui des graphes conceptuels, proposé par John Sowa [Sowa 1984]. Son intérêt pour nous réside dans trois points. Tout d'abord, il offre la lisibilité et la facilité de manipulation manuelle qui sont propres aux réseaux sémantiques. Cette facilité d'appréhension s'accompagne cependant de la rigueur caractérisant les formalismes logiques du fait de l'existence d'une procédure de traduction des graphes conceptuels vers la logique des prédicats du premier ordre. Enfin, au delà du formalisme proprement dit, Sowa propose une structuration cohérente des différents types de connaissances sémantiques et même pragmatiques.

Dans ce qui suit, nous donnons un aperçu rapide du formalisme des graphes conceptuels et des opérations qui permettent de le manipuler. Nous ne présenterons que les aspects dont nous nous servons pour notre travail. On pourra se référer pour plus de détails aux nombreuses présentations existant sur les graphes conceptuels, dont bien sûr [Sowa 1984], [Sowa 1992] pour une vue plus synthétique, [Nogier 1991] et [Nazarenko 1994] pour une présentation en français.

1.2.1. Le formalisme

Comme il se doit pour un formalisme de représentation des connaissances sémantiques, la notion centrale des graphes conceptuels est celle de concept. Le concept résulte ici de l'association d'un type et d'un référent. Le référent spécifie l'ensemble des individus du monde de référence désignés par le concept. Il s'agit de sa définition en extension. Le type, quant à lui, indique quelle classe, contenant ces individus, est utilisée afin de les caractériser et renvoie donc à une définition en intension.

Notation des concepts : [<type_de_concept> : <référent>]

Un référent peut être :

- individuel, auquel cas il désigne un individu particulier.
  [Table: #12] : la table du monde de référence identifiée de façon unique par #12
  [Garçon: "Jean Spit"] : le garçon s'appelant Jean Spit (en supposant que le nom n'est pas ambigu dans le contexte)

- collectif. Il dénote alors un ensemble d'individus.
  [Fille: {"Julie Desbrosse", "Claudine Sage", "Juliette Toc"}] : l'ensemble formé par les filles Julie Desbrosse, Claudine Sage et Juliette Toc

- générique. On fait référence à un individu de la classe (ou à un ensemble d'individus dans le cas des référents collectifs génériques) sans en désigner un précisément.
  [Couteau : *] ou [Couteau] : un couteau
  [Fourchette: {*}] : un ensemble de fourchettes

Les types de concept sont quant à eux assez classiquement organisés de manière hiérarchique. Cette hiérarchie prend ici la forme d'un treillis dont tous les types sont inférieurs, relativement à une relation d'ordre partiel, au type universel (T) et supérieurs au type absurde ( ). La figure 4.1 donne un aperçu d'un tel treillis. Cette relation d'ordre partiel, en vertu de laquelle A < B, A et B étant des types de concept, traduit tout à la fois que B est plus général que A, que A hérite des traits de B et que l'extension de A est incluse dans celle de B. Chaque type de concept possède en effet une extension dans le monde de référence et la construction des concepts est soumise à la vérification de la relation de conformité : les individus dénotés par le référent du concept doivent faire partie de l'extension du type du concept.

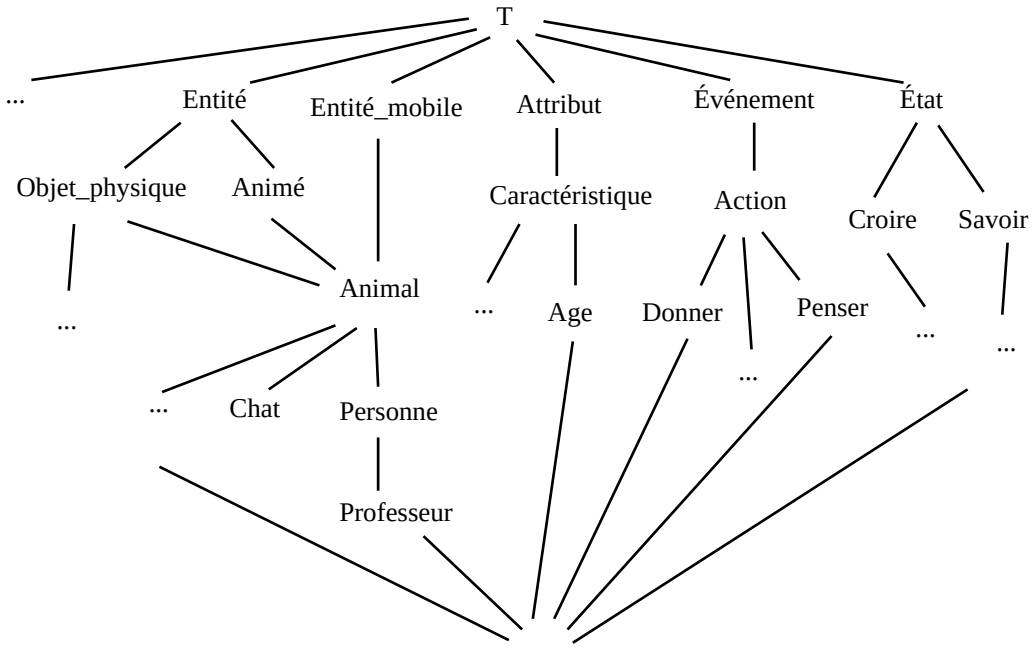


Fig. 4.1 - Aperçu du treillis de types de concept proposé par Sowa dans [Sowa 1984]

Afin de représenter des assertions, les concepts sont assemblés à l'aide de relations et forment ainsi des graphes conceptuels. Comme les concepts, les relations possèdent un type et peuvent être organisées en treillis. Cette dernière possibilité est cependant peu

exploitée car elle complexifie les opérations de manipulation[1]. Les graphes conceptuels se présentent donc comme des graphes bipartites (sommets de type concept et de type relation) auxquels on impose de surcroît la contrainte de connexité. Une relation peut faire intervenir de un à N concepts[2] mais en pratique, N se restreint le plus souvent à 2 et les relations binaires sont de loin les plus fréquentes pour représenter les propositions des textes. La figure 4.2 donne un exemple de graphe conceptuel, que l'on peut interpréter comme la représentation sémantique de la phrase *Jean conduit prudemment une voiture chère.*

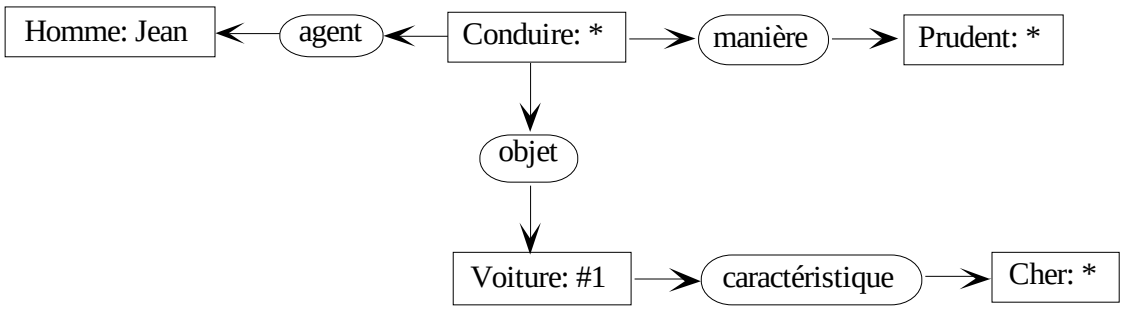


Fig. 4.2 - Un exemple de graphe conceptuel

Ce graphe peut être représenté également en adoptant la notation linéaire suivante.

```
[Conduire: *]
   {-> (agent) -> [Homme: Jean],
   -> (objet) -> [Voiture: #1] -> (caractéristique) -> [Cher: *],
   -> (manière) -> [Prudent: *]
   }
```

Nous avons légèrement modifié la notation linéaire proposée par Sowa dans [Sowa 1984] en synthétisant les extensions et les remarques présentées dans [Esch 1992], [Esch et alii 1994] et [Wermelinger 1995]. On trouvera à l'annexe A la grammaire complète de cette notation modifiée telle qu'elle utilisée par l'analyseur et le générateur que nous avons implantés.

Il faut ajouter que dans un graphe ou entre deux graphes appartenant à un même contexte, on peut spécifier une contrainte d'identité entre deux référents lorsqu'ils ne sont pas identifiés de façon unique, comme les référents individuels, grâce à l'utilisation de variables de co-référence. La figure 4.3 en offre une illustration. La variable *x permet

[1] Se référer à [Sabah & Vilnat 1991] pour une illustration de l'utilité d'une hiérarchie de relations pour l'analyse sémantique.

[2] L'arité d'une relation est définie au niveau de son type.

ainsi d'affirmer que l'homme qui mange le fruit et le même que celui qui possède un couteau.

La notion de variable de coréférence nous a amené à introduire celle de contexte et nous conduit à apporter des précisions sur les relations entre graphes conceptuels et logique. Un graphe représente une assertion que l'on peut exprimer, par l'intermédiaire d'une procédure de traduction, la fonction Φ, en une formule de la logique des prédicats. Le graphe (2) de la figure 4.3 se représente ainsi par la formule existentiellement quantifiée :

x, y, z Homme(x) Couteau(y) Posséder(z) Agent(x,z) Objet(y,z)

En général, on a besoin de représenter et de manipuler tout un ensemble cohérent d'assertions (liées implicitement par un ET logique), formant une base de connaissances. Dans le formalisme des graphes conceptuels, l'équivalent de la notion de base de connaissances est celle de contexte. Un contexte est concrètement un concept de second ordre, c'est-à-dire un concept ayant comme référent un graphe conceptuel ou un ensemble de graphes conceptuels. La boîte entourant les deux graphes de la figure 4.3 symbolise ainsi un concept de second ordre ayant le type de concept implicite *Proposition*.

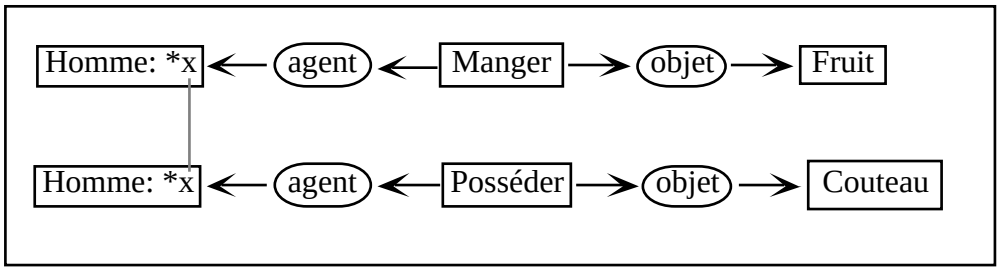


Fig. 4.3 - Un contexte rassemblant deux graphes

Ce statut de concept lui confère la capacité d'apparaître dans un graphe comme tout concept de base. La négation est ainsi exprimable par une simple relation unaire : (NON) -> [Proposition: ...]. On obtient de cette façon la même puissance d'expression que la logique des prédicats du premier ordre. Précisons que les contextes peuvent être emboîtés avec un niveau d'imbrication arbitraire, ce qui amène au delà du pouvoir expressif de la logique du premier ordre. En tout état de cause, tout graphe se situe toujours dans un contexte, au besoin dans celui de plus haut niveau, appelé "outermost context".

Sur le plan représentationnel, en particulier pour la sémantique des langues naturelles, on utilise les concepts de second ordre dès lors que l'on souhaite exprimer une attitude

vis-à-vis d'une assertion. Ils permettent donc de représenter tout ce qui a trait aux modalités.

1.2.2. Les opérations de manipulation

Sowa définit un corps d'opérations de base pour manipuler les graphes conceptuels qu'il compare aux règles d'une grammaire de graphes. Elles garantissent qu'à partir de graphes bien formés, on construit d'autres graphes bien formés, mais également plus spécialisés. En revanche, ce ne sont pas des règles d'inférence dans le sens où elles ne produisent pas nécessairement une assertion vraie à partir d'une assertion vraie. Ces opérations sont au nombre de quatre :

- la copie. Il s'agit simplement de réaliser une copie exacte d'un graphe. Cette opération puise sa raison d'être dans la nécessité de ne pas détruire les graphes d'origine lorsqu'on applique les autres opérations;

- la simplification. Cette opération permet de supprimer les relations redondantes lorsqu'entre deux mêmes concepts existent plusieurs relations de même type. Ce cas de figure s'obtient en particulier à l'issue de l'application de plusieurs jointures élémentaires;

- la jointure. La jointure permet sur un plan général de fusionner deux concepts identiques (même type et même référent) en rattachant au concept résultant les relations qui étaient liées aux deux concepts originels. Cette opération peut intervenir aussi bien à l'intérieur d'un graphe qu'entre deux graphes distincts. Dans ce second cas, le graphe obtenu est une réunion des deux graphes.

- la restriction. De façon générale, la restriction consiste à remplacer un concept d'un graphe par un concept plus spécifique. Celui-ci peut avoir un type inférieur au type du concept originel. Il s'agit alors d'une restriction de type. Il peut également avoir un référent moins large (recouvrant moins d'individus) ou plus spécifique, auquel cas on est face à une restriction de référent. Les deux types de restriction peuvent être combinés. Par ailleurs, l'opération de restriction se doit de conserver la relation de conformité entre type et référent.

La figure 4.4 illustre chacune de ces quatre opérations, qui ont bien sûr pour vocation de se combiner afin de réaliser des transformations plus complexes. Lorsqu'un graphe *v* est le résultat de l'application sur un graphe *u* d'une telle succession de ces opérations, on dit que *v* dérive de *u*. La copie et la simplification étant assimilables avant tout à des opérations de gestion, la dimension significative de ces macro-transformations est donc assurée par une combinaison de restrictions et de jointures. Ce sont d'ailleurs ces deux

opérations qui sont responsables de la spécialisation des graphes, d'une part en spécialisant leurs concepts et d'autre part, en ajoutant aux graphes de nouvelles branches, exprimant des propriétés supplémentaires. L'assertion *un homme jeune court dans le stade* est ainsi considérée comme une spécialisation de l'assertion *un homme court* et inversement la première est une généralisation de la seconde.

Plus formellement, la dérivation d'un graphe à partir d'un autre introduit une relation d'ordre partiel entre les graphes. Si *v* dérive de *u* alors on a : $v \leq u$. Du fait de l'absence d'opérations destructives, la simplification ne faisant que supprimer des informations redondantes, il est possible d'identifier au sein de *v* le sous-graphe, spécialisé du point de vue de ses concepts, correspondant à *u*. Ce sous-graphe est appelé la *projection* de *u* dans *v*. L'opération de construction de ce sous-graphe, notée π, est elle-même appelée *projection* et se comporte de la façon suivante vis-à-vis des concepts et des relations :

- pour chaque concept c de *u*, πc est un concept de *v*, c'est-à-dire de πu, tel que πc est une restriction de c;
- pour chaque relation r de u, πr est une relation de même type que r et telle que si le ième arc de r[1] pointe vers le concept c dans u, alors le ième arc de πr pointe vers le concept πc de πu.

La figure 4.5 fournit un exemple de projection. Il est à noter que la projection d'un graphe dans un autre peut fournir plusieurs résultats.

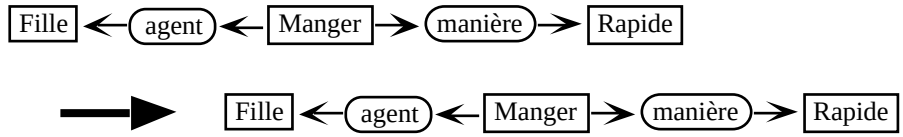
Sur un plan pratique, on voit que cette opération permet en quelque sorte de poser des questions. Entre deux graphes, la question serait du type : est-ce que l'assertion exprimée par le graphe recevant la projection est un cas particulier de l'assertion exprimée par le graphe à projeter? Avec un ensemble de graphes comme récepteur de la projection, il s'agit plutôt de sélectionner ceux pour lesquels on peut répondre par l'affirmative à la question précédente.

La dernière opération que nous évoquerons ici est la *jointure maximale*. Les quatre opérations que nous avons présentées en premier lieu constituent, comme nous l'avons dit, des primitives. Il est en pratique plus aisé de manipuler des opérations de plus gros grain dans la mesure où comme le suggère la figure 4.3 la jointure de deux graphes intervient souvent sur plus d'un concept et s'accompagne d'autre part fréquemment de plusieurs restrictions. La jointure maximale peut être vue comme une de ces macro-opérations. Grossièrement, elle consiste à réaliser une jointure entre deux graphes, non pas sur un concept identique mais sur le plus grand sous-graphe commun à ces deux

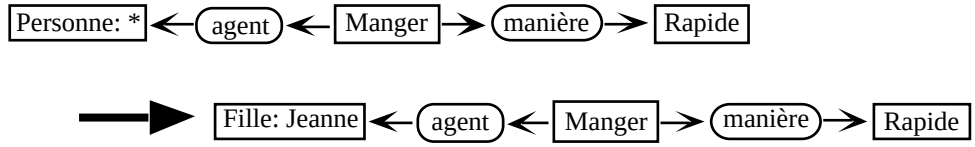
[1] Les arcs d'une relation vers les concepts qu'elle relie sont numérotés.

graphes, les concepts formant ce sous-graphe étant les restrictions minimales compatibles avec les concepts correspondant au niveau des deux graphes originels.

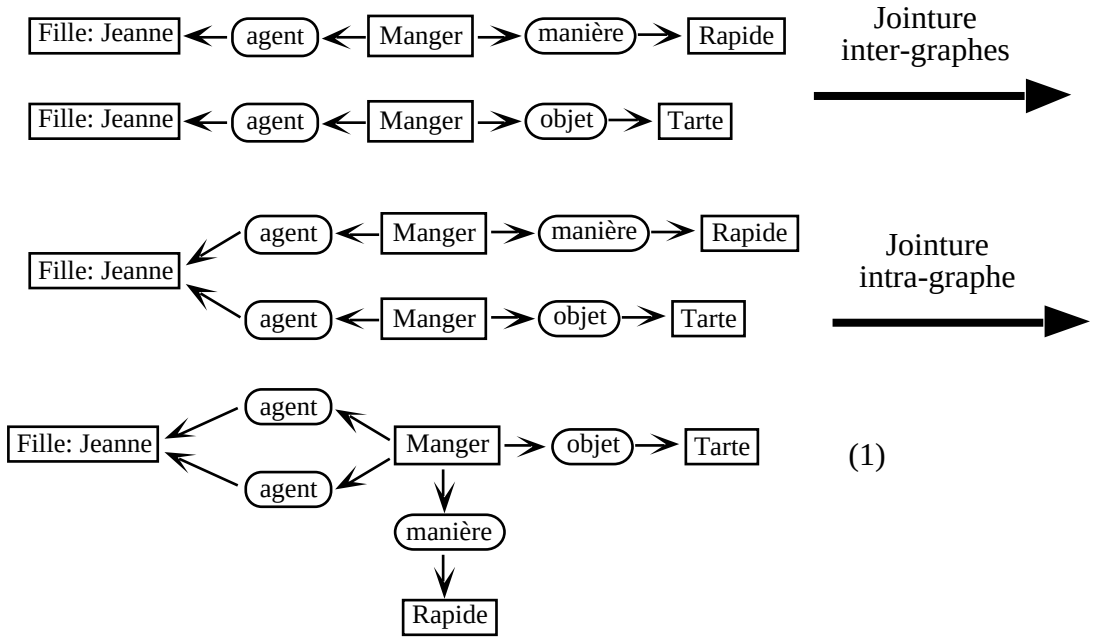
Copie



Restriction



Jointure



Simplification

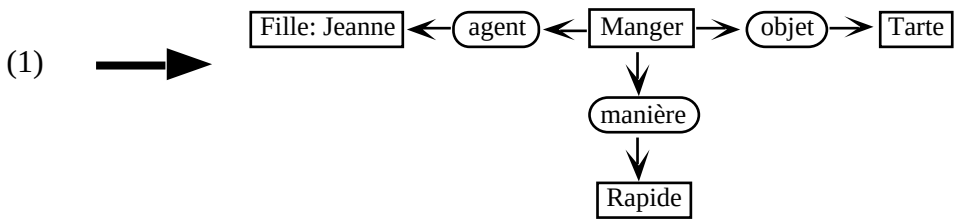


Fig. 4.4 - Les quatres opérations de base sur les graphes conceptuels (d'après [Sowa 1984])

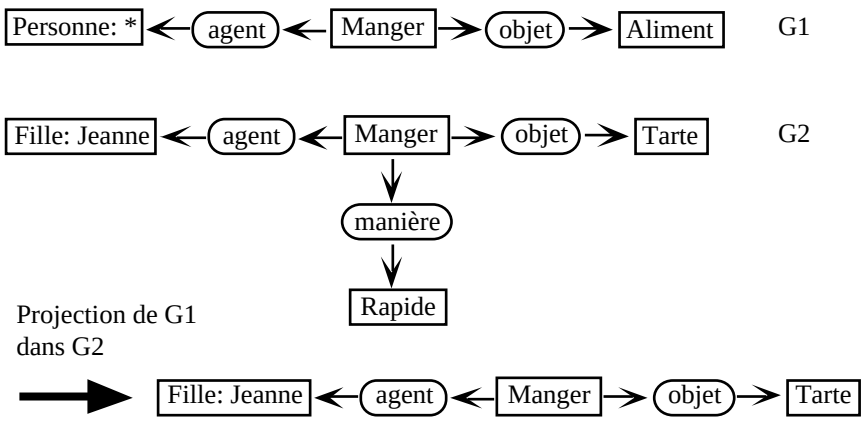


Fig. 4.5 - Un exemple de projection

La réalisation de la jointure maximale de deux graphes s'effectue en trois étapes. Il faut tout d'abord trouver une généralisation commune à ces deux graphes qui soit maximale, c'est-à-dire trouver le plus grand graphe pouvant se projeter dans chacun de ces deux graphes. Il est possible que ce plus grand graphe, uv_{max}, ne soit pas unique et la jointure maximale peut donc donner, comme la projection, plusieurs résultats. La seconde étape consiste à construire, par des restrictions sur les concepts, la spécialisation commune la plus faible entre les projections de uv_{max} dans chacun des deux graphes originels. La dernière étape, enfin, est chargée de raccrocher à cette spécialisation les morceaux des deux graphes qui ne faisaient pas partie de uv_{max}. La figure 4.6 illustre le résultat de cette opération.

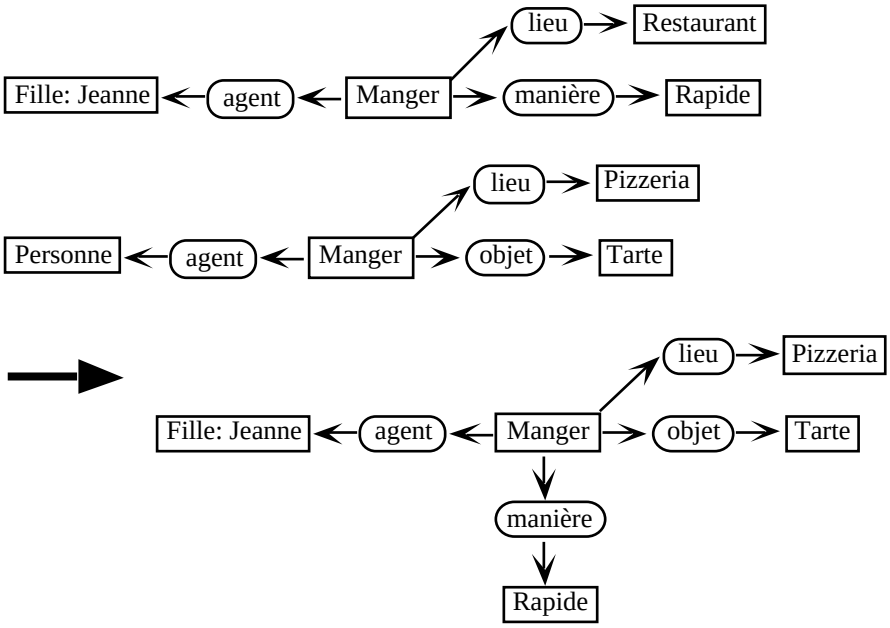


Fig. 4.6 - Un exemple de jointure maximale

1.2.3. La structuration des connaissances

Comme nous l'avons dit en préambule, un des points forts du formalisme des graphes conceptuels est la proposition par Sowa d'une structuration cohérente des différents types de connaissances. Cette structuration est organisée autour de la hiérarchie des types de concept. Chacun d'entre eux est en effet caractérisé, en dehors de son extension dans le monde de référence, par quatre types de connaissances, représentées toutes sous forme de graphes conceptuels :

- une définition en intension, chargée d'exprimer les traits différenciant le type de concept considéré de son ou de ses surtypes. Par référence à Aristote, le surtype d'un type est appelé *genus* et sa définition, *differencia*. Une définition est exprimée par une lambda abstraction ayant pour corps un graphe conceptuel. Le lien entre les paramètres de la lambda abstraction et les concepts représentant le genus est réalisé par des variables de co-référence. Ainsi, le type de concept Usine se définit par rapport au type de concept Bâtiment de la façon suivante :
  type Usine(x) is [Fabriquer]
  {-> (agent) -> [Ouvrier: {*}],
  -> (objet) -> [ObjetManufacturé: {*}],
  -> (lieu) -> [Bâtiment: *x],
  }.

- des contraintes impératives d'usage des concepts de ce type dans des graphes. Ces contraintes spécifient les relations, ainsi que les concepts qui sont liés à ces relations, qui peuvent être associés aux concepts ayant le type considéré. Les graphes représentant ces contraintes sont appelés *graphes canoniques* et l'ensemble de ceux-ci forment une *base canonique*. De par la définition de ces graphes, tout graphe conceptuel construit doit dériver, par l'intermédiaire des opérations exposées précédemment, de la base canonique. Le graphe canonique du type de concept Manger, donné par exemple par [Être_vivant] <- (agent) <- [Manger] -> (objet) -> [Aliment], devra ainsi pouvoir se projeter dans tout graphe faisant intervenir un concept de ce type;

- les contextes typiques dans lesquels intervient ce type de concept. Il s'agit d'exprimer par exemple que les concepts ayant pour type Chirurgien sont souvent impliqués dans des graphes faisant intervenir des concepts ayant pour type Infirmière ou Hôpital. Un schéma attaché au type Chirurgien pourra donc être un graphe exprimant qu'un chirurgien travaille avec des infirmières dans un hôpital. Ces connaissances sont appelées des schémas et se rapprochent des connaissances pragmatiques représentées par les MOPs. Alors que les graphes de définition et les

graphes canoniques sont uniques pour un type, celui-ci peut avoir un nombre illimité de schémas;

- les traits que l'on prête par défaut à un individu du type considéré. Il s'agit de préciser des traits qui ne sont pas intrinsèquement liés à ce type de concept, contrairement à ceux présents au niveau de sa définition, mais que l'on trouve régulièrement associés aux concepts de ce type. On pourra ainsi exprimer qu'une voiture possède quatre roues ou que le foie gras coûte cher. Le graphe rassemblant ces différents traits pour un type de concept est appelé un prototype.

Parmi ces quatre types de connaissances, on définit et on utilise prioritairement au niveau sémantique les définitions des types ainsi que leurs graphes canoniques. Nous verrons au §2 comment nous prenons en compte dans notre travail les connaissances exprimées par les schémas et les prototypes.

En dehors de leur importance dans la construction de la hiérarchie des types, les définitions ont également une grande importance au niveau de la forme de la représentation sémantique des textes. Les graphes canoniques fournissent les briques que l'on spécialise et que l'on assemble pour construire la représentation des propositions. Les graphes de définition interviennent quant à eux dans la définition de la nature exacte des briques utilisées. Cela est rendu possible par l'existence de deux opérations concernant les définitions de type : l'*expansion* et la *contraction* de type. L'expansion de type permet de remplacer dans un graphe un concept par le graphe de définition de son type. La contraction de type est l'opération strictement inverse. Si l'on reconnaît dans un graphe un sous-graphe correspondant à la définition d'un type, il est possible de remplacer ce sous-graphe par un concept de ce type.

Dans le cas de l'expansion, on peut avoir à réaliser des jointures entre des concepts de la définition et des concepts équivalents dans le graphe contenant le concept expansé. Expanser par exemple le concept de type Usine dans le graphe

[Travailler]
{-> (agent) -> [Ouvrier: {*}],
-> (lieu) -> [Usine],
-> (manière) -> [Intensif]
}

en utilisant la définition donnée précédemment donne le graphe suivant (en supposant que Fabriquer est un sous-type de Travailler) :

[Fabriquer]

{-> (agent) -> [Ouvrier: {*}],

-> (objet) -> [ObjetManufacturé: {*}],

-> (lieu) -> [Bâtiment: *x],

-> (manière) -> [Intensif]

}.

Dans le cas de la contraction, il est parfois nécessaire de conserver des éléments de la définition reconnue après la contraction dans le cas où ceux-ci sont plus spécifiques ou si l'absence de possibilité de raccrochage des parties pendantes ne garantit plus la connexité du graphe résultant.

Il est à noter que les types de relation possèdent eux aussi des définitions et que les opérations d'expansion et de contraction de type les concernent donc de la même façon.

On voit qu'en utilisant la contraction et l'expansion de type, on peut faire varier de façon conséquente la forme des représentations sémantiques. À l'extrême, on peut choisir de ne manipuler des graphes ne contenant que des types de concept et des types de relation primitifs, en nombre assez restreint, en réalisant des expansions successives de leurs concepts et de leurs relations jusqu'à obtenir ces types primitifs. Cela suppose bien entendu de ne pas introduire de circularité dans la définition des types et d'adopter une démarche de définition opérant strictement par niveaux. Si une telle façon de faire facilite la comparaison des graphes, elle est néanmoins d'un coût assez prohibitif en termes de complexité calculatoire.

1.2.4. Implémentation

Dans le cadre de notre travail, nous avons utilisé une plate-forme de gestion de graphes conceptuels développée en Smalltalk au sein du groupe Langage et Cognition du LIMSI. Cette plate-forme permet de représenter des graphes simples (c'est-à-dire des graphes ne comportant pas de contexte) ainsi des graphes comportant des contextes, éventuellement imbriqués. Les opérations développées pour la manipulation des graphes rassemblent l'essentiel de celles présentées dans [Sowa 1984] : les opérations de base – copie, simplification, jointure, restriction de type et de référent – les opérations plus avancées – projection et jointure maximale – les opérations en liaison avec les définitions de type – expansion de types de concept et de relation, contraction de types de concept et de relation – ainsi que des macro-opérations, comme la jointure dirigée sur un concept, qui représentent un assemblage couramment utilisé d'opérations de base.

Il est à noter qu'aucune de ces opérations n'est pour le moment sensible aux contextes. En outre, cette implémentation des graphes conceptuels ne prend pas encore en compte les opérations de nature logique et ne gère pas les types définis dynamiquement par des

λ-abstractions. Les référents de nature ensembliste sont également traités de façon très sommaire.

La plate-forme dispose par ailleurs d'un ensemble d'outils facilitant l'interaction avec l'utilisateur : gestionnaire de base de graphes, gestionnaire d'un treillis de types et des connaissances associées aux types, environnement de test des opérations sur les graphes. Les graphes conceptuels peuvent dans chacun de ces outils être visualisés et édités soit sous forme graphique, soit sous forme linéaire.

Notre contribution à cette plate-forme a consisté à raffiner certaines opérations ainsi que principalement, à définir un formalisme de représentation des graphes conceptuels sous forme linéaire (cf. Annexe A) en réalisant une synthèse des différentes extensions avancées à la suite de la proposition initiale de Sowa. Cette définition s'est accompagnée de la réalisation d'un compilateur permettant de construire des graphes à partir de cette forme linéaire ainsi que d'un générateur assurant à l'inverse le passage des graphes à leur forme linéaire, soit pour les visualiser, soit pour les stocker.

1.3. Quelques éléments à propos de l'utilisation des graphes conceptuels pour la modélisation des connaissances sémantiques

Notre objectif n'est pas ici de faire une étude rigoureuse de la façon dont on peut modéliser les connaissances sémantiques grâce aux graphes conceptuels, ce qui est le sujet de plusieurs thèses à lui tout seul, mais de préciser quelques aspects de la façon dont nous avons utilisé ce formalisme au niveau sémantique.

L'examen du problème de la représentation de la multiplicité des points de vue est une façon d'aborder assez globalement ces différents aspects. Le formalisme des graphes conceptuels s'appuie sur un treillis de types de concept, donc sur une organisation hiérarchique permettant de définir un type par rapport à plusieurs autres types plus généraux. Cette possibilité est au premier abord intéressante dans la mesure où il est nécessaire de pouvoir représenter plusieurs points de vue sur un même type de concept. Par exemple, une tomate peut être vue tantôt comme un fruit, tantôt comme une marchandise, une nourriture ou même un projectile à destination des cantatrices peu talentueuses. Cette dernière définition illustre d'ailleurs le flou de la frontière entre dimension sémantique et pragmatique. Si la définition de la tomate en tant qu'objet physique solide lui permet sans doute d'entrer dans la catégorie sémantique des projectiles, la relation avec les cantatrices est en revanche davantage du ressort d'un usage prototypique relevant de la pragmatique.

Même en en restant à la catégorie des projectiles sans aborder celle des cantatrices, on voit bien qu'un même type de concept peut être défini de multiples façons, chacune de ces définitions s'inscrivant dans un contexte plus ou moins large. Plusieurs solutions sont possibles pour faire face à cette nécessité. La solution la plus courante consiste à définir plus ou moins arbitrairement et plus ou moins implicitement la taille minimale du contexte d'application de la définition, ce qui revient à ne retenir qu'un nombre très restreint de surtypes (voire souvent un seul) pour chaque type de concept. Elle convient avant tout à des travaux opérant dans des domaines bien circonscrits.

La solution inverse consisterait à essayer de représenter le plus de points de vue possible pour chaque type de concept. C'est en pratique difficilement réalisable en ne s'appuyant que sur le treillis des types. On obtiendrait en effet un enchevêtrement assez inextricable de liens qui serait très difficile à maintenir. Il faut de fait constamment veiller à respecter la structure de treillis et surtout, trouver des définitions qui le soutiennent. La présence d'une seule définition par type de concept ne permet pas en outre de faire apparaître véritablement des points de vue à la fois multiples et suffisamment différenciés.

Nous préconisons une solution exploitant la structuration des connaissances proposée par le formalisme des graphes conceptuels. Son principe consiste à retenir un point de vue privilégié à partir duquel est construit la hiérarchie des types de concept et de représenter la multiplicité des points de vue au travers du graphe canonique qui leur est associé.

Exemple : le type de concept *Médecin* se définit alors comme un sous-type de *Profession* et possède le graphe canonique [Médecin] <- (aPourProfession) <- [Humain]. Ce dernier permet d'exprimer le raccourci métonymique en vertu duquel on a l'habitude de préciser qu'un médecin est un être humain. On évite ainsi de rendre *Médecin* à la fois sous-type de *Humain* et sous-type de *Profession.*

On restreint de cette façon le treillis des types de concept à une structure assimilable globalement à un arbre[1]. On obtient ainsi une organisation des types qui d'une part, est plus facilement manipulable par le modélisateur et d'autre part, réduit la complexité des opérations portant sur les graphes. La contre-partie de ces deux avantages est la production par l'analyse sémantique de graphes plus complexes pour représenter les propositions et la multiplication du nombre des types de relation. La représentation sémantique de la phrase "Le médecin court" sera en effet donnée par le graphe

[Médecin] <- (aPourProfession) <- [Humain: #123] <- (agent) <- [Courir]

au lieu d'être simplement donnée par le graphe

[Médecin: #123] <- (agent) <- [Courir].

[1] Ce n'est pas véritablement un arbre du fait de la présence persistante du type absurde devant être inférieur à tous les autres types de concept.

La multiplication des types de relation devrait en principe s'accompagner de leur hiérarchisation. Mais ceci accroîtrait la complexité des opérations portant sur les graphes et l'on peut se contenter en pratique de conserver un ensemble non structuré de types de relation. Il faut ajouter que cette démarche conduit à associer des graphes canoniques[1] à tous les types de concept alors que cela n'est réalisé généralement que pour les types de concept prédicatifs (c'est-à-dire en liaison avec des verbes ou des adjectifs).

1.3.1. Un exemple de hiérarchie des types de concept : les concepts verbaux

Le problème posé par la solution proposée ci-dessus consiste à définir un point de vue privilégié susceptible de soutenir la définition systématique de tout un ensemble de types de concept. Nous présentons ici brièvement un tel travail réalisé pour les types de concept verbaux par Karim Chibout [Chibout & Vilnat 1997]. Ce travail présente l'avantage de s'appuyer sur des principes clairement identifiés et d'avoir été mené sur une échelle assez large, puisqu'il a abouti à la construction d'une hiérarchie de 1000 types de concept verbaux environ, accompagnés chacun de leur graphe de définition et de leur graphe canonique. Il s'inscrit dans la ligne tracée par Wordnet [Miller et alii 1989].

Le fil directeur de ce travail est la modélisation de la polysémie verbale. La méthode sur laquelle il s'appuie est l'utilisation des définitions données par un dictionnaire terminologique. En dépit des variabilités que ces définitions comportent, on peut en effet y retrouver une structure générale assez bien établie. Chaque verbe est de fait défini à la fois par similitude par rapport à un hyperonyme immédiat et par différenciation vis-à-vis de celui-ci. Cette différenciation prend la forme de cas sémantiques récurrents.
Exemple : La définition du verbe couper s'écrit ainsi de la façon suivante :
couper : *diviser* (hyperonyme immédiat) un *objet solide* (objet) en *plusieurs parties* (résultat) à l'aide d'*un instrument tranchant* (moyen) en *traversant l'objet* (méthode).

En dehors des relations casuelles habituelles du type agent, objet, ..., il est à noter que les différenciations de sens sont plus spécifiquement fondées sur les cas résultat, méthode, but et manière.

En réalisant ce travail d'analyse sur un ensemble important de définitions et en suivant systématiquement le lien d'un hyperonyme à l'autre, Chibout a pu mettre en évidence l'existence d'une structuration des verbes s'articulant autour d'une quinzaine d'états primitifs. Ceux-ci correspondent à des états que l'on peut assimiler aux relations casuelles habituellement en usage : être possesseur de (possesseur), être dans un lieu (lieu),

[1] Pour être strictement rigoureux, il faudrait parler de graphes canoniques non triviaux dans la mesure où un type de concept a toujours un graphe canonique constitué au minimum d'un concept de ce type.

contenir quelque chose (contenant), être le patient de (patient), expérimenter quelque chose (expérienceur), ...

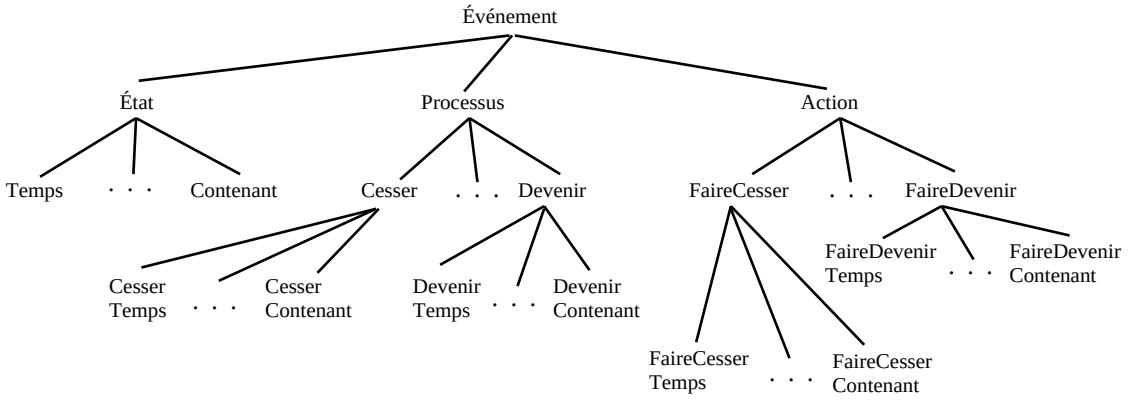


Fig. 4.7 - Le sommet de la hiérarchie des types de concept verbaux

Les actions et les processus sont définis par rapport à ces états. Les processus se répartissent ainsi entre ceux correspondant à l'entrée dans un de ces états (Devenir + état) et ceux qui en signifient une sortie (Cesser + état). Dans la sous-arborescence ayant comme sommet **Devenir Contenant**, on trouve ainsi des types de concept tels que *SeRemplir*, *Manger* ou *Boire* tandis que celle ayant pour sommet **Cesser Contenant** abrite des types de concepts comme *SeVider* ou *Vomir*. Au niveau des actions, on fait la distinction entre celles qui font entrer dans un des états primitifs (Faire Devenir + état) et celles qui en font sortir (Faire Cesser + état). *Créer* ou *Enfanter* héritent de cette manière de **Faire Devenir Temps** tandis que Interrompre ou Tuer se placent sous **Faire Cesser Temps**. La figure 4.7 illustre cette organisation du sommet de la hiérarchie des verbes.

La figure 4.8 donne quant à elle un aperçu de la hiérarchie des actions. Chaque type de concept verbal apparaissant sous les primitives (Possesseur, Devenir Expérienceur, Faire Cesser Expérienceur, ...) est défini selon les principes énoncés ci-dessus en le différenciant par rapport à son hyperonyme, cette différenciation étant réalisée par la donnée de valeurs spécifiques pour un certain nombre de cas.

Transposé dans le formalisme des graphes conceptuels, cette définition s'exprime au travers du graphe de définition associé à chaque type de concept. Il faut remarquer que la démarche adoptée par Chibout est très proche de la façon dont Sowa, au travers des notions de genus et differencia, envisage la définition des types. La figure 4.9 montre plus formellement comment le type de concept *Couper* est défini par un graphe conceptuel.

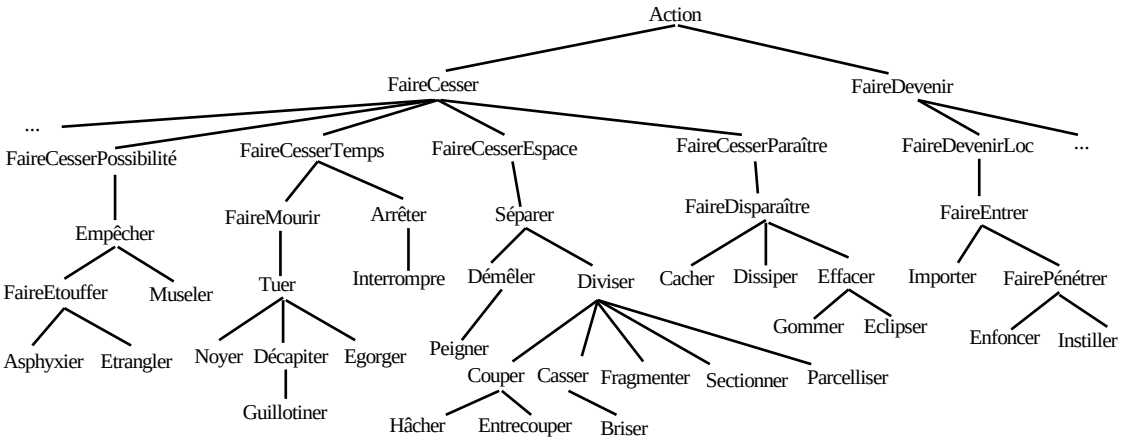


Fig. 4.8 - Aperçu de la hiérarchie des actions

Même s'il est intéressant de par la perspective qu'il trace, le travail de Chibout ne résout pas, comme on peut s'en douter, tous les problèmes posés par la modélisation des connaissances sémantiques. En particulier, il ne porte que sur les types de concept verbaux. On voit, notamment au travers des graphes canoniques, que ceux-ci posent quelques contraintes sur les autres types de concept mais cela ne donne pas lieu pour le moment à des principes systématiques de modélisation.

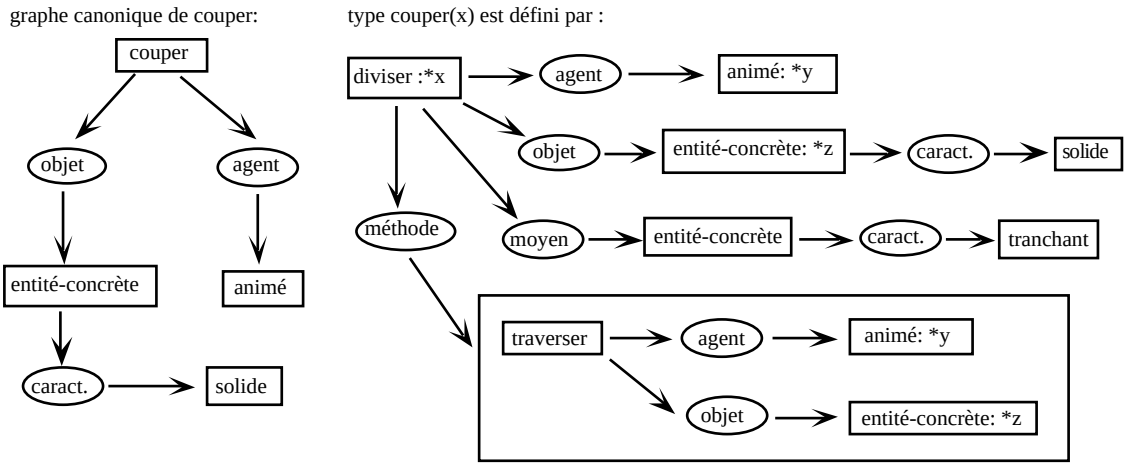


Fig. 4.9 - Graphe canonique et graphe de définition associés au type de concept verbal *Couper*

En fait, il est assez facile de créer intuitivement des sous-hiérarchies pour de petits domaines assez spécialisés (comme par exemple spécifier qu'un appareil photo numérique est une sorte d'appareil photo, qui est lui-même une sorte d'appareil de prise de vues) mais il est plus difficile de les réunir de façon véritablement cohérente. Chibout a ainsi hiérarchisé plus ou moins complètement près de 3000 types de concept nominaux en partant des définitions d'un dictionnaire terminologique mais sans retrouver de principes

aussi systématiques que pour les types de concept verbaux. Aucun définition ne leur a en particulier été associée.

Au niveau sémantique, nous restons donc dans la perspective qui est celle des domaines restreints, en attendant un travail de modélisation plus extensif du type Wordnet[1] adapté à la langue française.

1.4. L'hypothèse simplificatrice de l'utilisation des graphes conceptuels

Comme nous l'avons précisé au chapitre 1, notre travail s'inscrit dans le cadre plus général du modèle MoHA, lequel vise non seulement à l'apprentissage automatique de connaissances pragmatiques mais également à celui de connaissances sémantiques. Conformément aux principes de MoHA, ces connaissances sémantiques se doivent de mettre en œuvre une conception du sens qui soit dynamique à la fois du point de vue de l'interprétation d'un énoncé et du point de vue de sa capacité d'évolution en fonction de la confrontation à de nouvelles expériences. Le premier aspect de cette plasticité est équivalent à poser que la définition d'un concept doit s'adapter au contexte dans lequel il se trouve plongé. Par exemple, une table est vue comme un objet d'ameublement si l'on parle de l'aménagement d'une salle à manger mais doit être considérée comme une marchandise si l'on se place d'un point de vue économique.

Le second aspect du caractère dynamique des connaissances sémantiques répond quant à lui à un double souci. Il s'agit d'une part de rendre compte du développement de nouveaux concepts et d'autre part de l'évolution de concepts déjà affirmés.

Même en supposant un effort extensif de modélisation, il n'est pas possible de définir un ensemble clos de connaissances sémantiques. De façon évidente, à côté des concepts quotidiens, il existe une multitude de concepts spécialisés, attachés à des domaines spécifiques, dont il est impossible d'avoir une couverture complète. Par ailleurs, même dans le domaine des concepts largement partagés, il existe des créations constantes du fait du changement continu du monde qui nous entoure (changements sociaux, scientifiques et technologiques, artistiques, ...). On ne peut donc s'appuyer sur un ensemble fermé de concepts fournis a priori.

Les changements du monde nourrissent d'autre part la transformation des concepts déjà existants. Il est fort probable qu'un concept comme celui de liberté, par exemple, n'a pas véritablement le même contenu à l'heure actuelle que celui qu'il pouvait avoir à la fin du XVIII^{e} siècle ou chez les Grecs dans l'Antiquité. Même si un fond commun persiste

[1] Il faut préciser que Wordnet est d'abord un réseau de type lexical et que ce type de réseau ne contient pas autant de connaissances qu'un treillis et une base canonique de graphes conceptuels, notamment en ce qui concerne les définitions.

entre ces différentes formes, les situations auxquelles renvoie le concept en question changent. De ce fait, une partie de son contenu change également, et ceci, souvent sur une échelle de temps beaucoup plus courte que celle évoquée à propos du concept de liberté.

Face à ces deux exigences, adaptabilité par rapport au contexte d'usage et adaptabilité vis-à-vis de l'évolution de l'objet modélisé, l'utilisation des graphes conceptuels n'apparaît pas comme la solution la plus appropriée au premier abord. On a vu comment il est possible de rendre compte d'une certaine multiplicité des points de vue avec ce formalisme mais il s'agit là d'une proposition concernant sa représentation et non son utilisation pratique. Le recours aux graphes canoniques n'explicite pas en effet comment sélectionner un point de vue plutôt qu'un autre en fonction du contexte courant.

Par ailleurs, le formalisme n'intègre pas particulièrement de dimension ouvrant la voie à l'apprentissage. En particulier, il ne va pas du tout dans le sens de l'hypothèse sur laquelle repose MoHA et en vertu de laquelle les deux exigences d'adaptabilité posées ne peuvent être remplies qu'en supposant un ancrage des connaissances dans un niveau des expériences dont elles sont issues. Au mieux, les graphes conceptuels peuvent être vus comme une forme d'abstraction de connaissances sémantiques plus conformes à ces principes mais une abstraction ayant rompu tout lien avec son origine.

Le recours au formalisme des graphes conceptuels doit donc être considéré comme une solution temporaire permettant d'illustrer les principes mis en avant ici. Cela revient en quelque sorte à figer les connaissances sémantiques pour étudier l'apprentissage des connaissances pragmatiques, sachant que faire évoluer ces deux dimensions simultanément est une tâche de trop vaste ampleur pour être abordée directement. Même si cette façon de procéder n'est pas exempt de biais, nous estimons qu'elle permet au moins de spécifier les solutions retenues de façon suffisamment précise pour que celles-ci puissent ensuite être transposées dans un cadre plus réaliste au sein duquel un amorçage pourra intervenir à la fois sur la dimension sémantique et la dimension pragmatique. Nous donnerons d'ailleurs à la fin de la présentation de ce travail des indications concernant la manière dont les travaux de Jean-Pierre Gruselle sur l'émergence de concepts, qui sont parallèles aux nôtres sur l'émergence de connaissances pragmatiques, peuvent y être connectés afin de concrétiser plus largement les principes de MoHA.

2. Mémoire pragmatique

2.1. Nature et rôle de la mémoire pragmatique

Le rôle premier de la mémoire pragmatique est d'abriter les connaissances pragmatiques stables concernant la représentation des situations prototypiques du monde de référence. Le parallèle avec les MOPs est à cet égard immédiat, tant sur le plan du fond que sur celui de la forme, puisque ces connaissances sont représentées ici aussi par des schémas. Elles constituent d'autre part une forme d'aboutissement du système ANTHAPSI. À ce titre, elles ont donc une influence certaine sur les formes de connaissance qui en sont les précurseurs, autrement dit les UTs agrégées, et par voie de conséquence, sur les UTs elles-mêmes. Nous aurons l'occasion de revenir sur cette similarité de structure dans les chapitres 5 et 6.

2.1.1. Nature de la mémoire pragmatique

Pour être plus précis, on peut différencier la mémoire pragmatique de la mémoire épisodique sur trois critères : le mode d'évolution, la structuration et le niveau d'abstraction. La mémoire épisodique s'apparente globalement à une sorte de chaudron ou de creuset en ébullition dans lequel on injecte en continu des ingrédients et qui produit de façon progressive de nouveaux composés intéressants, insérés au sein d'une gangue formée des déchets des réactions intervenues. Elle constitue donc un milieu assez instable mais caractérisé par un mode d'évolution continu. La mémoire pragmatique présente les caractéristiques strictement inverses. Ses connaissances sont stables et leur évolution, qu'il faut plus envisager sous l'angle de restructurations de type spécialisation ou généralisation de schémas [Grau & Sabah 1985] que sous celui de changements de nature profonds, s'effectue de façon discrète.

Cette différence recèle en fait une complémentarité. La mémoire épisodique est adaptée à l'émergence de nouvelles unités. Mais celles-ci sont des produits bruts dont on ne peut faire un usage très pointu, du fait justement de la gangue qui les accompagne. Une UT agrégée résulte en effet de l'accumulation des représentations d'une même situation par plusieurs textes et les détails propres à chacun d'entre eux s'y retrouvent donc au milieu d'éléments plus centraux pour la situation. En gagnant en pureté, les schémas de la mémoire pragmatique perdent la possibilité d'évoluer continûment. En revanche, ils forment une connaissance beaucoup plus sûre : en utilisant un schéma, on n'a pas à déterminer si l'élément que l'on considère en fait véritablement partie. Cette question a été tranchée une fois pour toutes lors de l'abstraction de la ou des UTs agrégées dont ce schéma est issu.

Cette définition plus fine et plus sûre de ses unités de base permet, et c'est le second volet de la complémentarité évoquée, à la mémoire pragmatique de se concentrer particulièrement sur les relations que ces unités entretiennent entre elles. Au sein de la mémoire épisodique, on peut considérer que l'on a globalement une structure à plat : une UT agrégée rassemble les propositions de différents textes relatives à une même situation mais si plusieurs de ces propositions évoquent une situation plus élémentaire, on ne leur substitue pas un lien vers une UT agrégée représentant cette situation lorsque celle-ci existe en mémoire. À ce stade, les connaissances pragmatiques n'ont pas acquis leur autonomie de définition par rapport aux connaissances sémantiques.

Un schéma, au contraire, rend principalement compte d'une situation en restant au niveau pragmatique, c'est-à-dire au travers des relations que cette situation entretient avec d'autres : elle fait référence pour sa description à des situations plus élémentaires ou, à l'inverse, elle participe à la définition de situations plus importantes. Elle peut également représenter une spécialisation d'une autre situation ou la généraliser. La mémoire pragmatique est donc avant tout un vaste réseau de schémas et son évolution s'incarne pour une bonne part dans le changement de la structure de ce réseau. Les relations entre schémas ayant une signification bien déterminée, de tels changements se font selon des critères explicites et non par la simple modification de poids. C'est pourquoi la mémoire pragmatique évolue de manière discrète.

Nous avons vu plus haut que le passage d'une UT agrégée à un schéma s'accompagne de l'abandon d'un certain nombre de composants de la première jugés comme contingents. C'est un premier aspect du processus d'abstraction qui caractérise ce passage.

Le second concerne le niveau d'abstraction des composants de part et d'autre. Au sein d'une UT agrégée, deux propositions peuvent avoir été jugées similaires même si leurs concepts ne sont pas strictement identiques. Une certaine flexibilité existe en effet quant au type des concepts, pourvu qu'ils conservent entre eux une relation de hiérarchie. Une UT agrégée peut donc rassembler des situations proches mais de niveaux de généralité différents. Dans un schéma, un tel flou n'existe pas. Soit les situations ont été jugées suffisamment proches et l'on obtient un seul schéma dont les composants généralisent ceux de l'UT agrégée; soit elles sont suffisamment éloignées pour donner lieu à la création de plusieurs schémas, lesquels peuvent éventuellement être liés par une relation de spécialisation. On retrouve là encore au niveau des schémas la volonté d'être le plus précis et le plus explicite possible.

2.1.2. Utilisation de la mémoire pragmatique

Même s'ils représentent un certain point d'aboutissement pour ANTHAPSI dans sa forme actuelle, les schémas n'en sont pas pour autant une fin en soi. Leur intérêt réside en effet dans l'usage que l'on peut en faire pour l'analyse des textes. Dans le cadre de l'analyse thématique, sur laquelle nous nous centrons ici, les schémas interviennent doublement. D'une part, ils sont utilisés par le mécanisme d'analyse de MLK au même titre que les UTs agrégées. D'autre part, ils servent de support à une analyse thématique plus élaborée qui leur est spécifiquement dédiée.

L'analyse thématique de MLK est conçue pour exploiter des UTs agrégées mais ainsi que nous le verrons au chapitre 8, elle peut également faire usage de schémas lorsqu'ils sont disponibles, soit en tant que source unique de connaissances, soit en conjonction avec les UTs agrégées. Cette possibilité est ouverte du fait de la proximité entre schémas et UTs agrégées mais elle ne peut être considérée comme véritablement importante dans la mesure où elle n'exploite pas les spécificités des schémas en termes de niveau d'abstraction et de structuration.

En revanche, Brigitte Grau, dans [Grau 1983], présente un mécanisme d'analyse thématique reposant spécifiquement sur la notion de schéma. Ce mécanisme établit de façon fine quelles relations de suivi thématique lient les différentes propositions d'un texte. Le type de la relation unissant deux propositions est directement déduit des relations que l'on peut trouver entre les schémas évoqués par chacune d'elle au sein de la base de connaissances pragmatiques de référence. Cette base prend la forme d'un réseau de schémas tout à fait comparable à celui que nous présentons ici.

Ce type d'analyse représente une évolution par rapport à celui proposé par MLK dans la mesure où ce dernier se contente principalement de regrouper les propositions d'un texte relative à un même thème mais ne dispose pas des moyens de suivre finement le développement des thèmes. En conservant la perspective d'amorçage qui est la nôtre, on peut donc envisager ANTHAPSI, par sa capacité à faire émerger des schémas, comme un moyen d'amorcer une analyse thématique à base de schémas. Nous verrons néanmoins au chapitre 7 que cet amorçage pose quelques problèmes spécifiques.

2.2. Forme de la mémoire pragmatique

La mémoire pragmatique est constituée dans notre cas d'un type unique de composant, le schéma, chargé de décrire de façon abstraite et déclarative une situation prototypique du monde de référence. Un schéma rassemble dans ce but les représentations des actions et des états qui caractérisent la situation ainsi que les personnages qui en sont les acteurs. Il

explicite également les relations de nature causale et temporelle qui lient ces actions et ces états.

Tant sur le plan de la structure générale d'un schéma que sur celui de l'organisation des schémas au sein de la mémoire pragmatique, nous reprenons les principes énoncés dans [Grau 1983]. La principale modification que nous y apportons réside dans l'utilisation du formalisme des graphes conceptuels pour faire le lien avec les connaissances sémantiques et exprimer un certain nombre de contraintes sur les concepts impliqués dans les schémas.

2.2.1. Structure des schémas

Une première vue d'ensemble d'un schéma conduit à distinguer trois parties principales : un en-tête, un corps et des rôles.

En-tête des schémas

L'en-tête est un graphe conceptuel permettant d'exprimer la façon dont la situation est évoquée sur le plan sémantique. À ce titre, il peut être vu comme une sorte de nom attribué au schéma possédant la particularité d'être significatif du point de vue du système qui le manipule. Ce dernier point le différencie de l'identifiant qui est effectivement attribué à chaque schéma (le schéma de la figure 4.10 porte ainsi le nom de "Kidnapping") et qui permet d'y faire référence. Celui-ci n'est en effet qu'une étiquette qui n'a de sens au mieux que pour la personne qui a construit le schéma, lorsque celui-ci est le fruit d'une modélisation manuelle.

L'en-tête d'un schéma a également un rôle de déclencheur de ce schéma lors de l'analyse des textes, ou plus précisément un rôle de confirmation de son déclenchement. Lors de la recherche des liens entre propositions, on commence par sélectionner un ensemble de schémas susceptibles de s'appliquer à la situation évoquée par le texte en utilisant généralement les concepts de ces propositions comme clés de recherche dans une structure d'indexation des schémas. Il faut ensuite en choisir plus spécifiquement un seul sur des critères plus stricts. L'en-tête d'un schéma en offre la possibilité. En réalisant une opération de projection (au sens des graphes conceptuels) de l'en-tête d'un schéma dans la représentation sémantique d'une proposition (qui est aussi un graphe conceptuel ici), on peut en effet déterminer si cette proposition est compatible avec la représentation sémantique de la situation correspondant à ce schéma. Si la projection réussit, il semble ainsi raisonnable de retenir ce schéma pour interpréter la proposition sur le plan pragmatique.

Plus globalement, l'entête d'un schéma permet donc de faire la liaison entre les représentations sémantiques et les représentations pragmatiques.

Corps des schémas

Le corps d'un schéma permet quant à lui de décrire sur le plan pragmatique le contenu de la situation qu'il représente. Il est principalement formé d'un ensemble de références vers d'autres schémas de la mémoire pragmatique représentant des actions et des états plus spécialisés[1]. Le schéma de la figure 4.10 fait ainsi référence pour se définir à des schémas tels que *Séquestration, DemandeDeRançon* ou *LibérationOtage* qui décrivent des événements plus élémentaires intervenant lors d'un kidnapping.

Références vers les schémas

Lien avec le schéma référencé

Chaque référence est au moins constituée de deux éléments. Elle comporte de façon évidente le schéma référencé, mais également un graphe conceptuel qui est une copie de son en-tête. Cette copie permet de définir des contraintes d'identité entre les rôles d'un schéma et ceux du schéma auquel il fait référence. On utilise pour cela le mécanisme de coréférence entre concepts existant au niveau des graphes conceptuels. Celui-ci permet de spécifier que deux concepts appartenant respectivement à deux graphes plongés dans un même contexte[2] font référence à une même entité du monde.

Dans notre cas, chaque schéma se voit associé un contexte rassemblant tous les graphes impliqués dans sa définition. On peut de cette façon préciser que la victime du kidnapping, désignée par la variable *x1* au niveau de l'entête du schéma, est la même personne que la victime de la séquestration, apparaissant quant à elle avec la même variable *x1* dans la copie de l'en-tête du schéma *Séquestration* présente au niveau de la référence faite à ce schéma. Bien entendu, cette victime pourra être désignée par une tout autre variable dans la définition du schéma *Séquestration* puisque les contextes de deux schémas sont indépendants.

---

[1] Le lecteur pourrait légitimement trouver que les dénominations sont un peu confuses : on parle tantôt de situations et tantôt d'actions et d'états. Toutefois, il faut bien voir qu'il ne s'agit là que des deux faces d'une même médaille. Ainsi que nous l'avons vu au chapitre 1, un événement tel que "Aller au cinéma" peut aussi bien être envisagé sous un angle sémantique (définition et contraintes d'usage des types de concept Aller et Cinéma) que sous un angle pragmatique (décomposition de cet événement en événements plus élémentaires), donc être vu aussi bien comme une action que comme une situation.

[2] Les coréférences ont dans le cas général un champ d'application plus large puisque les graphes peuvent appartenir à deux contextes différents pourvu que ceux-ci soient emboîtés (pas nécessairement de manière directe d'ailleurs).

Sur le plan pratique, la mise en correspondance de rôles appartenant à deux schémas distincts, A et B, avec A référençant B, s'opère grâce à une projection du graphe associé à la référence faite à B dans A dans le graphe constituant l'en-tête de B. Cette opération permet d'établir une liste d'équivalences entre variables. Le mécanisme retenu pour cette mise en relation des rôles présente en outre l'avantage de s'affranchir d'une éventuelle différence des schémas en termes de niveau d'abstraction. Par exemple, la rançon est représentée ici par un concept ayant le type *Argent* dans l'en-tête du schéma *DemandeDeRançon* mais il apparaît sous les traits d'un concept ayant le type *ObjetDeValeur* dans l'en-tête du schéma *RemiseDeRançon*. Cette différence (peut-être un peu artificielle ici) n'empêche pas néanmoins la mise en correspondance des rôles car le mécanisme des coréférences peut intervenir entre des concepts n'ayant pas le même type, pourvu que la relation de conformité entre type et référent soit respectée.

Poids associés à une référence

Les références comportent également deux autres éléments, moins essentiels. Ce sont tous deux des poids normalisés entre 0 et 1. Le premier, appelé *importance*, caractérise l'importance que revêt le schéma référencé dans la définition du schéma source de cette référence. Il permet ainsi de stipuler si la présence de l'événement désigné est obligatoire pour que le schéma soit reconnu ou si cet événement peut être absent sans empêcher pour autant le déclenchement du schéma. Par exemple, il semble évident que l'enlèvement et la séquestration de la personne cible du chantage ainsi que la demande de rançon qui les accompagne forment le cœur d'une situation de kidnapping telle que celle décrite par la figure 4.10, alors que la fuite des ravisseurs est un événement plus contingent, dépendant en particulier d'une éventuelle intervention de la police pour libérer l'otage.

Le second poids, appelé *spécificité*, traduit quant à lui le degré de spécificité de l'événement considéré par rapport à la situation représentée par le schéma. On peut également voir ce poids comme l'expression du degré de typicalité de cet événement relativement à cette même situation. Cette notion est bien distincte de celle d'importance. Un schéma peut être en effet spécifique par rapport à un autre, c'est-à-dire ne pas intervenir dans la définition de beaucoup d'autres schémas en dehors de celui-ci, mais ne pas y être forcément important. Le schéma *LibérationOtage* est ainsi plus spécifique qu'il n'est important dans la figure 4.10 dans la mesure où cet événement est moins fréquent que la séquence enlèvement – séquestration – demande de rançon tout en étant néanmoins propre à un kidnapping. À l'inverse, le fait que la victime soit riche est une condition importante pour le déroulement d'un kidnapping crapuleux mais c'est un état peu discriminant vis-à-vis de cette situation, chose qui devrait rassurer les personnes possédant un peu de fortune.

La spécificité revêt par ailleurs un intérêt tout particulier lors de l'analyse d'un texte pour la sélection des schémas les plus pertinents relativement à un contexte donné. En effet, si un événement est très typique d'une situation, sa présence dans un texte conduit naturellement à sélectionner le schéma qui la représente afin d'interpréter le passage du texte où il apparaît. Ce problème de la sélection des connaissances, partagé avec la mémoire épisodique, sera plus particulièrement développé dans le chapitre 6.

Chacun de ces deux poids répond ainsi à deux besoins complémentaires de l'application des schémas à l'analyse de textes : la spécificité offre le moyen de sélectionner rapidement un ensemble de schémas présumés pertinents tandis que l'importance constitue le support de la vérification du bien-fondé de cette sélection. Il faut préciser que la distinction de ces deux poids reprend la dichotomie opérée par Lebowitz dans IPP (cf. chapitre 2) entre l'assurance d'une part (assimilable à l'importance) et la prédictivité d'autre part (équivalent à la spécificité), notion plus particulièrement développée dans [Lebowitz 1983].

Lorsqu'un schéma est issu d'une modélisation manuelle, ces deux poids peuvent être fixés par le modélisateur en fonction de sa connaissance de la situation. C'est ce qui s'est passé pour le schéma de la figure 4.10 par exemple. En revanche, lorsqu'un schéma est construit automatiquement, comme avec l'abstraction des UT agrégées dans ANTHAPSI, il faut trouver une méthode elle aussi automatique de détermination de ces poids. En ce qui concerne l'importance, il est possible d'exploiter les poids que possèdent les événements dans les UTs agrégées, même si la forme de ces événements est un peu différente. En effet, en caractérisant un degré de récurrence, ces poids rendent compte également d'une notion d'importance. Nous renvoyons le lecteur au chapitre 8 sur l'abstraction des UTs agrégées pour plus de détails sur ce point.

En ce qui concerne la spécificité, nous proposons une méthode d'évaluation de base présentant l'avantage de pouvoir être utilisée aussi bien dans le cadre d'une modélisation manuelle que d'une construction automatique. Cette méthode s'appuie sur une évaluation de la spécificité globale des schémas : on fait en pratique l'hypothèse que plus un schéma est spécifique et plus le nombre de schémas le référençant dans le cadre de leur définition est faible. Son contexte d'application est en effet supposé plus étroit. Notre méthode de calcul de la spécificité d'un schéma I au sein d'un schéma J tient compte à la fois de cette spécificité globale et du positionnement de celle-ci par rapport à la spécificité des autres schémas intervenant dans la définition de J. I est donc d'autant plus spécifique vis-à-vis de J que I est à la fois spécifique à l'échelle de la mémoire pragmatique et que cette spécificité est forte parmi celle des autres schémas servant à le définir. Pour estimer ce

dernier rapport, on prend comme référence commune la plus petite valeur de spécificité pour l'ensemble de J. On obtient ainsi la mesure suivante :

$$specif(s_i, s_j) = \sqrt{\frac{1}{Nref_i} \frac{\min_k(Nref_k)}{Nref_i}} = \frac{\sqrt{\min_k(Nref_k)}}{Nref_i}$$

avec

specif(s_i,sj) : spécificité du schéma s_i au sein du schéma s_j;

$Nref_i$: nombre de références faites au schéma s_i au sein de la mémoire pragmatique;

k : indice énumérant tous les schémas intervenant dans la définition de s_j.

La spécificité globale du schéma s_i est donnée par le terme ($1/Nref_i$) tandis que le second terme met en balance cette spécificité globale avec la plus forte parmi celles des schémas de s_j. L'utilisation de la racine carrée est un moyen de compenser l'effet du produit combinant ces deux termes afin de ne pas avoir des valeurs trop petites.

Structuration des références

Les différentes références d'un schéma sont structurées de deux façons. La première est globale. Elle consiste à les répartir entre trois grands attributs caractérisant chacun quelle dimension de la situation est décrite :

- *circonstances* : cet attribut rassemble les états qui sont vrais avant que la situation ne prenne place et que l'on peut considérer comme indicatifs de sa survenue. Certains de ces états pourront ne pas être altérés par la situation tandis que d'autres ne seront plus valides à son terme;

- *description* : l'attribut *description* est constitué des actions caractérisant la situation. Il en forme le corps véritable;

- *états incidents* : cet attribut regroupe les états qui sont vrais à l'issue de la situation. Il s'agit en général des états induits par la situation et non de ceux qui sont restés vrais tout au long de celle-ci, lesquels figurent plutôt dans l'attribut *circonstances.*

Cette tri-partition, qui reprend celle présente classiquement en planification au travers du triptyque pré-conditions – actions – résultats, se retrouve sous des formes plus ou moins élaborées dans l'essentiel des schémas représentant le même type de connaissances (cf. [Mooney & DeJong 1985] et [Ram 1993] par exemple).

La seconde structuration des références est donnée par les relations temporelles et causales qu'elles entretiennent. Ces relations rendent compte de l'enchaînement des événements au sein de la situation ainsi que des raisons pour lesquelles certains d'entre eux interviennent. Elles peuvent lier des références situées au sein du même attribut

comme des références appartenant à deux attributs distincts. La relation de précédence temporelle entre les schémas *Enlèvement* et *Séquestration* au sein de l'attribut *Description* du schéma de la figure 4.10 est un exemple du premier cas. La relation causale entre le schéma *RemiseDeRançon* de l'attribut *Description* et le schéma *DevenirRiche* de l'attribut *ÉtatsIncidents* est un exemple du second. Il faut d'ailleurs noter que généralement, les relations temporelles d'ordonnancement des événements sont cantonnées au sein de l'attribut *Description* alors que les relations causales, qui ont surtout pour objet de lier les actions aux états qui les motivent ou qui en sont les résultats, prennent place entre l'attribut *Description* et les attributs *Circonstances* ou *ÉtatsIncidents*.

Les relations possèdent comme les références un poids marquant leur degré d'importance dans le schéma mais n'ont pas de degré de spécificité. Cette dernière notion n'a pas en effet été retenue pour les relations dans la mesure où elles ne sont pas impliquées dans la sélection des schémas. Ce sont en effet ces derniers qui doivent les mettre à jour.

Graphes d'expression de contraintes

Comme le montre la figure 4.10, les attributs ne contiennent pas que des références vers d'autres schémas. Ils peuvent également abriter des graphes conceptuels chargés d'exprimer des contraintes sur les rôles du schéma. Ces contraintes sont de nature sémantique et ne renvoient pas à des situations. C'est pourquoi elle n'apparaissent pas sous la forme de références vers des schémas. Il s'agit par exemple d'exprimer qu'un objet est une partie d'un autre objet ou qu'il présente une caractéristique spécifique. De par cette fonction, ces graphes sont placés dans l'attribut *Circonstances*.

Dans le schéma de la figure 4.10, ils permettent ainsi de spécifier que les humains agents du kidnapping sont des ravisseurs, que l'humain patient de ce même kidnapping est un otage tandis que l'objet de valeur impliqué dans la situation sert de rançon. Ces précisions sont rendues nécessaires par la façon dont nous avons organisé les connaissances sémantiques (cf. §4.1). Les graphes canoniques associés aux types de concept prédicatifs[1] ne contiennent que des types très généraux (comme *Humain, ÊtreVivant* ou *ObjetPhysique*). Il est donc nécessaire de faire le lien entre cette caractérisation très générale d'une entité du monde et le rôle plus spécifique qu'elle joue dans une situation.

[1] Les types de concept considérés comme prédicatifs sont ceux s'exprimant en langue sous la forme de verbes ou d'adjectifs. Les concepts ayant un statut de prédicat (c'est-à-dire possédant un type de nature prédicative) sont désignés au sein des graphes par une annotation spécifique (prédicat: vrai), suivant une extension de la forme linéaire des graphes conceptuels.

Schéma KidnappingCrapuleux

spécialisationDe: Chantage

Entête:
[Humain: *x1] ← (patient) ← [Kidnapper; prédicat: vrai] → (agent) → [Humain: {*} *x2];

Rôles

agent -> [Humain: {*} *x2];
valDef: Humain;
spécificité: 0.1;
importance: 1.0.

patient -> [Humain: *x1];
valDef: Humain;
spécificité: 0.1;
importance: 1.0.

relation -> [ObjetDeValeur: *x3];
valDef: Argent;
spécificité: 0.25;
importance: 0.9.

relation -> [Humain: *x4];
valDef: Humain;
spécificité: 0.1;
importance: 0.9.

FinRôles

Attribut Circonstances

Schéma ÊtreRiche
importance: 0.75;
spécificité: 0.25;
[ÊtreRiche; prédicat: vrai] → (expérienceur) → [Humain: *x1].

Schéma ÊtreUnProcheDe
importance: 0.75;
spécificité: 0.1;
[ÊtreUnProcheDe; prédicat: vrai]
{→ (source) → [Humain: *x4],
→ (patient) → [Humain: *x1]
}.

Graphe Otage
importance: 1.0;
spécificité: 1.0;
[Humain: *x1] → (caractéristique) → [Otage].

Graphe Ravisseur
importance: 1.0;
spécificité: 1.0;
[Humain: {*} *x2] → (caractéristique) → [Ravisseur].

Graphe Rançon
importance: 1.0;
spécificité: 1.0;
[ObjetDeValeur: *x3] → (caractéristique) → [Rançon].

FinAttribut Circonstances

Attribut Description

Schéma Enlèvement
importance: 1.0;
spécificité: 1.0;
[Humain: *x1] ← (patient) ← [Enlever; prédicat: vrai] → (agent) → [Humain: {*} *x2].

Schéma Séquestration
importance: 1.0;
spécificité: 1.0;
[Humain: *x1] ← (patient) ← [Séquestrer; prédicat: vrai] → (agent) → [Humain: {*} *x2].

Schéma DemandeDeRançon
importance: 0.9;
spécificité: 1.0;
[Exiger; prédicat: vrai]
{→ (objet) → [Argent: *x3],
→ (agent) → [Humain: {*} *x2],
→ (receveur) → [Humain: *x4]
}.

Schéma RemiseDeRançon
importance: 0.7;
spécificité: 1.0;
[Donner; prédicat: vrai]
{→ (objet) → [ObjetDeValeur: *x3],
→ (receveur) → [Humain: {*} *x2],
→ (agent) → [Humain: *x4]
}.

Schéma LibérationOtage
importance: 0.6;
spécificité: 1.0;
[Humain: *x1] ← (patient) ← [Relâcher; prédicat: vrai] → (agent) → [Humain: {*} *x2].

Schéma Fuite
importance: 0.5;
spécificité: 0.5;
[SEnfuir; prédicat: vrai] → (agent) → [Humain: {*} *x2].

FinAttribut Description

Attribut ÉtatsIncidents

Schéma DevenirRiche
importance: 0.5;
spécificité: 0.4;
[SEnrichir; prédicat: vrai] → (expérienceur) → [Humain: {*} *x2].

FinAttribut ÉtatsIncidents

RelationsIntraSchéma

Description.Enlèvement -> Description.Séquestration:
type: précédenceTemporelle;
importance: 1.0.

Description.Séquestration -> Description.LibérationOtage:
type: précédenceTemporelle;
importance: 1.0.

Circonstances.ÊtreUnProcheDe -> Description.RemiseDeRançon:
type: motivation;
importance: 0.75.

Description.RemiseDeRançon -> Description.LibérationOtage:
type: cause;
importance: 0.5.

Description.RemiseDeRançon -> ÉtatsIncidents.DevenirRiche:
type: cause;
importance: 0.5.

Description.DemandeDeRançon -> Description.RemiseDeRançon:
type: précédenceTemporelle;
importance: 1.0.

Description.RemiseDeRançon -> Description.Fuite:
type: précédenceTemporelle;
importance: 1.0.

FinRelationsIntraSchéma

FinSchéma Kidnapping

Fig. 4.10 - Représentation sous forme linéaire d'un schéma représentant une situation de kidnapping destiné à extorquer de l'argent

Les graphes se voient associer comme les références un poids caractérisant leur importance et un poids rendant compte de leur spécificité. Les principes énoncés à propos de ces poids au niveau des références restent globalement valides pour les graphes. Nous nous contenterons de signaler que la méthode présentée pour l'évaluation automatique de la spécificité est plus difficilement applicable ici dans sa forme première. Les schémas sont des entités à part entière de la mémoire pragmatique. Il en résulte qu'un schéma n'est pas dupliqué à chaque fois que l'on y fait référence. On peut donc déterminer aisément combien de références y sont faites. La situation est différente pour les graphes de contraintes. Ils sont propres aux schémas au sein desquels ils sont présents et existent donc en autant d'exemplaires séparés qu'il y a de références les concernant. Le recensement des références faites à un graphe est donc très coûteux puisqu'il oblige en théorie à parcourir tous les schémas de la mémoire et à réaliser pour chacun des projections afin de vérifier la compatibilité des graphes.

Sachant que l'intérêt de ces graphes, du point de vue de la sélection, réside dans les types de concept qu'ils contiennent et non dans les graphes eux-mêmes, nous proposons de retenir comme valeur de spécificité pour un graphe de contrainte la valeur de spécificité la plus forte parmi celles de ses types de concept. La valeur de spécificité d'un type de concept dans un schéma est définie dans le paragraphe suivant sur les rôles.

Rôles

La dernière composante d'un schéma qu'il convient d'introduire est la notion de rôle. Un rôle est chargé de représenter une entité de la situation incarnée par le schéma et donc

d'assurer le lien entre toutes les références qui sont faites à cette entité au travers de concepts présents dans les graphes conceptuels intervenant dans la définition du schéma, que ce soit en tant que contraintes ou bien en tant qu'en-têtes de schémas. Pour assurer ce lien d'identité, on fait appel comme on l'a vu précédemment aux variables de coréférence entre concepts.

Un rôle se compose de façon minimale d'une relation et d'un graphe conceptuel réduit à un seul concept générique. Ce concept porte la variable de coréférence qui identifie l'entité représentée au sein du schéma. Son type est la plus faible généralisation incluant tous les types des concepts qui coréfèrent par l'intermédiaire de cette variable[1]. La relation vise à attribuer une forme de fonction globale à l'entité que figure le rôle. Pour les entités apparaissant dans l'en-tête du schéma, on prend la relation qui les lient au prédicat de cet en-tête. Pour les autres, tout dépend du contexte de création du schéma. Si celle-ci intervient de manière automatique, comme dans le cas de l'abstraction des UTs agrégées de la mémoire épisodique, il sera certainement très difficile de synthétiser une relation exprimant la fonction globale de l'entité. On peut alors se contenter de faire appel à la relation générique qui constitue le sommet de la hiérarchie des relations (type de relation *Relation*), ainsi que cela est fait dans le schéma de la figure 4.10 pour les rôles correspondant aux variables *x3* et *x4*. Si les schémas résultent d'une modélisation manuelle, rien ne s'oppose à l'attribution d'une relation plus spécifique.

En plus de ces deux constituants de base, les rôles se voit également attribuer un type par défaut. Celui-ci permet de préciser quel est le type de concept prototypique des concepts occupant ce rôle. Le type de ce dernier, en généralisant celui de toutes les occurrences du rôle, perd en effet en spécificité, donc en contenu informationnel. Le type par défaut est le moyen de contrebalancer cette perte. Lorsqu'un schéma est déclenché, en l'absence d'indication venant du texte sur la nature exacte d'une entité de la situation, elle est supposée avoir comme type le type par défaut. Cela permet en particulier de réaliser certaines inférences et donc, de faire progresser la compréhension. Ces inférences seront éventuellement remises en question si des précisions contradictoires sont apportées

[1] Tous les schémas de la mémoire épisodique n'ont pas le même niveau de généralité (cf. §4.2.2). Néanmoins, toutes les situations ne sont pas décrites de façon systématique pour tous les différents niveaux de généralité existant. La définition d'un schéma peut donc nécessiter de faire référence à un schéma se situant à un niveau de généralité différent de celui qui caractérise le schéma défini. Ce phénomène est particulièrement fréquent lorsque les schémas sont le produit d'un apprentissage. Par exemple, un schéma évoquant ce qui se passe lorsque l'on prend des vacances peut mentionner que l'on commence par voyager, sans apporter plus de précisions sur le moyen de transport employé compte tenu de la multiplicité des cas observés. Dans le même temps, il indique que l'on visite des musées, ce qui pourrait se généraliser, pour être compatible avec le niveau du précédent point, en la pratique d'activités culturelles. Cependant, si la visite de musée est la seule activité de ce type à avoir été rencontrée, c'est elle qui figure dans le schéma et les concepts de celui-ci n'ont pas de ce fait toujours le même type pour désigner des entités identiques.

ultérieurement par le texte. Compte tenu du caractère prototypique du type par défaut, on peut néanmoins estimer que le mécanisme est plus souvent avantageux qu'il n'est inefficace.

De manière parallèle à ce qui prévaut pour les références et les graphes de contrainte, les rôles comportent enfin deux poids, caractérisant d'une part l'importance du rôle, et d'autre part sa spécificité par rapport au schéma défini. La définition de ces notions reste identique à celle donnée pour les références mais comme pour les graphes de contrainte, il est nécessaire d'apporter des compléments sur la méthode d'évaluation automatique de la spécificité. En effet, les types de concept, qui sont les éléments informationnellement distinctifs des rôles, ne sont pas non plus des entités autonomes de la mémoire pragmatique. En revanche, et au contraire des graphes de contrainte, ils possèdent un tel statut en dehors de celle-ci. La mémoire conceptuelle s'articule de fait autour du treillis qu'ils forment. En outre, la notion de base canonique préconise que chaque type de concept renvoie à un certain nombre de schémas (cf. §4.1) afin de rendre compte de son utilisation en contexte. En systématisant cette relation, on peut donc savoir, comme pour un schéma, combien de schémas référencent un type de concept donné. Dès lors, l'application de la mesure de spécificité présentée plus haut est directe pour les rôles, et au delà, pour les graphes de contraintes.

La répertoriation des schémas au niveau des types de concept s'opère plus précisément de la façon suivante : pour un rôle, la référence au schéma est faite à partir du type de concept constituant le type par défaut du rôle; pour un graphe de contrainte, elle est réalisée à partir du type de concept le plus spécifique (dans la hiérarchie des types) apparaissant dans le graphe en dehors des concepts de celui-ci représentant des entités définies comme rôles.

2.2.2. Structure de la mémoire pragmatique

La mémoire pragmatique se présente avant tout comme un vaste réseau de schémas. Comme on a pu le voir lors de l'examen de leur structure, ceux-ci se définissent essentiellement par les références qu'ils font à d'autres schémas. Ces liens de référence forment ainsi la trame de base du réseau de la mémoire pragmatique. Les schémas présentés ici reprennent sur ce point les principes que Schank a cherché à concrétiser au travers des MOPs. Plutôt que de constituer de grosses unités monolithiques tels que les plans décrits dans [Schank & Abelson 1977], il s'agit de former des unités plus petites, caractérisant des situations plus élémentaires, mais pouvant être assemblées pour représenter une grande variété de situations complexes. Les schémas diffèrent en cela des UTs agrégées, qui sont conçues comme des entités plus autonomes puisque ne faisant appel pour leur définition qu'à la représentation sémantique de propositions.

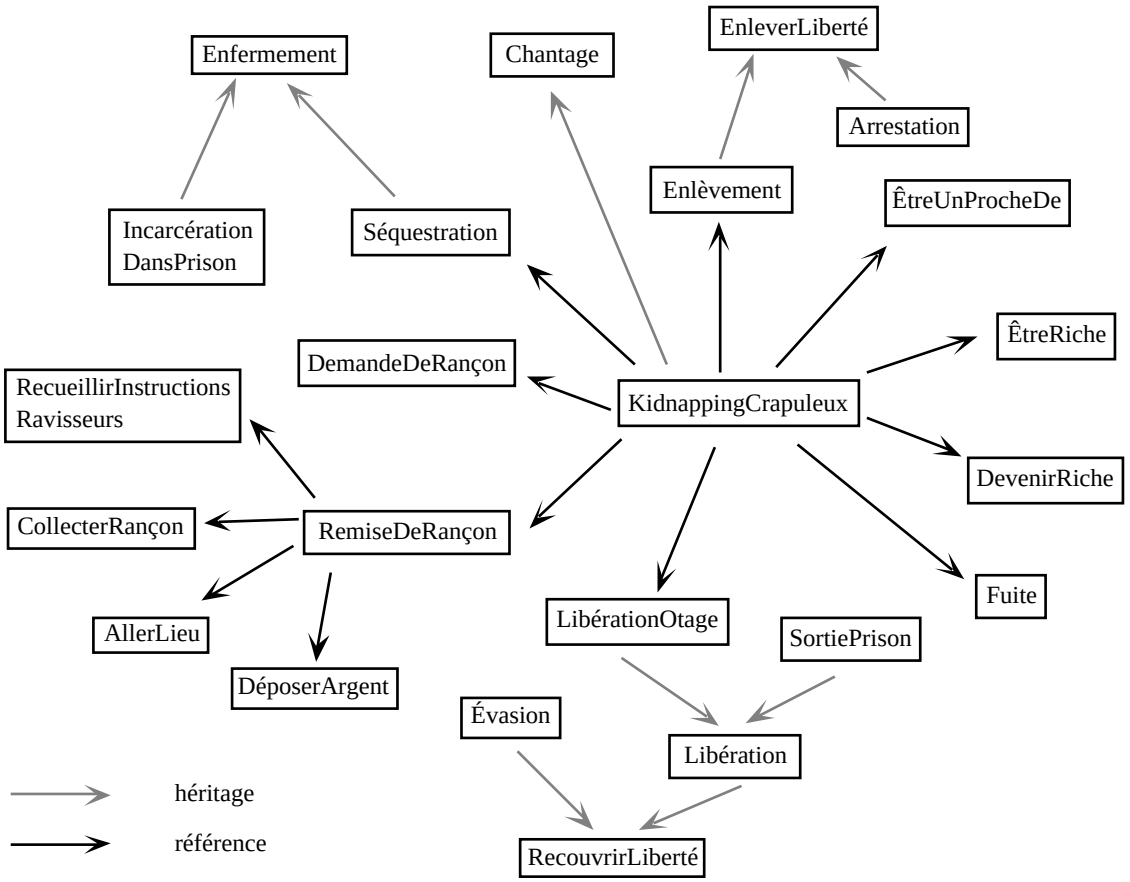


Fig. 4.11 - Un aperçu des relations au sein de la mémoire pragmatique

La relation de référence introduit par ailleurs une première forme d'ordre sur les schémas. Elle est en effet proche de la relation *partie_de* : dire que le schéma A référence le schéma B revient à dire que B est un morceau de A. Cela signifie également que A ne peut apparaître dans la définition de B : on ne décrit pas une situation par des événements qui sont eux-mêmes composés de cette même situation. En suivant les liens de référence, on parvient donc à des schémas terminaux, supposés représenter des événements suffisamment élémentaires pour ne pas être décrits sur le plan pragmatique. Ces schémas ne possèdent ainsi pas de corps. Ils sont uniquement formés d'un en-tête et des rôles correspondant aux entités introduites par celui-ci. Bien entendu, le statut de schéma terminal est le résultat du choix arbitraire d'un niveau de granularité, susceptible d'ailleurs d'évoluer dans un contexte où l'apprentissage a sa place : le fait de manger un aliment est vu comme un événement élémentaire dans une situation décrivant ce qui se passe lorsque l'on va au restaurant; il est en revanche lui-même détaillé si on l'envisage davantage sous un angle physiologique (mastication, déglutition, ...).

Comme le montre la figure 4.11, la mémoire pragmatique est également structurée par une relation d'héritage entre schémas. Lorsqu'un schéma B hérite d'un schéma A, les éléments apparaissant dans la définition de B (aussi bien les schémas référencés, les

graphes de contraintes que les rôles) sont soit des spécialisations d'éléments déjà présents au niveau de A, soit de nouveaux éléments ajoutés pour apporter des précisions. Dans le schéma de la figure 4.10, le schéma *DemandeDeRançon* est ainsi une spécialisation du schéma *ExigerAvantage*, plus général et présent au niveau du schéma *Chantage*, dont hérite le schéma *KidnappingCrapuleux*. La spécialisation d'un schéma ne se concrétise d'ailleurs pas toujours directement par un schéma mais apparaît parfois implicitement sous la forme de son contenu. Ainsi les schémas *Enlèvement* et *Séquestration* peuvent être vus comme des constituants d'un schéma virtuel *FairePressionParEnlèvement* qui serait la spécialisation du schéma *FairePression*, figurant au niveau du schéma *Chantage*. Enfin, si un élément de A n'est pas explicitement spécialisé par B, alors il y est implicitement présent par héritage. Il ne s'agit ici que d'un héritage simple : un schéma n'hérite que d'un seul schéma. On évite ainsi d'avoir à imposer des mécanismes de gestion de conflit, toujours un peu artificiels et compliqués.

Plus globalement, la présence de cette relation de généralisation participe à la volonté de disposer de petites unités qui puissent être les plus adaptées possible à la fois au degré de granularité et au degré de généralité des situations évoquées dans les textes.

2.2.3. Liens avec d'autres représentations des connaissances sur les situations

La mémoire dynamique

Ainsi que nous l'avons souligné ci-dessus, notre conception de la notion de schéma s'apparente à celle promue par Schank au travers des MOPs. La mémoire dynamique schankienne et notre mémoire pragmatique diffèrent néanmoins quant à la variété des objets qu'elles mettent respectivement en jeu. Nous avons vu au chapitre 1 que la mémoire dynamique différencie des types de connaissances en fonction de leurs relations vis-à-vis d'autres types de connaissances : les MOPs sont composés de scènes et de façon similaire, ils forment des méta-MOPs. Par ailleurs, les MOPs sont généralisés par des U-MOPs qui font référence à des scènes généralisées. Que ce soit pour les relations de composition ou pour les relations de généralisation, la source et la cible d'une relation sont donc toujours de natures différentes.

La différenciation de toutes ces structures nous apparaît en pratique assez difficile à opérer. La notion de situation est à la base difficile à caractériser et il ne nous semble de ce fait pas nécessaire de la raffiner de multiples façons en l'absence de critères précis de définition. C'est pourquoi nous n'avons retenu ici qu'un type de schéma unique. Les structures qui composent ou qui généralisent un schéma sont donc des schémas de même

nature que ce schéma. Nous avons vu que cela est vrai y compris pour les schémas terminaux du point de vue de la relation de référence, c'est-à-dire les schémas ne possédant pas de corps. Ils pourraient être assimilés à ce que sont les scènes des MOPs mais du fait des évolutions de la mémoire induites par l'apprentissage, le caractère terminal de ces schémas n'a pas à être fixé de façon définitive par un statut particulier car il est potentiellement transitoire.

Les graphes conceptuels

La présentation du formalisme des graphes conceptuels que nous avons faite au §4.1 rappelle que dès son origine, celui-ci a été impliqué dans la représentation des connaissances pragmatiques. Dans [Sowa 1984], Sowa fait référence de façon explicite aux schémas de Schank lorsqu'il évoque les graphes, qu'il appelle d'ailleurs schémas, qui peuvent être rattachés aux types de concept afin de rendre compte de l'utilisation qui en est habituellement faite. Néanmoins, les exemples qu'il donne sont plutôt simplistes et ne dépassent guère le niveau des en-têtes des schémas que nous avons présentés ici. Par ailleurs, Sowa voit les schémas comme un moyen d'étendre la définition des types de concept mais ne semble pas leur accorder le statut de connaissances autonomes, ayant leur intérêt propre. Cela justifie certainement l'absence de structure des schémas qu'il fournit comme exemples.

Notre choix sur ce point est moins guidé par les connaissances sémantiques. Les situations sont représentées de façon autonome dans une mémoire spécifique et par des schémas structurés en fonction des impératifs propres à ce type de modélisation. Des liens existent entre les deux dimensions mais celles-ci sont nettement séparées. Les schémas font référence aux types de concepts pour désigner les objets, les actions et les états qu'ils organisent. Réciproquement, on répertorie au niveau des types de concept les schémas dans lesquels ils interviennent, ce qui permet notamment de conserver la modélisation mise en avant par Sowa.

Tout en conservant cette séparation, on peut se demander si l'utilisation des graphes conceptuels, cette fois-ci en tant que pur formalisme de représentation, n'aurait pas pu être étendue à l'ensemble d'un schéma, plutôt que de faire de ceux-ci des entités hybrides. Comme en témoigne la figure 4.12, Sowa a fait lui-même des propositions dans ce sens en généralisant le typage des contextes. Dans notre cas, rien n'empêcherait ainsi d'avoir des contextes de type Circonstances, Description ou ÉtatsIncidents.

Nous ne nous sommes pas engagé dans cette voie pour le moment dans la mesure où l'adoption d'un formalisme de représentation ne se justifie que si des opérations de manipulation qui lui sont propres peuvent être utilisées. Or, il n'existe pas encore de

définition véritablement établie pour les opérations que nous avons présentées au §4.1 lorsqu'elles s'appliquent à des graphes contenant des contextes. En leur absence, il était plus simple, à la fois sur le plan de la représentation et celui de la manipulation, de s'en tenir à une solution moins générique.

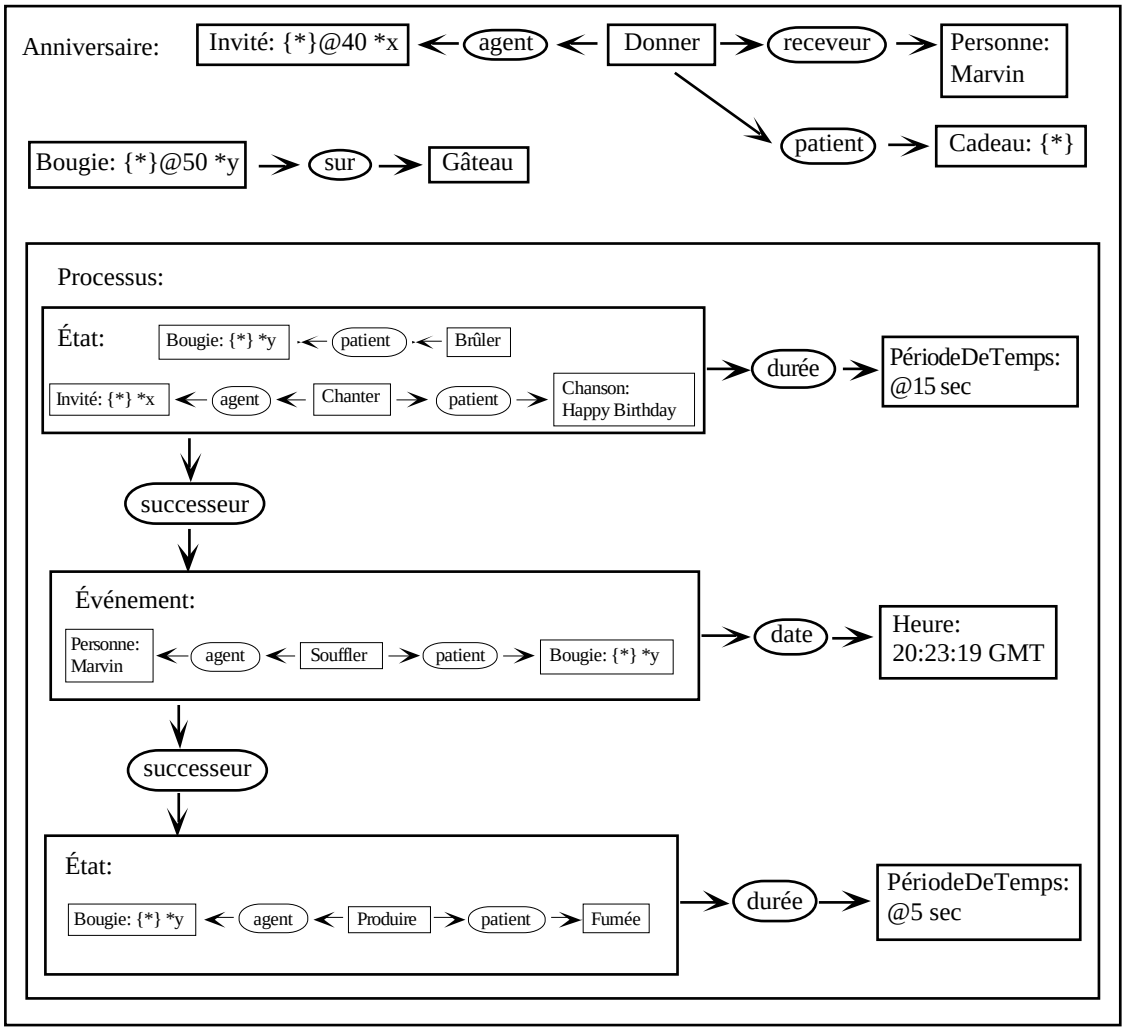


Fig. 4.12 - La représentation d'une occurrence de situation sous forme de graphes conceptuels (d'après [Sowa 1992])[1]

2.2.4. Implémentation

Nous avons implémenté la mémoire pragmatique ainsi que les schémas qui la composent en Smalltalk à partir de la plate-forme de graphes conceptuels présentée au §4.1. La construction d'un schéma peut être réalisée soit par l'intermédiaire d'une interface de programmation, soit en utilisant, comme pour les graphes conceptuels, une forme linéaire textuelle pour laquelle nous avons écrit un compilateur et un générateur. On

[1] Chaque boîte représente ici un contexte, dont le type figure dans son coin supérieur gauche.

trouvera à l'annexe B la grammaire de cette forme linéaire (le schéma de la figure 4.10 s'y conforme).

À côté d'une sélection par propagation d'activation reposant sur la mesure de spécificité des schémas les uns par rapport aux autres (cf. chapitre 6), nous avons conservé une indexation plus classique dans l'optique d'une utilisation la plus large et la plus diversifiée possible de la mémoire pragmatique. Chaque schéma est ainsi indexé par un triplet de types prédicat – relation – objet extrait de son en-tête. Des opérations de recherche permettent de retrouver un seul ou un ensemble de schémas à partir de la donnée d'un triplet similaire. Pour chaque type du triplet, on peut spécifier une égalité stricte ou autoriser un sur-type ou un sous-type.

Des interfaces graphiques facilitent enfin la gestion de la mémoire pragmatique aussi bien en ce qui concerne la création ou la destruction de schémas que pour leur indexation, leur recherche ou leur visualisation. Un aperçu de ces outils est donné à l'annexe B.

Récapitulatif

Ce chapitre nous a permis d'exposer ce sur quoi s'appuie MLK, c'est-à-dire les connaissances sémantiques, ainsi que ce qu'il cherche à construire, en l'occurrence les connaissances pragmatiques sur les situations.

Les connaissances sémantiques sont représentées en utilisant le formalisme des graphes conceptuels, qui propose à la fois un formalisme de représentation, des opérations de manipulation et une façon de structurer les connaissances. Nous complétons les principes généraux édictés par ce formalisme par un certain nombre de précisions concernant la façon dont nous l'avons utilisé pour modéliser les connaissances sémantiques. Le problème de la multiplicité des points de vue a ainsi particulièrement retenu notre attention. Nous proposons à ce sujet d'adopter une conception que l'on pourrait qualifier de métonymique : il s'agit de dissocier systématiquement l'être ou la substance d'une entité de son rôle ou de sa fonction. Cette ligne directrice nécessite toutefois que soit définie la colonne vertébrale que constitue cette substance. Nous en montrons un exemple concernant les types de concept verbaux.

Les connaissances pragmatiques sur les situations sont représentées quant à elles par un type unique de schéma, inspiré de l'idée, présente dans les MOPs de Schank, de petites unités facilement agençables pour représenter des situations complexes. Ces schémas sont stockés dans une mémoire spécifique au sein de laquelle ils se définissent

essentiellement par les relations qu'ils entretiennent entre eux : un schéma est un assemblage de références faites à d'autres schémas, assemblage venant spécialiser un schéma plus général. L'originalité des schémas que nous présentons réside dans la combinaison réalisée avec les graphes conceptuels, première étape en direction d'une représentation entièrement à base de graphes conceptuels. Cette combinaison offre l'avantage de pouvoir entretenir un lien particulièrement naturel avec les représentations sémantiques.

Chapitre 5

Les représentations de texte

L'objet de ce chapitre est de présenter les représentations de texte manipulées par MLK. Situées à la frontière entre analyse de textes et apprentissage de connaissances, elles jouent à cet égard un rôle central dans MLK. Nous abordons donc en détail la forme qu'elles revêtent ainsi que les problèmes plus généraux qu'elles posent quant à ce que doit être leur contenu relativement aux contraintes imposées à la fois par l'analyse thématique et l'apprentissage de connaissances sur les situations développés ici.

1. Nature des représentations de textes

Dans un système tel que MLK, qui cherche à mêler étroitement analyse de textes et apprentissage de connaissances, les représentations de textes occupent de façon évidente une position centrale. Elles sont en effet à la fois le produit de l'analyse de textes et le matériau à partir duquel l'apprentissage opère. Elles se trouvent donc à l'interface entre les deux dimensions du système.

Dans MLK, la notion qui sert de point de jonction entre analyse de textes et apprentissage de connaissances est celle de situation, ou plus précisément la représentation qui en est faite. Nous avons vu dans les chapitres précédents que cette représentation se définit, quel que soit le niveau d'abstraction ou de généralité auquel on se place, comme le rassemblement des événements (au sens large, c'est-à-dire aussi bien des actions, des états que des processus) et des acteurs caractérisant la situation représentée. Elle est désignée sous l'appellation générique d'Unité Thématique (UT).

Les représentations de texte manipulées dans le cadre de MLK sont donc naturellement centrées autour de la notion de situation. Elles sont plus précisément le produit de la structuration thématique du résultat de l'analyse sémantique des textes. À ce titre, elles font apparaître les différentes situations évoquées par un texte, définissent quels éléments du texte se rapportent à quelle situation et enfin caractérisent la façon dont ces différentes situations s'organisent les unes par rapport aux autres. Traduits en termes de représentation, les deux premiers points se matérialisent par la construction d'UTs tandis que le dernier conduit à déterminer le rôle de chacune de ces UTs dans la représentation globale ainsi que ses relations avec les autres UTs de cette représentation.

Dans une UT, les représentations sémantiques des événements et de leurs acteurs ne sont pas disposées pêle-mêle. Les UTs en général, et donc celles des représentations de texte en particulier, possèdent de fait une structure destinée à mettre en évidence le rôle des différents événements qui les composent. Compte tenu de l'accent mis sur la dimension proprement thématique de l'analyse des textes, il ne s'agit pas du point principal développé dans notre travail. Cette structuration s'avère néanmoins particulièrement intéressante pour la construction finale d'une représentation abstraite et générale d'une situation telle qu'elle est incarnée par les schémas que nous avons décrits au chapitre précédent. L'existence de cette relation entre schémas et UTs des représentations de texte amène ainsi à s'inspirer fortement, ainsi que nous allons le voir, de la structure des premiers pour définir celle des secondes.

2. Nature des textes

La prééminence de la notion de situation qui prévaut dans MLK, issue de la volonté d'apprendre des connaissances sur des situations prototypiques du monde, nous a conduit à adopter comme support de travail un type de textes particulier. Nous travaillons en effet à partir de textes de style narratif tels que celui de la figure 5.1. Nous avons en fait retenu des textes qui d'une part, évoquent ces situations, ce qui exclut par exemple les textes techniques, et d'autre part, le font en mettant en avant leur composante événementielle, ce qui laisse plus généralement de côté les textes expositifs, les textes argumentatifs et les textes descriptifs.

Rozenne (10 ans) en avait assez ...

Une fillette de 10 ans, Rozenne, demeurant à Lyon, a été portée disparue pendant 24 heures, mobilisant une partie des services de police lyonnais, avant d'être retrouvée hier matin endormie dans un immeuble.
Partie normalement à l'école lundi matin, l'enfant n'y arriva jamais. Vers 14h, les parents prévenaient les services de police et des patrouilles étaient organisées.
En fait, la petite Rozenne avait laissé une lettre chez ses parents avant de disparaître, lettre dans laquelle elle écrivait notamment : "J'en ai assez, je pars avec mon lapin, je reviendrai quand je serai majeure".
Et c'est effectivement avec son lapin, qui dormait à ses côtés, que Rozenne a été retrouvée.

Fig. 5.1 - Récit de presse illustrant le type des textes considérés dans MLK
(repris de [Adam 1984])

Cette seconde contrainte est particulièrement forte puisqu'il est en pratique difficile de trouver des textes dont le style est purement narratif. Même au sein des récits, la composante événementielle est le plus souvent noyée au milieu de descriptions qui constituent en réalité l'essentiel du texte. La façon de faire la plus générale consisterait à séparer automatiquement ce qui est du domaine de la relation des événements et ce qui est du ressort de la description des personnages, des objets et des lieux. Chaque portion de texte pourrait de cette façon donner lieu à un type d'analyse spécifique, se prolongeant par un type d'apprentissage également adapté. Alors que les parties typiquement narratives permettent de faire progresser l'apprentissage des situations, les parties descriptives sont en effet plus spécifiquement favorables à l'apprentissage des concepts.

En l'absence de telles capacités, nous avons opté en faveur de textes autant que possible de style narratif pur mais de ce fait même, relativement courts. Le prototype de ces textes est la dépêche d'agence de presse, telle que celle de la figure 5.1. De courts récits comme les anecdotes forment aussi un matériau facilement exploitable comme le montre l'exemple de la figure 5.2.

3. Forme des représentations de textes

3.1. L'unité de base

Le constituant élémentaire retenu pour former les représentations de texte est la proposition. Il nous a semblé en effet que celle-ci constitue le meilleur compromis entre la capacité d'expression offerte par l'articulation des mots et la nécessité de représenter les situations de façon homogène, en s'appuyant sur des unités facilement comparables. Le choix d'un apprentissage fondé sur l'agrégation d'unités similaires impose en effet que celles-ci soient les plus homogènes possibles afin de pouvoir comparer des choses véritablement de même nature.

La proposition est en outre une unité possédant un ancrage syntaxique suffisant pour que sa reconnaissance, sans être exempte de certaines incertitudes, ne soit pas trop difficile. Une grammaire de base [Bonnard 1981] la définit ainsi comme "le groupe formé par un sujet et un verbe, et tous les mots se rapportant directement ou non au sujet ou au verbe"[1]. Nous verrons à la fin de ce chapitre que nous étendons dans les faits un peu cette

[11] Les définitions que l'on peut trouver de la proposition, dans un contexte grammatical, se recoupent fortement. Le Dictionnaire du français contemporain la définit comme "une unité constitutive d'un énoncé, composée en général d'un groupe nominal et d'un groupe verbal et formant une partie d'une phrase, sinon la phrase toute entière" tandis que J. Marouzeau, dans le Lexique de la Terminologie linguistique, en parle comme d'un "énoncé constitué essentiellement par un prédicat, ordinairement verbal,

définition; mais en première approximation, elle peut être retenue comme tout à fait représentative. Parmi les possibilités envisageables, la phrase aurait constitué une unité encore plus facilement reconnaissable mais son contenu est dans le même temps beaucoup trop dépendant de la manière dont le rédacteur exprime sa pensée pour répondre aux exigences d'homogénéité posées ici.

Au niveau où nous nous situons dans le cadre de MLK, nous ne manipulons pas directement des propositions, au sens grammatical du terme, mais plutôt la représentation sémantique de ces propositions. Celle-ci prend la forme de graphes conceptuels puisque c'est le formalisme adopté pour représenter les connaissances sémantiques de MLK (cf. chapitre précédent). Une proposition mise en évidence par l'analyse syntaxique donne ainsi lieu à la construction d'un graphe conceptuel. On passe des mots aux concepts et des fonctions syntaxiques aux relations casuelles. Un tel passage implique également la résolution des coréférences si l'on veut être capable d'associer un concept à un pronom personnel par exemple. Nous reviendrons un peu plus loin dans ce chapitre sur ce problème de l'analyse thématique et de la résolution des coréférences, que l'on peut globalement désigner comme le problème de la construction d'une représentation pré-thématique des textes.

3.2. Structure des représentations de texte

Comme nous l'avions laissé entendre en préambule de ce chapitre, les représentations sémantiques des propositions des textes sont structurées en fonction des situations évoquées par le texte auxquelles elles se rapportent. Une représentation de texte dans MLK, appelée également épisode, est donc formée d'un ensemble d'UTs, chacune étant un regroupement de graphes conceptuels.

La figure 5.3 montre la structuration en UTs de l'épisode représentant le texte de la figure 5.2. Elle met ainsi en évidence les trois situations successivement évoquées par le texte :

- une séance de dédicace dans un grand magasin,
- une tentative de meurtre,
- un séjour à l'hôpital.

Au niveau des UTs, on ne fait figurer ici que les graphes représentant les propositions explicites du texte. Nous verrons par la suite que certaines inférences sont possibles et se

---

mais qui peut être aussi nominal (quant à ce qu'il a dit, *sornettes !*), accompagné habituellement d'un sujet et de termes rapportés l'un à l'autre".

traduisent alors par l'ajout de nouveaux graphes. Les graphes conceptuels sont représentés dans les UTs par le numéro de la proposition qui leur correspond dans le texte. Chacune de ces propositions est délimitée de façon approximative au niveau la figure 5.2.

Il y a quelques années, **[**je me trouvais dans un grand magasin de Harlem**](1)**, **[**entouré de quelques centaines de personnes**](2)**. **[**J'étais en train de dédicacer des exemplaires de mon livre "Stride toward Freedom"**](3)**, **[**qui relate le boycottage des autobus de Montgomery en 1955-56**](4)**. Soudain, tandis que **[**j'apposais ma signature sur une page**](5)**, **[**je sentis quelque chose de pointu s'enfoncer brutalement dans ma poitrine**](6)**. **[**Je venais d'être poignardé à l'aide d'un coupe-papier, par une femme**](7)** **[**qui devait être reconnue folle par la suite**](8)**. **[**On me transporta d'urgence à l'Hôpital de Harlem**](9)** où **[**je restai de longues heures sur un lit**](10)** tandis qu'**[**on faisait mille préparatifs**](11)** **[**pour extraire l'arme de mon corps**](12)**.

Fig. 5.2 - Extrait de "Révolution non-violente" par Martin Luther King

La représentation d'un texte fait non seulement apparaître une structuration en situations mais elle met également en évidence les relations existant entre les situations, donc entre les UTs. Il s'agit plus précisément de relations de suivi thématique rendant compte de la façon dont s'effectue le passage d'un thème à un autre dans un texte.

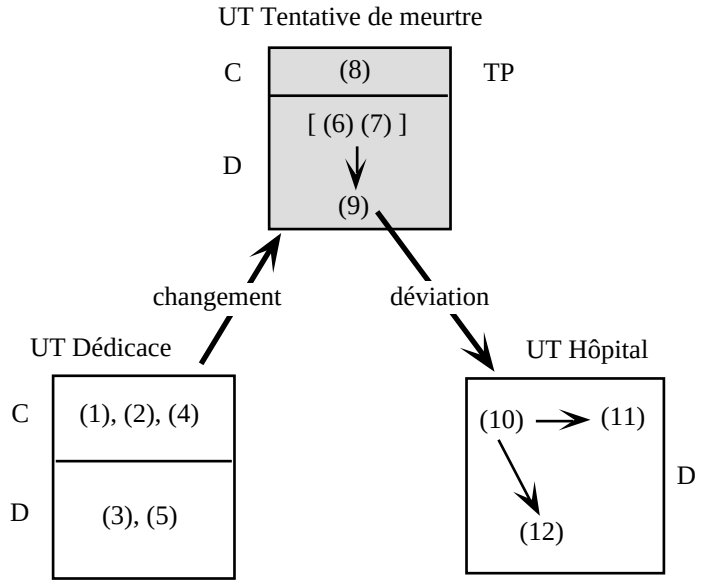


Fig. 5.3 - Représentation de texte construire à partir du texte de la figure 5.2[1]

[1] Au niveau de la représentation sémantique, les propositions 6 et 7 sont considérées comme équivalentes. Elles ne donnent donc lieu qu'à un seul graphe conceptuel. Cette équivalence est détectée via le graphe de définition associé à chacun des types de concept. La même équivalence pourrait éventuellement être établie entre les propositions 3 et 5. Leurs sens sont néanmoins plus distants.

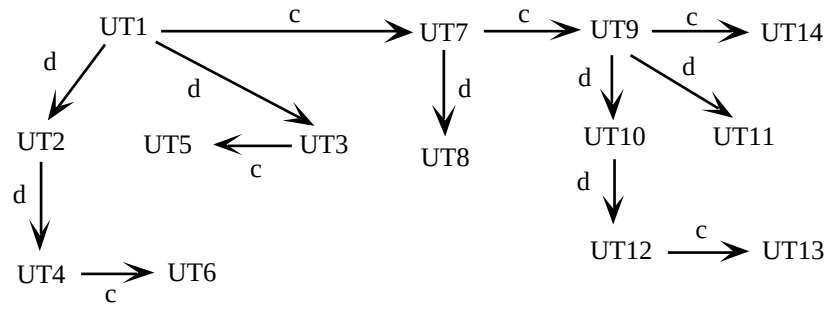
Mais ces relations sont illustratives dans le même temps du rapport existant entre les deux thèmes en dehors de leur apparition dans un texte. [Grau 1983] et [Dahlgren 1993], parmi d'autres auteurs, soulignent la dépendance étroite existant entre ces relations et les connaissances sur les situations abordées. C'est pourquoi on peut parler à propos de ces relations de suivi thématique de relations entre situations. Un texte ne contient certes pas les informations permettant d'établir systématiquement quelles relations existent entre toutes les situations qu'il évoque, prises deux à deux. C'est d'ailleurs ce qui distingue une représentation de texte d'une connaissance générale et abstraite telle que les schémas. Mais les relations de suivi thématiques mises en évidence du fait de l'agencement des situations opéré par le texte nous renseignent tout de même sur une partie des relations effectives entre situations.

En accord avec [Grau 1983], nous avons retenu deux grandes relations, sans rechercher une différenciation plus poussée que ne supporteraient pas les moyens d'analyse volontairement limités que nous souhaitons mettre en œuvre ici :

- *déviation thématique* : la présence de cette relation signifie que l'on passe d'un thème à un thème qui lui est proche, d'où le terme de déviation. Plus exactement, le second thème constitue en général le développement d'un point particulier du premier. Il existe donc entre les deux situations considérées une relation allant au delà de la simple présence de personnages communs. Cette relation est en quelque sorte équivalente à la relation de référence présente entre les schémas de la mémoire pragmatique.
  Sur le plan des représentations, une telle déviation signifie qu'une UT est détaillée par une autre UT. Dans notre exemple de la figure 5.3, la situation de l'hôpital est ainsi considérée comme étant une déviation de la situation de tentative de meurtre du fait de l'existence d'une relation de causalité entre elles : le séjour de Martin Luther King à l'hôpital résulte d'une blessure causée par la tentative de meurtre.
  Une déviation est plus spécifiquement attachée à l'un des graphes de l'UT source. Elle représente le développement sur le plan pragmatique de la situation évoquée par ce graphe. Ici, l'UT Hôpital est reliée au graphe pivot (9) exprimant que Martin Luther King est conduit à l'hôpital.

- *changement thématique* : cette relation caractérise le passage d'un thème à un thème qui ne lui est pas lié, ou tout du moins qui ne lui est pas lié que de façon intrinsèque. Plus précisément ici, cette relation s'identifie à l'introduction d'une nouvelle situation, donc d'une nouvelle UT. Dans l'exemple de la figure 5.3, il existe ainsi un changement thématique entre la situation de la séance de dédicace et celle de la tentative de meurtre du fait que celles-ci ne sont liées que de façon contingente, fort heureusement pour les écrivains.

Les relations de suivi thématique donnent aux représentations de texte une structure d'arbre, ainsi que l'illustre la figure 5.4. Une situation S2, qui est la déviation d'une situation S1, ne peut pas par exemple avoir S1 comme déviation : un détail d'une situation ne peut pas être développé en faisant référence à cette même situation, que ce soit directement ou indirectement. S2 ne peut pas non plus être la source d'une relation de changement de thème vers S1 : le fait de revenir à S1 après avoir évoqué S2 est seulement indicatif de l'inclusion de S2 dans S1.



d : déviation
c : changement de thème

Fig. 5.4 - Structure arborescente d'une représentation de texte

Cette structure confère à une UT au moins de chaque épisode le statut spécial de thème principal (TP). Intuitivement, le thème principal d'un texte est le sujet principal, central d'un texte. Sa définition formelle reprend cette notion de centre : le thème principal est l'UT qui est la plus proche de la racine de la structure arborescente de l'épisode et qui est la source du plus grand nombre de relations de déviation thématique. Cette définition est davantage une indication de ce qu'est le thème principal qu'un moyen opératoire de reconnaissance infaillible. Dans le texte de figure 5.2, la position centrale de l'UT Tentative de meurtre au sein du réseau des relations thématiques la désigne ainsi comme thème principal.

Le thème principal d'un texte est en outre caractérisé par une propriété spécifique du point de vue du suivi thématique : le thème principal d'un texte est le seul thème pouvant être réintroduit à tout endroit de ce même texte. Cette propriété, exploitée dans [Grau 1983] pour contraindre la reconnaissance de l'enchaînement des thèmes, peut être utilisée ici comme un moyen de reconnaissance de la nature des UTs. Une situation évoquée à plusieurs endroits différents dans un texte constitue probablement le thème principal de ce texte.

Lorsque le texte considéré est assez long, la structure de la représentation de texte prend souvent l'apparence d'un ensemble de sous-arbres liés de proche en proche par des

relations de changement thématique, comme le montre la figure 5.4. La représentation de texte possède alors plusieurs thèmes principaux. En pratique, elle en compte autant que de sous-arbres. En supposant qu'il existe un seul thème englobant pour un texte[1], la présence d'un seul thème principal ou de plusieurs dépend en fait étroitement de ce qui est explicité dans le texte.

Sur le plan représentationnel, il est tout à fait possible d'avoir une structure d'arbre complète : une situation racine suffisamment générale peut recouvrir l'ensemble d'un texte et être détaillée par différentes déviations de thème, si besoin est de façon récursive. Du fait de l'approche que nous avons retenue, notre représentation des textes reste toutefois guidée par les informations apportées par les textes plus que par les connaissances du système. Pour qu'une UT soit créée, il faut donc qu'elle possède un ancrage dans le texte sous forme de propositions faisant référence à la situation représentée par cette UT. Si le thème le plus général du texte est explicité, on peut construire, à condition de le discerner, une UT le représentant et y rattacher les autres UTs venant le détailler. Si tel n'est pas le cas, le lien entre ces UTs n'est pas perçu. Elles ne sont alors liées que par des relations de changement de thème et l'on obtient une structure similaire à celle de la figure 5.4. Précisons que dans ce cas, la règle présentée plus haut concernant la possibilité de réintroduction des thèmes principaux ne s'applique a priori plus qu'au sein du sous-arbre contenant le thème considéré.

3.3. Structure des Unités Thématiques

3.3.1. Description

La structure interne des UTs est proche de celle des schémas, au point que les UTs peuvent être considérées comme des instances de schémas n'ayant pas encore d'existence. On retrouve donc à leur niveau la même structuration en trois grands attributs qui existe pour les schémas. Au lieu de regrouper un ensemble de références vers d'autres schémas, chacun d'entre eux rassemble ici un sous-ensemble des propositions de l'UT :

- *circonstances* (C) : l'attribut *circonstances* contient les propositions de l'UT relatives à la définition du contexte dans lequel intervient la situation;
- *description* (D) : les propositions de cet attribut expriment les événements qui se déroulent dans le cadre de la situation représentée par l'UT. C'est le seul attribut obligatoire d'une UT;

[1] Si un texte aborde différents grands thèmes sans rapport les uns avec les autres sur le plan thématique, on peut raisonnablement douter de sa cohérence.

- *états incidents* (E) : cet attribut recueille les propositions rendant compte des résultats des événements intervenus lors du déroulement de la situation liée à l'UT.

La figure 5.5 fournit le détail de cette structuration pour les UTs de la représentation de texte de la figure 5.3. Il s'agit plus précisément de la forme linéaire définie pour les représentations de texte à partir de laquelle elles peuvent être reconstruites.

De manière parallèle encore une fois à ce que l'on trouve dans les schémas, on fait figurer dans les UTs les relations temporelles et causales qu'entretiennent entre elles leurs propositions. Ces relations peuvent intervenir aussi bien au sein d'un même attribut (par exemple, des relations temporelles entre les événements de l'attribut *description*) qu'entre propositions appartenant à deux attributs distincts (comme pour les relations causales liant une circonstance à un événement ou un événement à l'un de ses résultats). Elles sont explicitées bien entendu dans la mesure où elles ont pu être découvertes de manière automatique dans les textes. Cette mise à jour est réalisée par des moyens d'analyse spécifiques qui se trouvent en dehors du champ de notre travail. La seule contrainte qui pèse sur eux de notre point de vue est la même que celle pesant sur l'analyse thématique : ils doivent disposer de la capacité de travailler, au moins initialement, en l'absence de connaissances pragmatiques.

Enfin, les UTs textuelles comprennent également des rôles. Chaque rôle représente une entité, ou un ensemble d'entités du monde de référence. Il peut s'agir aussi bien d'une personne, d'un objet que d'un lieu. Il est identifié comme pour les schémas par une variable de coréférence. Ainsi, un rôle d'une UT est constitué de l'ensemble des concepts apparaissant dans les graphes conceptuels de cette UT et possédant la même variable de coréférence. Comme on peut le voir au niveau de la figure 5.5, les rôles n'apparaissent pas autrement dans la forme linéaire des représentations de texte qu'au travers de cette variable. Ils sont en fait définis automatiquement grâce à elle lors de la construction de l'UT dont ils font partie.

Cette notion de rôle existe de façon similaire au niveau des épisodes. Une entité est donc référencée par la même variable à l'échelon d'un épisode entier. Ceci résulte de ce que tous les graphes conceptuels formant ses différentes UTs appartiennent au même contexte, propre à cet épisode. Les rôles d'un épisode sont formés plus précisément de tous les rôles des UTs correspondant à des entités communes à au moins deux UTs. On ne retient en fait comme rôle d'épisode que les entités apparaissant dans plusieurs situations.

Épisode TentativeAssassinatMLK

Commentaire: ce texte rend compte d'une tentative d'assassinat intervenue sur la personne de Martin Luther King. Ce dernier a été gravement blessé au cours de cet acte, perpétré par une déséquilibrée;
Texte: TentativeAssassinatMLK;

UT SéanceDédicace

Commentaire: MLK participe à une séance de dédicace dans un grand magasin;
Type: secondaire;

Attribut Circonstances

Graphe C1
Proposition: P1;
[Être_dans; prédicat: vrai]
{→ (lieu) → [Grand_Magasin: *x1] →
(localisation) → [Quartier: Harlem #2],
→ (source) → [Homme: MLK #1]
}.

Graphe C2
Proposition: P2;
[Entourer; prédicat: vrai]
{→ (agent) →[Humain: {*} *x2],
→ (patient) → [Homme: #1]
}.

Graphe C3
Proposition: P4_1;
[Boycotter: *x3; prédicat: vrai]
{→ (agent) → [Humain: {*} *x4],
→ (objet) → [Bus: {*} *x5] →
(localisation) → [Ville: Montgomery #2],
→ (date) → [Année: {1955,1956} *x6]
}.

Graphe C4
Proposition: P4_2;
[Relater; prédicat: vrai]
{→ (agent) → [Homme: #1],
→ (objet) → [Événement: *x3],
→ (instrument) → [LivreŒuvre: “Stride towards Freedom” #3]
}.

FinAttribut Circonstances

Attribut Description

Graphe D1
Proposition: P3;
[Dédicacer; prédicat: vrai]
{
→ (agent) → [Homme: #1],
→ (objet) → [Exemplaire: {*} *x7] →
(instance) → [LivreŒuvre: #3]
}.

Graphe D2
Proposition: P5;
[Apposer; prédicat: vrai]
{→ (agent) → [Homme: #1],
→ (objet) → [Signature: *x8] →
(possesseur) → [Homme: #1],
→ (destination) → [Page: *x9]
}.

FinAttribut Description

FinUT SéanceDédicace

UT TentativeAssassinat

Commentaire: une déséquilibrée poignarde MLK. Celui-ci est emmené à l'hôpital;
Type: principale;

Attribut Circonstances

Graphe C1
Proposition: P8;
[Être_fou] → (source) → [Femme: *x10].

FinAttribut Circonstances

Attribut Description

Graphe D1
Proposition: P6-7;
[Poignarder; prédicat: vrai]
{→ (agent) → [Femme: *x10],
→ (patient) → [Homme: #1],
→ (instrument) → [Coupe_papier: *x11],
→ (manière) → [Brutal],
→ (objet) → [Poitrine: *x12] →
(partie_de) → [Homme: #1]
}.

Graphe D2
Proposition: P9;
[Transporter; prédicat: vrai]
{→ (agent) → [Humain: *x13],
→ (patient) → [Homme: #1],
→ (destination) → [Hôpital: #4] →
(localisation) → [Quartier: Harlem #2],
→ (manière) → [Urgent]
}.

FinAttribut Description

Attribut ÉtatsIncidents

Graphe E1
Origine: inférenceSémantique;
[Être_dans; prédicat: vrai]
{
→ (lieu) → [Hôpital: #4],
→ (source) → [Homme: #1]
}.

Graphe E2
Origine: inférenceSémantique;
[Être_blessé; prédicat: vrai] → (source) →
[Homme: #1].

FinAttribut ÉtatsIncidents

RelationsIntraUT

Relation D1D2
Origine: analyseTemporelle;
Description.D1 → Description.D2:
précédenceTemporelle.

Relation C1D1
Origine: motivationPersonnages;
Circonstances.C1 → Description.D1:
motivation.

FinRelationsIntraUT

FinUT TentativeAssassinat

UT Hôpital

Commentaire: le séjour de MLK à l'hôpital;
Type: secondaire;

Attribut Description

Graphe D1
Proposition: P10;
[Rester; prédicat: vrai]
{→ (agent) → [Humain: #1],
→ (lieu) → [Lit: *x14],
→ (durée) → [Heures: {*} *x15] →
(caractéristique) → [Long]
}.

Graphe D2
Proposition: P11;
[Faire_préparatif; prédicat: vrai]
{→ (agent) → [Humain: *x16],
→ (manière) → [Intensif]
}.

Graphe D3
Proposition: P12;
[Extraire; prédicat: vrai]
{→ (agent) → [Humain: *x16],
→ (objet) → [Objet_physique: *x11] →
(fonction) → [Arme],
→ (origine) → [Corps: *x17] →
(possesseur) → [Homme: #1]
}.

FinAttribut Description

RelationsIntraUT

Relation D1D2
Origine: analyseTemporelle;
Description.D1 → Description.D2:
simultanéitéTemporelle.

Relation D2D3
Origine: analyseCausale;
Description.D2 → Description.D3: but.

FinRelationsIntraUT

FinUT Hôpital

RelationInterUTs

Relation
SéanceDédicace_TentativeAssassinat
Origine: analyseThématique;
SéanceDédicace → TentativeDeMeurtre:
changementDeThème.

Relation TentativeAssassinat_Hôpital
Origine: analyseThématique;
TentativeDeMeurtre.ÉtatsIncidents.E2 →
Hôpital: Déviation.

FinRelationInterUTs

FinÉpisode TentativeAssassinatMLK

Remarques :

- Dans le graphe C4, on fait référence au graphe C3 en utilisant une variable de coréférence entre le prédicat de C3 et le concept représentant le graphe entier dans C4. Il s'agit là d'une simplification imposée par le fait que nous prenons pas en compte les contextes dans le reste du travail présenté. Le concept [Événement: *x3] devrait en effet être un concept de second ordre prenant comme référent le graphe C3, celui-ci n'apparaissant alors plus au niveau le plus haut.
- Par rapport à la représentation de la figure 5.3, on a fait figurer ici les graphes pouvant être inférés sémantiquement à partir des actions apparaissant de façon explicite (graphes E1 et E2 de l'UT TentativeAssassinat). De ce fait, la déviation thématique entre l'UT TentativeAssassinat et l'UT Hôpital n'a plus pour source le graphe D2 de la première UT mais sa conséquence directe, incarnée par le graphe E1 de la même UT. Les graphes inférés sont différenciés des autres par la valeur du trait *Origine* qui leur est associé. Lorsque ce trait n'est pas explicité, il prend la valeur *Texte*.
- Le trait *Proposition* associé aux graphes est destiné à conserver un lien rudimentaire avec la forme initiale de surface des textes.

Fig. 5.5 - Forme linéaire de la représentation de texte de la figure 5.3

3.3.2. Discussion

La contingence des informations textuelles

Du fait que les UTs sont construites à partir de textes et ne résultent pas d'une modélisation a priori ou d'un apprentissage, elles contiennent généralement un certain nombre de propositions dont la présence n'est que contingente par rapport à la situation dont l'UT rend compte. Ce phénomène se manifeste pour tous les attributs mais il est particulièrement notable en ce qui concerne le contenu de l'attribut *circonstances*. L'UT Dédicace de la figure 5.3, par exemple, contient dans son attribut *circonstances* trois propositions dont l'une d'entre elles peut être considérée comme non liée de façon intrinsèque à la situation d'une séance de dédicace. En effet, si le fait de se trouver dans un grand magasin (proposition 1) et d'être entouré de nombreuses personnes (proposition 2) sont assez caractéristiques de cette situation, le fait que le livre dédicacé relate un boycottage des autobus est quant à lui tout à fait spécifique de la situation particulière abordée par le texte.

L'agrégation des UTs textuelles au sein de la mémoire épisodique, en conduisant à pondérer chacun de leurs graphes en fonction leur fréquence d'apparition, permet justement de faire la part entre les événements propres à la situation et ceux qui ne lui sont pas liés.

Inférences immédiates et représentation des actions

Toujours en observant la représentation de texte construite à partir de la figure 5.2, on notera qu'aucune des UTs définies à partir strictement des propositions du texte ne contient un attribut *états incidents*. Bien ce soit là partiellement le fruit du hasard lié au choix du texte, cette constatation est également supportée par les relations spécifiques existant entre l'attribut *description* et l'attribut *états incidents*. Une action peut en effet être exprimée soit directement par le prédicat qui la caractérise, soit de façon plus indirecte en évoquant une conséquence qui lui est intrinsèquement liée. Par exemple, dire "X a tué Y en le poignardant" est équivalent à dire "X a poignardé Y. Y en est mort". Dans le premier cas, on ne mentionne que des actions. Leur conséquence, en l'occurrence la mort de Y, doit être inférée, alors que dans le second cas, elle est clairement exprimée. En fonction de la façon dont le texte rend compte d'une situation, on peut donc avoir des attributs *états incidents* et *description* plus ou moins développés si l'on s'en tient uniquement à ce qui est explicité.

Les inférences permettant de passer d'une action à sa conséquence immédiate sont de nature sémantique. Elles sont en effet directement liées aux informations présentes au niveau des graphes de définition associés aux actions. Dans le modèle de représentation

sémantique des verbes présenté au chapitre 4, un cas spécifique, le cas But, permet de faire apparaître ces dépendances. De telles inférences peuvent donc être réalisées pour compléter les représentations de texte et les rendre plus homogènes. Une illustration en est donnée au niveau du détail de la représentation de texte de la figure 5.3 (cf. figure 5.5). Le graphe exprimant que Martin Luther King se trouve dans un hôpital est ainsi déduit de celui spécifiant qu'on l'y transporte tandis que celui indiquant qu'il est blessé est inféré du fait qu'il a été poignardé. Nous reviendrons plus globalement sur ce problème de normalisation du contenu des représentations de texte au §4.3.

Forme générale et forme minimale des représentations de texte

La description que nous avons faite précédemment des représentations de texte correspond à leur forme la plus riche et la plus détaillée possible. Dans le contexte relativement pauvre en connaissances dans lequel nous nous situons, on ne peut toutefois espérer disposer d'une analyse des textes toujours très performante. Ce manque se manifeste dans le traitement des éléments explicités par les textes. Des propositions devant être réparties dans des attributs différents au vu de la situation peuvent ainsi se retrouver dans le même attribut. Une UT B qui devrait être considérée comme la déviation d'une UT A peut de même ne pas être reconnue en tant que telle et voir son contenu complètement inclus dans A. Mais ce manque s'exprime également, voire surtout, dans la mise en évidence des éléments implicites des textes. La recherche des relations entre les propositions des UTs ou la réalisation de certaines inférences pour ajouter certaines propositions, notamment si elles vont un peu au delà de celles exposées ci-dessus, sont en effet étroitement conditionnées par les connaissances disponibles.

Les représentations de texte construites n'auront donc pas toujours la forme complète que nous avons décrite. D'ailleurs, l'exemple de la figure 5.5 en est déjà un témoignage puisque l'UT Hôpital ne contient qu'un attribut *description* et l'UT Séance de Dédicace ne possède pas d'attribut *états incidents*. Les représentations de texte ont de ce fait une géométrie variable. Dans des conditions extrêmes de dégradation, un épisode peut très bien n'être formé que d'une seule UT et celle-ci ne contenir que le seul attribut *description*. Bien entendu, si tous les textes sont représentés de cette façon, on ne peut attendre que des résultats limités. Mais la base même de notre approche consiste à faire l'hypothèse que la multiplicité et la diversité des textes à propos d'une même situation permet de compenser les manques propres à l'analyse d'un certain nombre d'entre eux.

Caractérisation des différents attributs

Le but n'est pas ici de donner des règles précises pour répartir les différentes propositions d'une UT parmi ses trois attributs mais de fournir quelques éléments sur la

façon de le faire. Il faut d'abord préciser qu'au dessus de la partition des propositions selon les trois attributs d'une UT, il existe un autre plan de répartition mettant d'un côté les propositions présentes dans les attributs *circonstances* et *états incidents* et de l'autre, celles dépendant de l'attribut *description*. Il s'agit en fait de la séparation entre le premier plan de la situation et son arrière plan. Nous reprenons ici une distinction opérée par l'"approche spatiale" de l'analyse du discours narratif [Irandoust 1997] entre figure et fond [Reinhart 1984]. Le premier plan est ce qui fait progresser l'action tandis que l'arrière plan est tout ce qui contextualise cette même action. En transposant dans notre cadre, l'action est constituée par le contenu de l'attribut *description*, sa contextualisation proprement dite est réalisée au travers de l'attribut *circonstances* tandis que l'attribut *états incidents* exprime l'incidence de la progression de l'action sur le contexte de celle-ci.

Cette distinction entre premier plan et arrière plan présente l'avantage de se manifester de façon assez nette au niveau de la dimension aspectuo-temporelle du discours narratif. En simplifiant, on peut dire que les procès ayant une valeur aspectuelle non bornée sont indicatifs de l'arrière plan tandis que ceux ayant une valeur aspectuelle bornée signalent des éléments du premier plan. Les travaux dans ce domaine ont montré que cette valeur aspectuelle est le résultat de la combinaison de la structure temporelle interne des procès impliqués, appelée également "Aktionsart", et de la perspective temporelle dans laquelle ils sont placés. Ce second aspect rend compte du point de vue subjectif de l'énonciateur et se traduit par l'opposition accompli / inaccompli ou perfectif / imperfectif.

Du côté de l'Aktionsart, il existe de nombreuses classifications plus ou moins raffinées des procès dérivant à des degrés divers des travaux de Kenny [Kenny 1963], puis de Vendler [Vendler 1967]. Le premier a en particulier mis en évidence la désormais classique tri-partition des verbes entre verbes d'état, verbes de processus et verbes d'action. De notre point de vue, nous considérerons que les verbes d'état rencontrés dans les textes se rattachent systématiquement aux circonstances ou aux états incidents d'une UT. En parlant de verbe seulement, on opère une simplification. Des travaux tels que [Dowty 1982] par exemple ont en effet mis en évidence que l'Aktionsart d'un procès n'est pas déterminée seulement par le verbe mais résulte plus généralement de l'interaction aspectuelle de tous les composants d'une phrase. Néanmoins, sur le plan de la réalisation pratique, il semble plus réaliste de s'en tenir uniquement à une classification des verbes comme première approximation.

Pour les actions et les processus, la séparation entre *circonstances* et *états incidents* d'un côté et *description* de l'autre repose sur la détermination de la valeur aspectuelle en contexte du procès considéré : une valeur aspectuelle non bornée conduit à placer une proposition d'une UT parmi les *circonstances* ou les *états incidents* tandis qu'une valeur

aspectuelle bornée l'oriente vers l'attribut *description*. Un certain nombre de travaux, tels que [Maire-Reppert 1990], [Ho 1990], [Vazov 1997] ou encore [Dormont & Gruselle 1993], ont montré que l'utilisation d'un ensemble d'indices linguistiques, conjuguée à une classification des verbes du type de celle de Vendler, permet de réaliser une telle détermination dans un grand nombre de cas.

Il faut néanmoins conserver à l'esprit que l'utilisation de la valeur aspectuelle est un moyen minimal de répartition des propositions parmi les attributs d'une UT. Il n'y a pas en effet nécessairement identité entre la façon dont un fait est relaté et son statut véritable du point de vue causal dans une situation. Or, la valeur aspectuelle d'une proposition est étroitement liée à la façon dont le récit qui la contient est construit.

En ce qui concerne la séparation entre les *circonstances* et les *états incidents*, nous ne proposons pas véritablement de critère systématique présentant une bonne fiabilité. Le critère le plus évident concerne la position relative de la proposition considérée par rapport aux actions de l'UT : les *états incidents* sont naturellement plutôt mentionnés à la suite des actions qui les engendrent qu'avant ceux-ci. De façon similaire, la contextualisation de ces actions, donc les *circonstances*, précèdent généralement leur évocation. Il peut néanmoins subsister une ambiguïté entre les *états incidents* d'une UT et les *circonstances* de l'UT qui la suit. Cette ambiguïté existe même parfois pour un modélisateur humain, en particulier si la seconde UT est une déviation de la première. Au niveau de la figure 5.5 par exemple, le graphe exprimant que Martin Luther King se retrouve à l'hôpital fait partie des *états incidents* de l'UT Tentative de Meurtre mais pourrait tout aussi bien figurer dans les *circonstances* de l'UT Hôpital.

Un second critère évident peut contribuer à lever cette ambiguïté dans certains cas. Il s'agit de la possibilité d'utiliser le graphe de définition des actions afin de déterminer si une proposition est une conséquence intrinsèque d'une action d'un attribut *description*. Si tel est le cas, elle vient naturellement prendre place parmi les *états incidents* de l'UT correspondante. On utilise là le même support que celui sur lequel reposent les inférences immédiates dont nous avons parlé précédemment. Il faut d'ailleurs remarquer que dans un nombre non négligeable de cas, les *états incidents* sont en fait constitués du résultat de ces inférences et que le problème de la discrimination vis-à-vis d'éventuelles *circonstances* ne se pose donc pas.

3.4. Liens avec les travaux sur la structuration du discours

De façon générale, les travaux sur la structuration du discours cherchent à rendre compte du fait que les propositions d'un texte ou d'un dialogue s'articulent les unes avec les autres afin de former un tout cohérent. Outre qu'ils se différencient sur le nombre et le type des relations qu'ils proposent pour rendre compte de cette cohérence, ces travaux se distinguent également par le degré de dépendance qu'ils accordent aux relations de cohérence par rapport aux conditions proprement externes aux textes. Si certains postulent l'existence d'une certaine autonomie de ces relations au niveau linguistique [Kamp 1981, Mann & Thompson 1987], d'autres les lient de façon très étroite aux connaissances sur le monde ou sur la situation considérée [Grosz & Sidner 1986, Schank 1986]. Entre ces deux pôles, on trouve tout un éventail de travaux, de [Hobbs 1979] et [Polanyi 1988] à [Asher 1993] et [Dahlgren 1993], mêlant ces deux visions en leur attribuant respectivement une part plus ou moins importante.

L'ensemble des travaux menés dans la lignée des réflexions de Schank constituent une forme extrême de la position faisant dépendre la cohérence textuelle de connaissances externes aux textes. Les représentations de texte dans ce cadre sont globalement assimilables à une instanciation de structures de connaissances possédées en mémoire, ou tout du moins à un assemblage de tels instanciations. Les relations mises en évidence entre les propositions se situent de fait uniquement sur le plan de la réalité évoquée (causalité, motivation, relations temporelles, ...) et aucunement sur le plan textuel (relations dites rhétoriques du type élaboration, contraste, ...).

La structure interne de nos UTs est assez directement inspirée de cette conception. Elles sont en effet assez proches de structures telles que les XPs [Schank 1986] que nous avons évoqués au chapitre 2 à propos de [Ram 1993]. On retrouve le même découpage en trois attributs selon le triptyque pré-conditions – actions – résultats et la possibilité d'avoir des relations binaires, notamment de nature causale et temporelle, entre les éléments de ces attributs.

Le modèle proposé par Grosz et Sidner prend davantage en compte le texte en tant que tel et sa structure. Il met clairement en évidence la notion de segment de discours et définit la façon dont les segments d'un texte ou d'un dialogue peuvent s'organiser. En revanche, il réfute la possibilité de définir un ensemble clos et autonome de relations entre segments et inféode complètement la structure du discours aux intentions véhiculées par ce discours ainsi qu'à la façon dont elles sont organisées. Reconnaître qu'un but est un sous-but d'un but plus large permet alors de déterminer qu'un segment est inclus dans un autre.

À l'échelon des épisodes, nos représentations de texte sont assez proches de ce que propose ce modèle. Les UTs correspondent à des segments de texte et les relations de suivi thématique, en particulier la relation de déviation, concrétisent les opérations de "push" and "pop" affectant la ou les piles de focalisation de l'attention[1]. La principale différence réside en fait dans ce que caractérise la notion de segment. Dans notre cas, un segment est défini par un critère thématique : il rassemble toutes les propositions du texte relatives à la même situation, alors que dans le modèle de Grosz et Sidner, un segment rassemble toutes les propositions du texte contribuant au même but. Il existe d'ailleurs une certaine ambiguïté sur la nature exacte de la notion de segment dans les travaux existants sur la structuration du discours, en particulier lorsqu'il s'agit de textes. Beaucoup y font référence mais il n'est pas toujours évident de cerner ce qu'elle recouvre exactement. Une des raisons probables de cette ambiguïté vient de ce que la notion d'intention et celle de thème se recoupent assez fortement dans les textes, chose que l'on n'observe pas pour les dialogues[2].

Notre représentation des textes est plus éloignée des autres travaux mentionnés initialement. En particulier, elle ne cherche pas à rendre compte de tous les liens de cohérence existant entre les propositions d'un texte puisqu'elle met l'accent sur leur seule dimension thématique. La RST (Rhetorical Structure Theory) [Mann & Thompson 1987] est une des plus répandues parmi ces théories qui cherchent à rendre compte de la cohérence des textes par un ensemble défini a priori de relations. Elle se heurte cependant à deux difficultés. Bien que l'on puisse subdiviser les relations qu'elle propose entre celles qui se placent au niveau du contenu et celles, véritablement rhétoriques, qui se situent plus au niveau de la forme, en l'occurrence la façon dont le contenu est exposé, [Moore & Pollack 1992] fait remarquer que la théorie ne permet pas de maintenir ce double plan de façon constante : chaque proposition n'est impliquée que dans une seule relation, qui est soit d'un type, soit de l'autre. Il serait ainsi difficile de maintenir un point de vue thématique sur l'ensemble du texte en y ayant recours.

Par ailleurs, la théorie ne définit pas les moyens de reconnaître les relations de cohérence à partir des textes. C'est sans doute ce qui explique que cet ensemble de relations ne soit pas très stable et ait fait l'objet de diverses extensions en fonction de besoins spécifiques.

---

[1] Comme nous l'avons dit précédemment, ces relations ont été reprises de [Grau 1983] où elles sont établies par une analyse thématique utilisant également un mécanisme à base de pile de focalisation de l'attention.

[2] Il faut rappeler que même si le modèle de Grosz et Sidner se veut un modèle général de structuration du discours, il est tout de même influencé par la problématique du dialogue.

C’est notamment dans le but de lever toute ambiguïté sur la nature de ce type de relations et sur la façon de les mettre en évidence que Asher, mais également Lascarides et Oberlander, ont défini la SDRT (Segmented Discourse Representation Theory) en prolongement de la DRT (Théorie des Représentations Discursives) de Kamp. Cette clarification a été réalisée au prix de l’introduction d’une base de connaissances sur le monde. La théorie décrit alors comment l’interprétation d’un énoncé se construit à partir de l’interaction de ces connaissances avec des connaissances proprement linguistiques.

Néanmoins, la SDRT constitue un cadre tout à fait spécifique dont le niveau d’élaboration va au delà de ce que nous recherchons pour représenter le contenu thématique des textes. Compte tenu de notre contexte de travail, nous préférons en effet une analyse relativement grossière, mais robuste, à une analyse et une représentation des textes plus sophistiquées mais à l’application également plus limitée du fait de ses pré-requis. En particulier, la dépendance de la SDRT vis-à-vis d’une base de connaissances sur le monde établie a priori rend son intégration difficile dans notre problématique. Enfin, et ce n’est pas le moindre des obstacles, la dimension thématique n’y est pas présente de façon explicite.

Pour achever cette comparaison rapide avec les types de représentations de texte existants, il faut également mentionner les propositions en provenance de la psychologie cognitive, notamment lorsqu’elles donnent lieu à des modèles informatisables. Les travaux de Schank que nous avons mentionnés précédemment s’inscrivent déjà dans ce cadre. Le modèle de Kintsch et van Dijk [Dijk & Kinstch 1983, Kintsch & Dijk 1978] est également l’une des grandes références dans ce domaine. Il propose de construire une représentation des textes à plusieurs niveaux. Le premier d’entre eux, appelé micro-structure, est assimilable à notre niveau de base puisqu’il résulte d’une analyse propositionnelle des textes. Allant dans le sens de cette proximité, Guha expose dans [Guha 1995] une étude de la pratique de l’analyse propositionnelle en psychologie cognitive, soulignant le flou qui l’entoure souvent, et propose un certain nombre de critères afin de la rendre plus facilement reproductible et tendre ainsi vers une analyse automatique telle que celle produite par un analyseur syntaxique du type LFG.

Le deuxième niveau est celui de la macro-structure. Il est formé d’un ensemble de macro-propositions représentant chacune la condensation d’un ensemble de micro-propositions. Plusieurs niveaux de macro-propositions traduisant différents niveaux de généralité existent pour un même texte. Ces macro-propositions forment un résumé du texte et en définissent la structure. Une macro-proposition s’assimile à ce titre assez directement à une UT, en particulier sur le caractère essentiellement thématique de leur nature. Certaines différences transparaissent néanmoins de cette comparaison. Une macro-proposition ne possède pas ainsi de structure interne puisqu’elle remplace les

micro-propositions plus qu'elle ne les structure. En revanche, elle est dotée d'une appellation explicite alors que les UTs sont anonymes. Enfin, il n'existe pas de relations entre les macro-propositions comme il existe des relations de suivi thématique entre les UTs. Il faut préciser néanmoins que les relations de déviation thématique rendent compte d'une différence de niveau de généralité entre les UTs[1] que l'on retrouve dans le modèle de Kintsch et van Dijk au travers des différents niveaux de macro-structure, sans d'ailleurs que les deux mécanismes soient équivalents.

Le passage de la micro-structure à la macro-structure s'effectue grâce à l'application d'un petit nombre de macro-règles. De notre point de vue, ces règles présentent elles aussi l'inconvénient de s'appuyer assez largement sur une base pré-établie de connaissances sur le monde.

Le dernier niveau du modèle est celui de la superstructure, sans équivalent véritable dans nos représentations de texte. Cette superstructure est destinée à caractériser l'organisation générale de la catégorie de textes auquel appartient le texte traité. Elle rejoint les travaux menés sur la structure du récit par Propp [Propp 1970] et les grammaires de textes de Rumelhart [Rumelhart 1977].

3.5. Implémentation

Comme dans le cas des schémas de la mémoire pragmatique, nous avons implémenté les représentations de texte en Smalltalk en nous fondant sur la plate-forme de graphes conceptuels présentée au §4.1. Une représentation de texte peut être créée soit en faisant appel à l'interface de programmation conçue à cet effet, soit par l'intermédiaire de sa forme linéaire textuelle, pour laquelle nous avons élaboré un compilateur et un générateur. On se reportera à l'annexe C pour la définition précise de cette forme linéaire dont la figure 5.5 donne un exemple assez complet.

L'interface de programmation est destinée à être utilisée par les processus, tels que l'analyse thématique de MLK, qui construisent des représentations de texte de façon automatique. La forme linéaire textuelle permet quant à elle de conserver le résultat de ces processus automatiques sous une forme humainement appréhendable tout en préservant les informations nécessaires à leur reconstruction à l'identique. Ce dernier point offre en particulier la possibilité de mener facilement des expériences sur la construction progressive de la mémoire épisodique. Le premier point permet quant à lui à un expérimentateur de définir des représentations de texte manuellement. Cette capacité est

[1] Il s'agit d'une relation de généralité particulière. Elle est assimilable davantage à la relation partie_de qu'à la relation sorte_de. La relation entre micro-propositions et macro-propositions est à cet égard plus riche.

exploitée afin de tester la mémoire épisodique et d'amorcer le processus d'analyse thématique en l'absence de processus capable de construire des représentations de texte de bout en bout.

Les représentations de texte sont stockées dans une base de textes au sein de laquelle on cherche à maintenir le lien entre les représentations de différentes natures d'un même texte : texte initial, représentation syntaxique, sémantique et thématique. Ce lien entre les différentes dimensions d'un texte n'est pour le moment qu'embryonnaire mais se trouve concrétisé au travers d'un outil spécifique de visualisation et de gestion des bases de textes (cf. Annexe C).

4. Contraintes pesant sur les représentations de texte

En préambule de ce chapitre, nous avons souligné le caractère central des représentations de texte dans MLK. Cette caractéristique tient tout autant à leur position au sein du système qu'à leur importance réelle dans son fonctionnement. C'est tout spécialement le cas en ce qui concerne la mémoire épisodique. Celle-ci est en effet toute entière le produit de l'agrégation de représentations de texte. Cette agrégation est elle-même dépendante de la mesure de similarité entre une représentation de texte et un agrégat de représentations de texte. Conformément à nos principes, cette mesure ne fait pas l'hypothèse de l'existence de connaissances a priori sur le domaine considéré. Elle est donc assez fortement tributaire de la forme même des représentations de texte. Par contre-coup, celle-ci a de ce fait une influence notable sur la forme de la mémoire épisodique. C'est pourquoi nous discutons de façon plus approfondie, dans les paragraphes qui suivent, d'un certain nombre de points touchant à la forme des représentations de texte.

4.1. Les représentations de texte pré-thématiques

Le pré-requis essentiel de MLK est la possibilité de disposer d'une analyse sémantique des textes conjuguée à une résolution, au moins partielle, des co-références. En ce qui concerne l'analyse sémantique proprement dite, nous nous appuyons sur les travaux décrits dans [Briffault et alii 1997]. Ceux-ci exploitent le résultat d'une analyse syntaxique de type LFG (Lexical-Functional Grammar) [Kaplan & Bresnan 1982] pour construire une représentation sémantique des phrases sous forme de graphes conceptuels. Comme dans [Zweigenbaum & Bouaud 1997], ils s'appuient pour cela sur la

spécification au niveau du lexique sémantique de liaisons entre rôle syntaxique et rôle sémantique ainsi que sur une opération de joint dirigé portant dans le premier cas sur les graphes canoniques associés aux types de concept liés aux mots de la phrase et dans le second cas, sur les graphes de définition de ces mêmes types.

Ces deux réalisations, parmi d'autres comme [Bérard-Dugourd et alii 1988, Schröder 1992], permettent au premier abord de penser que les pré-requis posés ne sont pas trop ambitieux. Ce jugement initial doit cependant être modulé comme le soulignent Zweigenbaum et Bouaud dans [Zweigenbaum & Bouaud 1997]. En effet, un mécanisme général tel que celui évoqué ci-dessus ne prend pas en compte les phénomènes de figure de style, tels que la métonymie ou la métaphore, pourtant très fréquents dans les langues naturelles. Même dans un cadre assez contraint comme celui du projet MENELAS [Zweigenbaum & MENELAS 1995], faisant intervenir un type de texte spécifique, en l'occurrence des comptes-rendus médicaux, dans un domaine restreint, celui des maladies coronariennes, il a été nécessaire de développer un module de résolution des métonymies exploitant des connaissances sur le domaine [Bouaud et alii 1996] pour aboutir à une représentation sémantique normalisée des textes.

Cela n'invalide pas nécessairement la possibilité d'une approche plus générale. Le travail décrit dans [Bouaud et alii 1996] est dépendant d'un domaine dans la mesure où le treillis des types de concept et la base canonique qu'il exploite sont eux-mêmes spécifiques du domaine considéré. En revanche, la méthode de traitement des métonymies est tout à fait générale. [Chibout 1993] présente d'ailleurs une méthode tout aussi générale pour résoudre ce même problème et dans une perspective similaire, [Ferrari 1993] fait des propositions pour traiter celui de la métaphore. Dans chacun de ces cas, les limites sont en fait déterminées par les connaissances sémantiques disponibles.

L'analyse sémantique à grande échelle se heurte donc principalement au problème de la modélisation d'un très vaste treillis de types de concept ainsi que de la base canonique qui doit l'accompagner. Un travail tel que [Chibout & Vilnat 1997] est une première avancée dans ce sens mais reste pour le moment insuffisant, notamment pour ce qui est des types de concept nominaux. Il reste de fait à montrer que l'on peut véritablement produire des connaissances sémantiques intéressantes pour la construction d'une représentation normalisée des textes sur un grand ensemble de domaines.

En supposant que l'on adopte une approche minimaliste de l'analyse sémantique en se contentant d'une désambiguïsation des mots et de liens sémantiques directement transposés des rôles fonctionnels d'une F-structure, il est envisageable de produire une représentation sémantique assez large. Dans ce cas, on reporte le problème de l'inhomogénéité des représentations au niveau de la construction de la mémoire

épisodique. De par ses principes de fonctionnement, celle-ci a pour vocation de faire face à une certaine hétérogénéité mais son ampleur reste en pratique à évaluer.

Du côté du traitement des co-références, il est également évident que l'hypothèse faite à la base est assez ambitieuse si l'on suppose que toutes les co-références d'un texte doivent être résolues. Il existe certes un certain nombre de méthodes, telles que celles décrites dans [Lappin & Leass 1994] ou [Popescu-Belis & Robba 1997], permettant de résoudre une part importante des anaphores pronominales sur des critères essentiellement morpho-syntaxiques. La possibilité de faire intervenir des contraintes sémantiques, comme dans [Sabah 1978] par exemple, augmente l'étendue des cas abordables tout en restant compatible avec nos hypothèses de travail[1]. On ne peut néanmoins prétendre traiter toutes les anaphores de cette manière. Les méthodes les plus poussées sont certainement celles s'appuyant sur la structure du discours, dans la lignée de [Sidner 1983] et [Grosz et alii 1983]. Les connaissances pragmatiques qu'elles requièrent limitent cependant leur capacité opérationnelle sur un plan général et rendent plus particulièrement leur application peu souhaitable au sein de MLK, tout au moins dans ses stades premiers de développement.

Il faut donc faire l'hypothèse, de manière similaire à ce qui se passe avec l'analyse sémantique, que la mémoire épisodique peut s'accommoder d'une certaine part d'imprécision dans la détermination des référents. Nous verrons au chapitre 6 que ce n'est pas forcément une hypothèse trop forte dans la mesure où dans les UTs agrégées, on ne cherche pas à conserver les liens de co-référence entre concepts et que, même en cas d'indétermination d'un pronom par exemple, on peut au moins utiliser les contraintes définies au niveau du graphe canonique du prédicat de la proposition dans laquelle ce pronom apparaît.

Dans ce qui précède, nous avons tout particulièrement insisté sur la volonté de ne pas utiliser de connaissances pragmatiques pour produire le type de représentation des textes à partir duquel MLK peut travailler, même si cela doit se traduire par une dégradation de la qualité de ces représentations par rapport aux spécifications fixées. Cette volonté répond au souci de construire un modèle d'apprentissage des connaissances pragmatiques rendant compte des phases initiales de l'acquisition de ces connaissances et pas seulement de leur spécialisation. Il est ainsi plus aisé de garantir son indépendance vis-à-vis du domaine considéré.

Cela n'implique pas cependant qu'une fois certaines connaissances pragmatiques acquises, elles ne puissent pas être utilisées par les processus de construction des représentations de texte. Ce point renvoie plus généralement à l'utilisation d'une théorie

[1] Le problème de la modélisation de connaissances sémantiques à une large échelle se pose alors de la même façon que pour la construction de la représentation sémantique des propositions.

du domaine à la fois incomplète et incertaine. Nous l'aborderons au chapitre 8 à propos de l'analyse thématique mais les processus intervenant dans l'élaboration de la représentation sémantique des textes se trouvent également concernés. Il s'agit d'un problème à part entière que nous n'avons cependant pas développé dans notre travail. Nous nous contenterons d'indiquer que la flexibilité offerte par les architectures multi-agents telles que celle présentée dans [Sabah & Briffault 1993] les rend par nature particulièrement aptes à prendre en compte à la fois l'incertitude et l'incomplétude des représentations manipulées, le cas le plus largement traité étant justement celui du dialogue entre les niveaux dits "supérieurs", comme le niveau pragmatique, et les niveaux dits "inférieurs", comme celui de la sémantique ou de la syntaxe.

4.2. Que doivent contenir les représentations de textes?

La représentation des textes que nous construisons ne cherche pas à rendre compte de toutes les dimensions exprimables par les langues naturelles. Nous privilégions ici la dimension thématique, autrement dit le fait qu'un énoncé fasse référence à un sujet, et plus particulièrement à une situation dans le cas des textes que nous considérons. Notre intérêt se porte donc sur les événements mentionnés lorsque la situation est évoquée et non sur ce qui est formulé à propos de ces événements. Tout ce qui a trait à l'expression d'un point de vue est donc absent des épisodes. C'est le cas par exemple des modalités. Lorsqu'elles sont intégrées dans une UT, les phrases (1), (2) et (3) (dans lesquelles X représente un événement) ont ainsi exactement la même représentation. Celle-ci n'est formée que de la seule la proposition exprimant l'événement X.

Il se peut que X (1)
Jean ne croit pas que X (2)
Pierre veut que X (3)

Ce filtrage s'exerce également à l'encontre de l'expression des points de vue particuliers relatifs aux différents personnages. Dans le texte de la figure 5.2 par exemple, l'auteur indique qu'il sent quelque chose de pointu s'enfoncer dans sa poitrine. L'information à retenir du point de vue thématique réside dans le fait que quelque chose de pointu s'enfonce dans la poitrine de l'auteur (assimilable au fait d'être poignardé via le graphe de définition du prédicat Poignarder) et non le fait qu'il le sente. C'est pourquoi cette dernière information n'est pas présente au sein de l'UT TentativeAssassinat.

Nous n'avons pas dressé l'inventaire exhaustif de tous les verbes (dont les verbes de modalité comme croire, falloir, vouloir, ... et les performatifs comme dire, ordonner, ...)

et de toutes les configurations syntaxiques dont la présence indique la nécessité d'un tel filtrage. À titre indicatif, voici néanmoins quelques règles parmi les plus évidentes :

- suppression de toute négation;
- verbe de modalité ou performatif + subordonnée conjonctive introduite par 'que' -> on ne retient que la subordonnée;
- juxtaposition de deux verbes, le second étant au participe passé -> on ne retient que le second verbe;
- verbe de modalité ou performatif + verbe à l'infinitif -> on ne retient que le second verbe;

Du point de vue de l'analyse thématique, opérer un tel filtrage n'a pas d'incidence puisque tout ce qui n'est pas retenu ne porte pas d'information sur le plan thématique. Par ailleurs, même lorsqu'on exprime la négation de X, on parle effectivement de X et l'on signale par là même que X est significatif vis-à-vis du domaine abordé. Dans un texte cohérent, le rédacteur s'attache de fait rarement à signaler que tout un ensemble d'événements n'ont pas lieu s'ils n'ont pas un lien avec la situation évoquée. Ce filtrage présente en outre l'avantage de simplifier les représentations de texte. Pour faire apparaître un point de vue sur un événement, il est en effet nécessaire de placer le graphe représentant cet événement dans un contexte. Or nous avons indiqué précédemment que nous préférions ne pas introduire cette notion dans nos représentations en l'absence d'une définition bien établie des outils permettant de les manipuler.

En revanche, cette simplification peut poser problème lors de l'abstraction des schémas dans la mesure où la vocation de ceux-ci ne se limite pas nécessairement à la dimension thématique. Un schéma renferme en effet des informations causales et/ou temporelles au travers des relations pouvant exister entre les références faites à d'autres schémas. Le fait de spécifier un événement ou sa négation n'est alors plus indifférent. Le point de vue des personnages impliqués n'est plus à négliger non plus dès lors que l'on souhaite construire une explication faisant intervenir leurs motivations. En l'absence de ces éléments au niveau des représentations de texte, les schémas qui sont abstraits à partir de la mémoire épisodique (cf. chapitre 7) ne pourront donc être exploités que de façon essentiellement thématique dans le cadre du travail considéré ici.

4.3. La normalisation des représentations de texte

Ainsi que nous l'avons signalé au début du §4, il est important que les représentations de texte soient comparables, donc normalisées. Les principaux éléments constitutifs des

UTs étant les graphes représentant les propositions des textes, cet effort de normalisation porte essentiellement sur eux, donc sur la représentation sémantique des textes.

Un des premiers aspects de cette normalisation réside tout simplement dans le découpage en propositions. Les graphes conceptuels composant les attributs contiennent chacun un concept prédicatif[1] désigné par un marquage spécifique comme le prédicat principal de la proposition qu'il représente. Ce prédicat correspond en pratique au verbe de la proposition. Les concepts qui lui sont liés peuvent eux-mêmes être précisés par des caractéristiques. Celles-ci, apparaissant dans la proposition sous la forme d'adjectifs, sont également des concepts prédicatifs. Cependant, nous ne construisons un graphe indépendant ayant la caractéristique concernée comme prédicat principal que si l'accent est mis sur cette caractéristique dans le texte. Cela revient à trouver une proposition portant sur l'explicitation de cette caractéristique, comme c'est le cas pour le graphe C1 de l'UT TentativeAssassinat de la figure 5.5. Ce graphe représente alors un état. Autrement, on se contente de lier le concept et sa caractéristique par une relation de type *caractéristique.* Le concept [Heures: {*} *x15] du graphe D1 de l'UT Hôpital ci-dessus en est un exemple.

Les figures de style telles que la métonymie ou la métaphore constituent un autre problème important du point de vue de la normalisation des représentations de texte. Dans le cas de la métonymie[2], il est ainsi nécessaire d'adopter systématiquement la forme la plus complète. La proposition '*ce livre est incompréhensible*' doit ainsi avoir la même représentation que la proposition '*le texte de ce livre est incompréhensible*', cette représentation faisant apparaître l'entité 'texte' qui n'est pas explicitée dans la première proposition. Pour les métaphores, il convient dans le même esprit de se ramener à l'entité cible afin d'avoir toujours la même référence. La proposition '*le moteur a rendu l'âme*' doit donc avoir une représentation identique à celle de la proposition '*le moteur a cessé définitivement de fonctionner*', la référence à adopter étant celle de l'arrêt du fonctionnement. Nous renvoyons le lecteur au §4.1 pour une discussion sur la possibilité seulement partielle de procéder à une normalisation pour la métonymie et la métaphore. Les autres figures de style telles que la litote, l'hyperbole ou l'ironie sont plus rares dans le type de textes que nous avons considéré et ne seront donc pas abordées.

Le dernier point relatif à la normalisation des représentations des texte concerne le niveau de description à adopter. Grâce à l'opération d'expansion des types de concept, un

---

[1] Nous ferons ici l'hypothèse simplificatrice consistant à définir de façon statique des types de concept prédicatif. En pratique, ce sont les types de concept s'instanciant en langue sous la forme de verbes ou d'adjectifs. On pourra se reporter à [Sabah 1978] pour une analyse faisant apparaître cette notion de prédicat de façon plus dynamique.

[2] Le terme 'métonymie' doit être pris ici dans son sens le plus large. On ne différencie pas ainsi la métonymie stricte de la synecdoque.

concept d'un graphe conceptuel peut être remplacé par le graphe de définition de son type. À l'inverse, si un sous-graphe correspond à la définition d'un type de concept, il peut être remplacé par un concept ayant ce type grâce à l'opération de contraction de type. Suivant la façon dont un texte est formulé, une même notion peut apparaître directement sous la forme d'un seul mot ou bien au travers d'une paraphrase, équivalente à sa définition. Les propositions 6 et 7 du texte de la figure 5.2 offrent l'exemple d'une telle diversité d'expression. La proposition 7 fait apparaître de façon directe l'action de poignarder tandis que la proposition 6 n'y fait référence que par sa définition : enfoncer un objet pointu dans le corps d'un homme. À supposer que l'on compare deux UTs dont l'une contient un graphe comportant explicitement le type de concept *Poignarder* tandis que l'autre abrite un graphe où n'apparaît que sa définition, on devra trouver que ces deux graphes sont équivalents.

La solution en apparence la plus directe pour résoudre ce problème consiste à ramener tous les graphes des UTs à un niveau de représentation commun. En pratique, cela suppose que les définitions des types de concept ne comportent aucune circularité. Chaque type de concept doit être défini en ne faisant appel qu'à des types de concept appartenant à des niveaux strictement supérieurs au sien. Cette définition par niveau suppose l'existence d'un niveau initial, constitué d'un ensemble de primitives sémantiques. Ce niveau initial constitue de fait un cadre de référence privilégié pour comparer les graphes. Pour se ramener à ce cadre, il faut réaliser pour chaque graphe d'une UT des expansions de type successives pour ses différents concepts jusqu'à ne plus avoir que des types de concept appartenant à ce niveau.

L'opération d'expansion d'un type de concept est cependant assez coûteuse en termes de complexité algorithmique. L'utiliser avec un tel systématisme risque donc d'être trop pénalisant en regard de contraintes minimales de performance. Par ailleurs, nous verrons dans le chapitre suivant que l'agrégation des graphes s'accompagne d'une généralisation des concepts qu'ils contiennent tout en conservant les concepts d'origine. Par définition, dans un graphe expansé de façon maximale comme nous le suggérons ci-dessus, les concepts ne sont plus généralisables puisque les types qu'ils possèdent forment le sommet de la hiérarchie des types de concept. En utilisant cette expansion, on gagne la possibilité de détecter des similarités entre graphes plus difficiles à établir mais on perd dans le même temps la dimension prototypique de ce qui est dit.

La solution que nous considérons comme la plus réaliste consiste à ne pas normaliser systématiquement les représentations de texte en se ramenant à un niveau de référence mais à reporter ces problèmes de différences d'expression sur la comparaison des graphes. Il existe en effet des moyens plus rapides que l'expansion des types de concept

pour vérifier que deux graphes d'une représentation de texte ne sont pas similaires, ce qui correspond en pratique à la majorité des cas[1].

Nous adoptons la même attitude face au problème des inférences immédiates que nous avons évoqué au §3.3.2. Lors de la construction d'une représentation de texte, on n'ajoute pas de façon systématique les graphes pouvant être inférés des propositions véritablement exprimées par l'intermédiaire des graphes de définition. En revanche, lors de la comparaison de deux UTs A et B, si l'on ne trouve pas, pour un graphe de l'UT A, un équivalent au niveau de l'UT B, on vérifie si ce graphe ne pourrait pas être inféré d'un des graphes constituant déjà l'UT B.

Récapitulatif

Ce chapitre nous a permis d'exposer la nature précise des représentations de texte qui sont produites par l'analyse des textes et qui servent dans le même temps de support à l'apprentissage des connaissances pragmatiques. Une représentation de texte, appelée également épisode, se présente comme un ensemble d'Unités Thématiques organisé par des relations de suivi thématique. Chaque UT rassemble les propositions d'un texte relative à une même situation. Elle possède une structure en trois attributs proche de celle des schémas de la mémoire pragmatique.

À côté de la structure proprement dite des épisodes, nous avons examiné plus en détail leur contenu et nous avons discuté des moyens de le caractériser, voire de le déterminer. Notre attention s'est portée plus spécifiquement sur le contenu des UTs. Ces dernières abritent la représentation sémantique des propositions des textes, laquelle constitue le pré-requis de MLK. Nous avons ainsi analysé dans quelle mesure l'obtention de ce pré-requis est envisageable. Cette étude laisse apparaître que des réalisations, comme le projet MENELAS par exemple, existent dans des domaines spécifiques. Même si rien n'interdit a priori leur extension, à condition de réaliser un effort très conséquent au niveau de la modélisation des connaissances sémantiques, nous restons conscient que des représentations de texte obtenues sur une large échelle n'obéiront pas nécessairement à toutes les contraintes de normalisation que l'on pourrait espérer.

[1] Tous les concepts d'un graphe ne jouent pas le même rôle au sein d'une représentation de texte. Le prédicat possède en particulier un statut particulier puisque la condition minimale pour que deux graphes puissent être similaires est qu'ils aient le même prédicat. C'est au niveau de celui-ci que les efforts doivent porter principalement. Pour cela, on peut se contenter d'expanser un type de concept uniquement par les types de concept de sa définition, sans se soucier de déterminer, pour une première vérification rapide, s'il existe également un appariement structurel.

Nous faisons néanmoins l'hypothèse que la nature même du modèle de mémoire épisodique que nous proposons le rend capable de résister à la variabilité résultant de cet 'étage' sémantique. Par ailleurs, des travaux comme Wordnet et son prolongement pour les langues européennes, EuroWordnet, ou des travaux tels que ceux de Chibout, que nous avons plus particulièrement détaillés au chapitre 4, montrent qu'un intérêt croissant existe pour une modélisation sémantique à large échelle et que des ressources générales seront de plus en plus largement accessibles dans ce domaine.

À côté du problème des représentations sémantiques, nous nous sommes également intéressé à la façon dont la structuration interne des UTs peut être menée. Nous avons en particulier mis en évidence la possibilité d'exploiter une analyse temporelle, même fondée sur la seule recherche d'indices linguistiques, afin de mener à bien cette tâche dans un nombre significatif de cas.

Nous avons également souligné que le contenu des épisodes se doit d'être purement de nature informationnelle. Tout point de vue exprimé à propos de la dimension locutoire des textes n'apparaît donc pas dans les représentations de texte.

Enfin, nous avons abordé le difficile problème de la normalisation des épisodes. Celle-ci cherche à contrer la variabilité d'expression d'une même notion face à la nécessité de comparer des graphes lors de l'agrégation des UTs. Une action peut en effet apparaître aussi bien sous la forme d'un concept spécifique, d'une proposition rapportant sa définition ou bien encore d'une proposition issue d'une inférence immédiate répertoriée au niveau sémantique. Le coût de l'adoption d'un niveau de représentation privilégié nous a amené à préconiser l'intégration de cette normalisation directement au sein de l'opération de comparaison des graphes.

Chapitre 6

La mémoire épisodique

Ce chapitre présente les différents aspects du modèle de mémoire épisodique que nous proposons afin de stocker les représentations de texte et les transformer de façon progressive en nouvelles connaissances pragmatiques. Il détaille la forme de cette mémoire, qui sous-tend MLK dans son entier, en montrant comment sa structure à base d'agrégats de représentations de texte répond aux principes énoncés au chapitre 1. Nous présentons également les mécanismes par lesquels cette mémoire est utilisable dans MLK. Dans la perspective de l'analyse thématique, nous exposons ainsi un mécanisme de sélection des connaissances abritées par cette mémoire. Celui-ci se fonde sur une propagation d'activation lui permettant de s'adapter naturellement à la fois au contexte de travail courant et à l'évolution de la mémoire. La dernière partie du chapitre est consacrée à la description du processus d'intégration des représentations de texte au sein de la mémoire et donc, à sa liaison avec l'apprentissage.[1]

1. Présentation de la mémoire épisodique

1.1. Caractéristiques

Comme toute mémoire, la mémoire épisodique[2] assure deux fonctions principales : le stockage des éléments qui lui sont fournis et le rappel des éléments qui lui sont demandés. En l'occurrence, les éléments sont des représentations de texte et le principal utilisateur de cette mémoire, tant demandeur que fournisseur, est le processus d'analyse thématique chargé de produire ces représentations de texte. La mémoire épisodique est cependant assez éloignée de l'image traditionnelle de la mémoire centrale d'ordinateur, c'est-à-dire un vaste tableau dont chaque case contiendrait dans le cas présent un épisode. La distinction s'opère aussi bien pour ce qui est du stockage que pour ce qui concerne le rappel.

Alors que dans le modèle de la mémoire d'ordinateur, le stockage d'un élément n'est qu'un rangement passif conservant strictement l'intégrité de cet élément, dans celui de la mémoire épisodique, le stockage est vu de façon active comme l'agrégation de cet élément

---

1 Les spécifications de la structure de la mémoire épisodique ainsi que celles des mécanismes de construction de cette mémoire (cf. §1, 2 et 4) sont le résultat d'un travail mené conjointement avec Brigitte Grau.

2 La terme de "mémoire épisodique" ne fait pas référence à la notion de même nom existant en psychologie cognitive mais constitue plutôt un équivalent de l'expression "mémoire des épisodes".

avec ceux qui, au sein cette mémoire, en sont le plus proches. C'est ce qui donne à ce modèle de mémoire une dimension créatrice et qui le conduit à produire de nouvelles connaissances. La restitution d'un épisode individuel n'est donc plus directe mais demande un effort spécifique de reconstruction. Au delà, le rappel dans son ensemble ne se fait pas par la donnée de l'adresse d'une case mais s'opère de manière associative. Rappeler une UT ayant été mémorisée s'effectue en fournissant les types de concept qui en sont le plus caractéristiques et pas en présentant son numéro. Par ailleurs, ce rappel ne se traduit pas par une réponse unique, qui prendrait la forme d'un épisode ou d'une UT, mais plutôt par un état spécifique de la mémoire, c'est-à-dire une configuration d'activation de ses différents constituants. Dans le cas du rappel, on privilégiera ainsi les éléments les plus activés.

Il s'agit là de la seconde forme d'activité de la mémoire épisodique. Elle lui permet notamment d'être sensible au contexte. Suivant la configuration d'activation caractérisant la mémoire, la réponse qu'elle fournit à une sollicitation n'est en effet pas la même. La fonction de rappel est donc sous la double influence de l'état de la mémoire d'une part, dépendant des sollicitations précédentes dont elle a fait l'objet, et des éléments de la situation courante qui lui soumis d'autre part.

1.2. Principes

Pour concrétiser le modèle de mémoire dont nous avons esquissé les caractéristiques ci-dessus, nous nous sommes appuyé sur un petit nombre de contraintes, très précisément trois. En premier lieu, ce modèle se doit de mettre en œuvre un mécanisme d'émergence par accumulation. Le principe général en est simple. Un ensemble d'entités, ici des épisodes et des Unités Thématiques, ayant été jugées suffisamment similaires sont rassemblées pour constituer des agrégats. Au sein de ceux-ci, ces épisodes et ces UTs sont appariées de manière à ce que leurs parties communes soient superposées. Cette superposition s'accompagne d'une pondération de leurs constituants en fonction de la fréquence avec laquelle ils apparaissent en leur sein. De cette façon, les traits récurrents des situations apparaissent progressivement et peuvent être abstraits ultérieurement dans le but de construire une représentation générale de ces situations. Ce principe d'accumulation donne ainsi lieu à une forme de pré-généralisation.

Cependant, et c'est la deuxième contrainte imposée, l'accès aux épisodes spécifiques doit être préservé. L'agrégation des épisodes ne signifient donc pas leur fusion irréversible. Cette contrainte est soutenue par une justification principale et trouve deux champs d'application importants. Ainsi que nous l'avons souligné précédemment, une phase de généralisation n'est pas supposée prendre place immédiatement à la suite d'une

phase de compréhension. Par conséquent, le seul moyen d'exploiter les connaissances venant des textes est bien souvent de travailler directement à partir du contenu de la mémoire épisodique. Les agrégats qu'elle abrite présentent de ce point de vue l'avantage de fournir une forme de jugement d'importance sur leurs constituants. Si une proposition possède un poids élevé dans un agrégat d'UTs, on peut raisonnablement penser qu'elle est plus caractéristique de la situation représentée par ces UTs qu'une autre, dotée d'un poids faible.

Cette capacité offerte par les agrégats s'exerce néanmoins au détriment de la précision inhérente aux épisodes individuels. Un agrégat représente en effet un amalgame dans lequel un élément A appartenant une entité X ayant contribué à la formation de l'agrégat et un élément B appartenant à une autre entité Y de cet agrégat se retrouvent ensemble alors que A n'apparaît pas dans Y et que réciproquement B est absent de X. Si rien n'est prévu dans ce sens, on perdra l'information que A et B ne sont jamais apparus dans un même épisode. Ceci n'est qu'un exemple du type d'informations que recèlent les épisodes individuels. Parmi les cas les plus intéressants, on trouve également la détection de sous-configurations disjointes : les éléments A, B, et C apparaissent simultanément de façon régulière mais ils ne sont jamais présents en même temps que les éléments C, D et E alors qu'ils font tous partie du même agrégat.

La mémoire épisodique doit donc conserver la possibilité de reconstruire les épisodes individuels, ou tout du moins de reconstituer les informations qu'ils recèlent intrinsèquement. Celles-ci sont exploitables aussi bien du point de vue de la compréhension que du point de vue de l'apprentissage. Dans la construction des représentations de texte, il peut être intéressant de s'appuyer sur les épisodes individuels afin d'alimenter un système de raisonnement à base de cas opérant en particulier sur le problème de la structuration interne des UTs. Sur le plan de l'apprentissage, il est nécessaire que le processus de généralisation ne s'appuie pas uniquement sur les similarités existant entre les épisodes. Pour être plus fin, il lui faut aussi prendre en considération les éléments qui les distinguent, ce qui suppose d'avoir accès au contenu précis des épisodes.

Il faut préciser qu'en dépit de son intérêt, cette caractéristique de la mémoire épisodique n'a presque pas été exploitée dans le travail exposé ici, aussi bien du point de vue de l'analyse des textes que de l'apprentissage de connaissances. Il ne s'agit là cependant que du signe d'un travail à compléter. La validité du principe n'en est pas affectée, ce qui explique pourquoi il a été mis en œuvre.

La troisième et dernière contrainte a trait à la structure générale de la mémoire et à la façon dont s'effectue le rappel des connaissances qu'elle contient. Dans beaucoup de

systèmes à base de cas, tels que AQUA [Ram 1993] par exemple que nous avons examiné au chapitre 2, les cas sont indexés par l'intermédiaire de connaissances abstraites, fournies a priori et organisées le plus souvent hiérarchiquement. Ces connaissances représentent alors les éléments caractéristiques des situations considérées qui sont jugés pertinents vis-à-vis d'une description profonde de ces dernières. Du fait de l'approche constructiviste que nous avons adoptée, ces éléments pertinents ne sont pas toujours connus à un certain stade de développement, ou tout du moins reconnus comme tels. Nous ne pouvons donc pas nous reposer sur eux pour élaborer une structuration a priori de notre mémoire. Cette première contrainte sur la structure de la mémoire en induit naturellement une seconde sur la façon dont on doit l'exploiter. Le processus de rappel des connaissances apprises, qui intervient principalement lors de l'analyse d'un nouveau texte pour la recherche des représentations des situations déjà rencontrées, doit de fait être capable s'adapter de lui-même à l'environnement évolutif que constitue cette mémoire.

2. Structure de la mémoire épisodique

2.1. Structures à l'échelle de la mémoire

Conformément au premier principe énoncé ci-dessus, tous les éléments de la mémoire qui ont été jugés similaires sont regroupés au sein d'une même structure et sont agrégés les uns aux autres. Il en résulte des structures spécifiques, appelés agrégats, au sein desquels la même démarche est appliquée récursivement en suivant la hiérarchie des constituants. Soient A, formé de A1 et A2 et B, formé de B1 et B2; si A et B sont agrégés et si A1 et B1 sont similaires, alors A1 et B1 seront eux-mêmes agrégés. Appliqué aux représentations de texte, ce principe donne lieu à des agrégats d'épisodes et à des agrégats d'UTs au premier niveau de décomposition. Les premiers seront appelés par la suite *épisodes agrégés* et les seconds, *UTs agrégées.* La figure 6.1 offre un aperçu de ce que peut être une mémoire épisodique à l'échelon des épisodes et des UTs tandis que la figure 6.2 donne l'ensemble des représentations de texte qui lui ont donné naissance.

Dans l'exemple de la figure 6.1, on voit que seuls les épisodes possédant le même thème principal ont été agrégés. Les épisodes *e1, e2* et *e3,* qui ont comme thème principal commun une tentative de meurtre au couteau, et les épisodes *e5* et *e6,* pour lesquels il s'agit d'une réparation automobile, sont dans ce cas. Pour leur part, les deux autres épisodes ont conduit chacun à la création d'un épisode agrégé différent. L'agrégation de plusieurs épisodes signifie qu'au moins leurs UTs principales s'agrègent elles aussi. C'est ainsi que se sont formées les UTs agrégées *Tentative_de_meurtre* et *Réparer_voiture* au sein de la mémoire de la figure 6.1.

Un épisode peut avoir plusieurs UTs principales lorsque les éléments qui permettraient de mettre en évidence un thème englobant ne sont pas explicités. Dans un épisode agrégé, il n'est cependant pas nécessaire que toutes les UTs principales de tous les épisodes qui le composent s'agrègent effectivement. Dans l'épisode *e2* par exemple, l'UT principale *Partie_de_cartes* ne s'agrègent avec aucune autre UT, principale ou pas, de l'épisode agrégé avec lequel cet épisode s'associe. On conserve néanmoins l'information qu'il s'agit d'une UT principale.

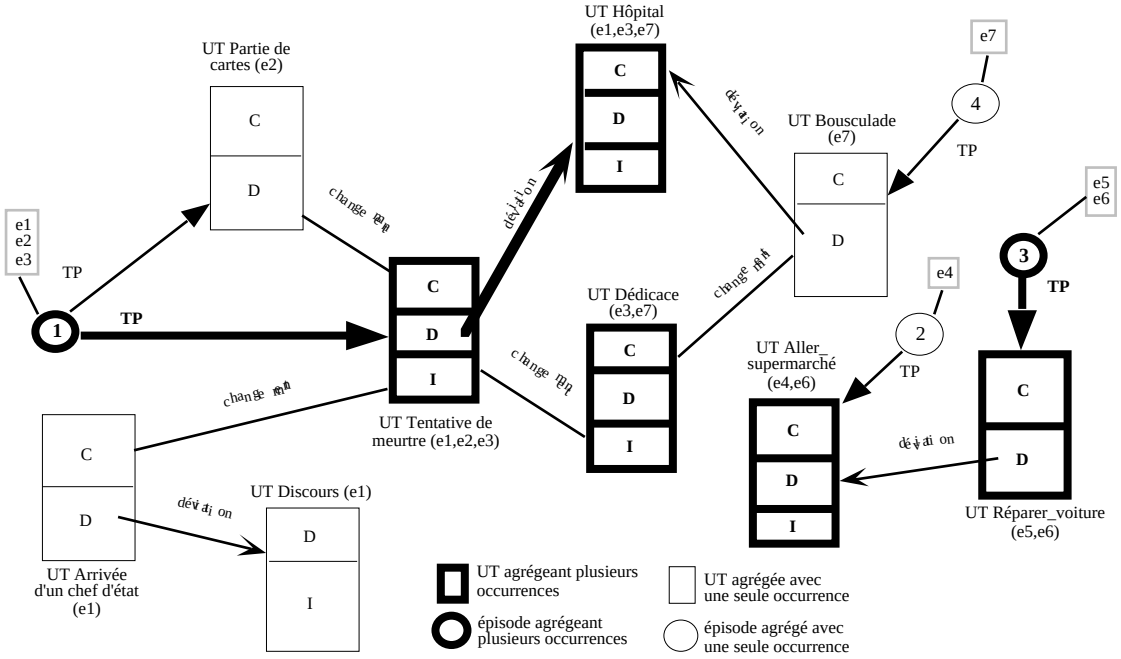


Fig. 6.1 - Un exemple de mémoire épisodique

Épisode e1
(1) Arrivée chef d'état
(2) Discours
(3) Tentative de meurtre (TP)
(4) Hôpital
(1) → (2): déviation
(2) → (3): changement
(3) → (4): déviation

Épisode e2
(1) Partie de cartes (TP)
(2) Tentative de meurtre (TP)
(1) → (2): changement

Épisode e3
(1) Dédicace
(2) Tentative de meurtre (TP)
(3) Hôpital
(1) → (2): changement
(2) → (3): déviation

Épisode e4
(1) Aller supermarché (TP)

Épisode e5
(1) Réparer voiture (TP)

Épisode e6
(1) Réparer voiture (TP)
(2) Aller supermarché
(1) → (2): déviation

Épisode e7
(1) Dédicace
(2) Bousculade (TP)
(3) Hôpital
(1) → (2): changement
(2) → (3): déviation

Fig. 6.2 - Les représentations de texte ayant servi à construire la mémoire épisodique de la figure 6.1

Lors de l'agrégation de plusieurs épisodes, des UTs n'ayant pas le statut d'UT principale – elles sont appelées UTs secondaires – peuvent également être agrégées si elles sont similaires. C'est le cas des UTs des épisodes *e1* et *e3* à l'origine de l'UT agrégée Hôpital.

Enfin, des UTs venant d'épisodes différents et ne s'agrégeant pas eux-mêmes peuvent être rassemblées dans une même UT agrégée, toujours dans le cas où elles ont été jugées similaires. Les UTs agrégées *Tentative_de_meurtre, Hôpital* et *Aller_supermarché* en offrent des exemples. Les deux premières intègrent des UTs venant de l'épisode *e7* d'une part, et des épisodes *e1, e2* et *e3* d'autre part. Or, *e7* n'apparaît pas dans le même épisode agrégé que *e1, e2* et *e3*. La dernière UT agrégée citée est le produit de l'agrégation d'une UT de l'épisode *e4* et d'une UT de l'épisode *e6,* également répartis dans des épisodes agrégés différents de la mémoire.

Cette possibilité d'agréger des UTs appartenant à des épisodes différents illustre une différence entre le statut des UTs au sein de la mémoire et leur statut dans le cadre des représentations de texte. Il faut rappeler en effet que ce modèle de mémoire vise en définitive à faire émerger des représentations de situations prototypiques analogues aux MOPs [Schank 1982]. L'optique consiste donc à préférer construire des unités de petite taille, bien ciblées et donc facilement réutilisables pour décrire d'autres situations, plutôt que des entités plus larges mais également moins souples d'utilisation car beaucoup plus spécifiques. Or, au sein de la mémoire épisodique, les précurseurs de ces représentations sont incarnés par les UTs. Les épisodes ne constituent que des contexte généraux donnant un éventail des différents éclairages sous lesquels ces situations peuvent être considérées. Les UTs bénéficient donc d'une autonomie au sein de la mémoire épisodique qu'elles ne possèdent pas dans les représentations de texte.

Les UTs ne sont pas néanmoins les seules entités à la base des épisodes concernées par l'agrégation. Au même titre qu'il existe des UTs agrégées, il existe également des relations thématiques agrégées comme on peut le voir au niveau de la figure 6.1. Sont concernées aussi bien les relations de déviation que celles indiquant un changement de thème, même si dans ce dernier cas, du fait même de leur définition, il sera certainement plus difficile de trouver des relations agrégées manifestant une forte récurrence. Un tel phénomène signifie en fait que l'apparition des deux UTs impliquées par la relation n'est pas seulement contingente mais régie par un déterminisme sous-jacent de nature non thématique. Les motivations et les plans des différents acteurs des situations concernées peuvent constituer un tel déterminisme.

Pour qu'une relation thématique d'un épisode s'agrège à une relation thématique agrégée, il est bien entendu nécessaire que les UTs qu'elle relie s'agrègent aux UTs

agrégées reliées par la relation thématique agrégée. Ce principe est illustré par la relation de déviation thématique figurant entre les UTs agrégées *Tentative_de_meurtre* et *Hôpital* de la figure 6.1. Dans le cas des relations de type déviation, il est également nécessaire que le graphe source de la relation au sein de l'UT de l'épisode s'agrège avec le graphe agrégé source de la relation thématique agrégée.

Il faut noter enfin que les relations thématiques ne font pas nécessairement de la mémoire épisodique un graphe connexe ainsi que l'illustrent les UTs agrégées *Réparer_Voiture* et *Aller_Supermarché* ci-dessus. La connexité n'est obtenue qu'en faisant intervenir le treillis des types de concept, impliqué dans la mémoire par le biais des concepts constituant les propositions des UTs agrégées.

Les derniers éléments d'un épisode dont nous n'avons pas examiné le sort sont les rôles. Ceux-ci sont par essence étroitement liés à la notion d'instance spécifique. Ils caractérisent en effet l'identité existant entre un acteur d'une situation et un acteur d'une autre situation. La notion de rôle est en revanche plus difficile à cerner lorsque l'on considère plusieurs épisodes similaires. Ceux-ci ne comprennent généralement pas exactement les mêmes situations, même si un certain recouvrement existe nécessairement entre les unes et les autres. Sauf à construire un rôle d'épisode agrégé pour chaque rôle d'épisode, ce qui n'a guère de sens du point de vue de l'agrégation, il faut accepter pour les rôles, comme pour les autres éléments des épisodes, une certaine perte de précision. Un rôle d'épisode agrégé se construit donc par l'agrégation d'un ensemble de rôles d'épisode similaires, et non identiques. Il a ainsi pour fonction de faire émerger les entités qui sont communes aux UTs agrégées les plus significatives, c'est-à-dire les plus récurrentes, d'un épisode agrégé.

En pratique, de même qu'un rôle d'épisode se définit par un ensemble de couples *(rôle d'UT, UT)*, un rôle d'épisode agrégé est formé par un ensemble de couples *(rôle d'UT agrégée, UT agrégée)*.

2.2. Principes de la pondération des constituants de la mémoire

Une des conséquences importantes de l'opération d'agrégation est de pondérer chacun des éléments de la mémoire épisodique. Cette pondération est double. Chaque constituant de la mémoire se voit d'abord associer un *poids absolu*. Celui-ci prend la forme simple du nombre d'occurrences rassemblées par ce constituant (pour alléger l'expression, on parlera par la suite du nombre d'occurrences du constituant). Le poids absolu d'un épisode agrégé est donc le nombre d'épisodes qu'il rassemble et de façon similaire, le poids absolu d'une UT agrégée est le nombre d'UTs qu'elle rassemble. Dans ce dernier

cas, comme dans celui des relations de suivi thématique agrégées, ce décompte s'effectue indépendamment de l'appartenance ou non des différentes occurrences au même épisode agrégé. Toutefois, afin de permettre l'évaluation de la seconde pondération, on conserve au niveau des UTs agrégées l'information du nombre d'occurrences par épisode agrégé.

Les épisodes agrégés 1, 2, 3 et 4 de la figure 6.1 ont ainsi respectivement les poids absolus 3, 1, 2 et 1. L'UT agrégée *Tentative_de_meurtre* possède un poids absolu de 3, résultant de l'agrégation d'UTs dépendant du même épisode agrégé, tandis que l'UT agrégée Hôpital, de même poids absolu, est le produit d'UTs venant de deux épisodes agrégés. Au niveau de la figure 6.1, toutes les relations de suivi thématique ont un poids absolu égal à 1, à l'exception de la relation de déviation entre l'UT agrégée *Tentative_de_meurtre* et l'UT *Hôpital* qui possède un poids absolu de 2.

Cependant, le poids absolu d'un constituant de la mémoire n'a pas intrinsèquement beaucoup de signification. Pour le rendre significatif, il faut le mettre en balance vis-à-vis du poids des autres éléments de la mémoire. Un second poids, appelé *poids relatif*, calculé sur la base du premier, permet ainsi de caractériser l'importance du constituant par rapport aux éléments avec lesquels il est en relation. La façon de le calculer dépend du type des relations entretenues. Dans un couple du type composant/composé, comme celui formé par une UT agrégée et un épisode agrégé, un constituant n'est relié qu'à un seul autre constituant et ceci, de façon orientée. Son poids se définit donc par rapport à ce constituant, en accord avec le sens de la relation. Le poids relatif d'un composant par rapport à un composé est donné en pratique par le rapport entre le nombre d'occurrences du composant et le nombre d'occurrences du composé. Ainsi le poids relatif de l'UT agrégée *Tentative_de_meurtre* par rapport à l'épisode agrégé 1 est-il égal à 1 tandis que le poids relatif de l'UT agrégée *Hôpital* par rapport au même épisode agrégé n'est égal qu'à 2/3. On peut assimiler ce poids à une forme de probabilité conditionnelle. On a ainsi :

Poids du composant par rapport au composé = P(composant/composé)

Ce poids est donc la probabilité d'avoir le composant sachant que l'on a le composé. La probabilité inverse, celle d'avoir le composé, ayant le composant, correspondrait davantage à ce que recouvre la notion de prédictibilité développée dans IPP.

Du fait du sens de la relation composant/composé, il n'y a pas de poids relatif du composé par rapport au composant. Les épisodes agrégés n'ont pas de poids relatif puisqu'il n'existe pas de structure de plus haut niveau au sein de la mémoire épisodique dont ils soient les composants.

Lorsqu'un constituant est relié à plusieurs autres, comme c'est le cas des relations de suivi thématique, le poids relatif est défini par rapport à l'ensemble de ces constituants. Tous les éléments se situent sur le même plan et il n'existe pas de sens privilégié. Le

poids d'une relation est ainsi donné par le rapport du nombre d'occurrences de cette relation sur le nombre de fois où la configuration formée par tous les constituants auxquels il est lié est réalisée. Comme dans le cas précédent, on évalue de cette manière la probabilité d'avoir le constituant en question sachant que l'on dispose des éléments auxquels il est lié lorsqu'il est présent.

Appliqué aux relations de suivi thématique, ce principe édicte que le poids d'une telle relation est donné par son nombre d'occurrences rapporté au nombre de fois où des UTs ayant contribué à la formation respective des deux UTs agrégées qu'elle relie appartiennent au même épisode. Soient UT1 et UT2, deux UTs agrégées reliées par une relation thématique agrégée R. UT1 a été formée par des UTs venant des épisodes a, c, d et g tandis que UT2 a été formée par des UTs venant des épisodes a, b, d et g. Si la relation R est représentée dans les épisodes a et d, son poids relatif a pour valeur 2/3 (2 occurrences / 3 cas de figure où UT1 et UT2, au travers de leurs instances, sont simultanément présentes). On remarquera que le calcul du poids relatif des relations de suivi thématique est, avec le calcul du poids relatif des relations intra-UTs, la seule occasion dans le travail présenté ici où l'on exploite la possibilité de revenir aux épisodes ayant originellement servi à construire la mémoire épisodique.

Dans l'exemple de la figure 6.1, le poids de toutes les relations agrégées est égal à 1, y compris celui de la seule relation thématique agrégée ayant plusieurs occurrences (déviation entre l'UT *Tentative_de_meurtre* et l'UT *Hôpital*). Les instances des UTs agrégées impliquées qui ne sont pas liées par une de ces occurrences appartiennent en effet à des épisodes différents.

2.3. Structures à l'échelle des UTs agrégées

2.3.1. Structure générale d'une UT agrégée

Le principe d'accumulation exposé précédemment s'applique également au sein des UTs agrégées, avec la simple restriction que l'agrégation des constituants des UTs ne s'effectue pas à l'échelle de la mémoire mais intervient uniquement dans le contexte d'UTs ayant été jugées similaires. Deux graphes conceptuels considérés comme similaires appartenant respectivement à deux UTs différentes ne pourront ainsi être agrégés que si les deux UTs ont été elles-mêmes jugées similaires. Cette contrainte est facilement compréhensible si l'on se ramène à ce que représentent ces structures. On ne voit pas en effet pourquoi on chercherait à rassembler des propositions intervenant dans la description de situations différentes si le but est de décrire les situations et non de rendre compte des emplois des propositions.

Circonstances	
(a) *Être_localisé* (0.2)	**(b) *SeQuereller* (0.2)**
(objet) (1.0)	**Événement (1.0)**
(objet) [1]	Événement [1]
(lieu) (1.0)	**Aéroport (1.0)**
(lieu) [1]	Aéroport [1]
(c) *Habiter* (0.4)	
(agent) (1.0)	**(Humain)**
(agent) [3,5]	Homme_politique (0.5) [3], Femme (0.5) [5]
(lieu) (1.0)	**Habitation (1.0)**
(lieu) [3,5]	Appartement [3], Maison [5]
(e) *Menacer* (0.2)	
(agent) (1.0)	**Homme_politique (1.0)**
(agent) [3] [3]	Homme_politique
(patient) (1.0)	**Homme_politique: {*} (1.0)**
(patient) [3]	Homme_politique: {*} [3]
(g) *Commander* (0.2)	
(agent) (1.0)	**Homme (1.0)**
(agent) [4]	Homme [4]
(objet) (1.0)	**Armée (1.0)**
(objet) [4]	Armée [4]

(agent) (1.0)	**Jeune_homme (1.0)**
(agent) [2]	Jeune_homme [2]
(objet) (1.0)	**Argent (1.0)**
(objet) [2]	Argent [2]
(co-agent) (1.0)	**Jeune_homme (1.0)**
(co-agent) [2]	Jeune_homme [2]
(d) *Croire* (0.2)	
(agent) (1.0)	**Homme_politique (1.0)**
(agent) [3]	Homme_politique [3]
(objet) (1.0)	**Idée: {*} (1.0)**
(objet) [3]	Idée: {*} [3]
(f) *Soutenir* (0.2)	
(agent) (1.0)	**Homme_politique: {*} (1.0)**
(agent) [3]	Homme_politique: {*} [3]
(objet) (1.0)	**Idée: {*} (1.0)**
(objet) [3]	Idée: {*} [4]
(h) *Dormir* (0.2)	
(agent) (1.0)	**Femme (1.0)**
(agent) [5]	Femme [5]
(temps) (1.0)	**Nuit (1.0)**
(temps) [5]	Nuit [5]

Fig. 6.3.a - UT agrégée *Tentative de meurtre* (attribut Circonstances)

Notation

La notation adoptée pour représenter les graphes agrégés reprend dans ses très grandes lignes seulement la notation des graphes conceptuels. Les libertés prises sont destinées à augmenter la lisibilité des graphes agrégés. On a plus précisément les conventions suivantes :

- **(g) *Commander* (0.2)** : prédicat du graphe agrégé g, accompagné du poids relatif du graphe.
- **(agent) (1.0)** : relation casuelle agrégée, accompagnée de son poids relatif.
- **Homme (1.0)** : concept agrégé, accompagné de son poids relatif. La notation **Homme: x** permet de spécifier que plusieurs concepts agrégés identiques au sein d'un même graphe agrégé correspondent en réalité à un seul et même concept, dissocié en plusieurs uniquement pour des problèmes de linéarisation de la représentation. Elle doit être différenciée de la notion de variable de co-référence, qui n'est pas utilisée au niveau des graphes agrégés. La même notation existe avec une signification identique pour les types d'occurrences de concept (cf. ci-dessous).
- Appartement [3] : type d'occurrences de concept, accompagné des identifiants des épisodes dans lesquels il est apparu. Un type d'occurrences rassemble l'ensemble des occurrences ayant les mêmes caractéristiques du point de vue de la généralisation. Pour les concepts, il rassemble les concepts ayant le même type et la même classe de référent. On distingue ici deux classes de référents : les référents individuels et les référents ensemblistes. Les concepts ayant un référent ensembliste sont suivi de la notation {*} alors que les autres n'ont pas de suffixe (cela concerne aussi bien les types d'occurrences que les concepts agrégés). Pour les relations, le type d'occurrences rassemble les relations ayant le même type.
- (lieu) [3,5] : type d'occurrences de relation casuelle, accompagné des identifiants des épisodes dans lesquels il est apparu.
- **(Humain)** : concept agrégé ayant comme type celui du concept jouant le même rôle au niveau du graphe canonique associé au prédicat du graphe. C'est le signe qu'aucune généralisation remplissant les

conditions fixées n'a pu être trouvée pour les concepts regroupés. Le poids est alors reporté au niveau des types d'occurrences, ce qui donne une notation telle que : Femme (0.5) [5].

Description	
(a) *Poignarder* (1.0)	
(agent) (1.0)	**Homme (0.8)**
(agent) [1,2,3,4,5]	Soldat [1], Jeune_homme [2], Femme [3], Homme [4,5]
(patient) (1.0)	**Homme: x (1.0)**
(patient) [1,2,3,4,5]	Chef_d'état: x [1], Homme [4], Homme_politique [3], Femme [5], Jeune_homme: x [2]
(objet) (0.4)	**(Partie_du_corps)**
(objet) [1,2]	Bras (0.2) [1], Ventre (0.2) [2]
(partie_de) (0.4)	**homme: x (0.4)**
(partie_de) [1], [1,2]	Chef_d'état: x Jeune_homme: x [2]
(instrument) (1.0)	**Arme_blanche (1.0)**
(instrument) [1,2,3,4,5]	Baïonnette [1], Cran_d'arrêt [2], Couteau [3], Épée [4], Couteau_de_chasse [5]
(manière) (0.2)	**Sauvage (0.2)**
(manière) [5]	Sauvage [5]
(c) *Attaquer* (0.2)	
(agent) (1.0)	**Soldat (1.0)**
(agent) [1]	Soldat [1]
(patient) (1.0)	**Chef_d'état (1.0)**
(patient) [1]	Chef_d'état [1]
(manière) (1.0)	**Soudain (1.0)**
(manière) [1]	Soudain [1]
(g) *SeBaigner* (0.2)	
(agent) (1.0)	**Homme_politique (1.0)**
(agent) [3] [3]	Homme_politique
(lieu) (1.0)	**Baignoire (1.0)**
(lieu) [3]	Baignoire [3]
(b) *Arrêter* (0.6)	
(agent) (1.0)	**Policier (0.66)**
(agent) [1,2,3]	Policier [1,3], Humain [2]
(patient) (1.0)	**Homme (0.66)**
(patient) [1,2,3]	Soldat [1], Jeune_homme [2] Femme [3]
(d) *Trébucher* (0.2)	
(agent) (1.0)	**Soldat (1.0)**
(agent) [1]	Soldat [1]
(e) *Frapper* (0.2)	
(agent) (1.0)	**Jeune_homme (1.0)**
(agent) [2]	Jeune_homme [2]
(patient) (1.0)	**Jeune_homme (1.0)**
(patient) [2]	Jeune_homme [2]
(f) *Pénétrer* (0.4)	
(agent) (1.0)	**Humain (1.0)**
(agent) [3,5]	Femme [3], Homme [5]
(destination) (1.0)	**Habitation (1.0)**
(destination) [3,5]	Appartement [3], Maison [5]
(manière) (1.0)	**(Illégal)**
(manière) [3,5]	Clandestin (0.5) [3], Par_effraction (0.5) [5]
(h) *Perdre* (0.2)	
(agent) (1.0)	**Homme (1.0)**
(agent) [4]	Homme [4]
(objet) (1.0)	**Bataille (1.0)**
(objet) [4]	Bataille [4]
(i) *Attacher* (0.2)	
(agent) (1.0)	**Homme (1.0)**
(agent) [5]	Homme [5]
(patient) (1.0)	**Femme (1.0)**
(patient) [5]	Femme [5]
(j) *Déchirer* (0.2)	
(agent) (1.0)	**Homme (1.0)**
(agent) [5]	Homme [5]
(objet) (1.0)	**Chemise_de_nuit (1.0)**
(objet) [5]	Chemise_de_nuit [5]

Relations intra-UTs agrégées :

D.b → I.a (1.0)
causalité [1, 2]

D.a → I.b (1.0)
causalité [1,5]

D.a → I.c (1.0)
causalité [2,3,4]

Fig. 6.3.b - UT agrégée *Tentative de meurtre* (attribut Description)

États Incidents	
(a) *Être_emprisonné* (0.4)	**(b) *Être_blessé* (0.4)**
(source) (1.0) **Homme (1.0)**	**(source) (1.0)** **Humain(1.0)**
(source) [1,2] Soldat [1], Jeune_homme [2]	(source) [1,5] Chef_d'état [1], Femme [5]
(c) *Être_mort* (0.6)	**(manière) (0.5)** **Léger (0.5)**
(source) (1.0) **Homme (1.0)**	(manière) [1] Léger [1]
(source) [2,3,4] Jeune_homme [2], Homme [4], Homme_politique [3]	**(d) *Être_Guillotiné* (0.2)**
	(source) (1.0) **Femme (1.0)**
	(source) [3] Femme [3]

Fig. 6.3.c - UT agrégée *Tentative de meurtre* (attribut États incidents)[1]

Remarques

Ainsi que nous l'avons rappelé au chapitre 5, l'analyse sémantique s'appuie sur les graphes canoniques associés aux types de concept pour construire la représentation sémantique des propositions. Ces connaissances lui permettent notamment de désambiguïser les mots des textes. Mais elles lui permettent également de rendre explicite des informations absentes des textes. Un ou plusieurs cas d'un prédicat peuvent ainsi ne pas être remplis par un élément venant du texte dans lequel ce prédicat apparaît. Le graphe canonique apporte alors une information par défaut sur la nature des concepts pouvant occuper ce rôle. Comme le montre le graphe (a) de l'attribut *Description*, nous n'exploitons pas directement cette dernière possibilité. Le cas *objet* n'y est en effet rempli que pour deux épisodes sur les cinq dans lequel une instance de ce graphe apparaît. Le fait de laisser de côté cette dimension de l'analyse sémantique se justifie dans la mesure où notre but n'est pas de faire apparaître dans les UTs agrégées les connaissances sémantiques que l'on suppose déjà posséder mais au contraire de faire émerger les éléments spécifiques de chacune des situations que ces UTs représentent. Dans le cas du graphe considéré, il n'est ainsi pas informatif de savoir que l'objet de *Poignarder* est une partie du corps puisque cette information figure déjà dans le graphe canonique de *Poignarder*. En revanche, il serait intéressant de faire apparaître quelle partie du corps est plus particulièrement visée dans une situation de tentative de meurtre.

Par définition, l'opération d'agrégation produit des structures qui sont assez proches de celle des éléments qu'elle prend pour objet. C'est particulièrement le cas en ce qui concerne les UTs agrégées. Leur structure reprend en effet exactement celle des UTs : elles contiennent des rôles, les trois attributs *Circonstances*, *Description* et *ÉtatsIncidents*, lesquels comprennent eux-mêmes des graphes conceptuels représentant des propositions. Ces graphes peuvent être liés, au sein d'un même attribut ou entre attributs, par des relations causales et/ou temporelles. Bien entendu, il s'agit à chaque fois d'éléments agrégés : rôles agrégés, attributs agrégés, graphes agrégés et relations intra-UT agrégées. Parmi ces quatre constituants, les attributs agrégés sont les seuls à ne pas présenter de particularité vis-à-vis de leurs homologues des UTs. L'opération d'agrégation respecte en effet les frontières que fixent les attributs au sein des UTs. On ne cherche pas ainsi à agréger les graphes de l'attribut *Description* d'une UT A avec les graphes de l'attribut

[1] Il faut préciser que l'UT *Tentative de meurtre* représentée ici est différente de l'UT *Tentative de meurtre* présente au niveau de la figure 6.1. Les deux ne sont pas néanmoins sans rapport puisque celle de la figure 6.1 résulte de l'agrégation des UTs des épisodes 1 et 2 de la figure 6.3, auxquelles il faut ajouter l'UT principale de la représentation de texte de la figure 5.5.

Circonstances d'une UT B. Puisque les instances d'un graphe agrégé ne peuvent venir que d'un seul et même attribut, on n'a pas besoin de conserver les différents attributs auxquels ces instances ont appartenu.

Les figures 6.3.a, 6.3.b et 6.3.c montrent le corps d'une UT agrégée résultant de l'agrégation de cinq UTs similaires. Seuls les rôles agrégés n'y sont pas représentés. On trouvera à l'annexe D le détail des UTs qui ont servi à la construire. On notera que la représentation sémantique des propositions ne se conforme pas exactement aux recommandations énoncées au chapitre 5. On trouve par exemple des prédicats ayant directement des fonctions, comme *Soldat, Policier* ou *Chef_d'état*, en tant qu'agent ou patient alors que l'on devrait plutôt avoir un motif du type [prédicat] → (agent, patient) → [Humain] → (fonction) → [Soldat, Policier ou Chef_d'état]. Cette simplification est destinée à alléger la présentation des exemples et n'influe pas sur leur validité dans la mesure où elle est systématique. Pour obtenir une forme "normalisée", il suffit d'expanser systématiquement les formes [prédicat] → (relation) → [fonction] en des formes du type [prédicat] → (relation) → [Humain] → (fonction) → [fonction], l'ajout de la partie [Humain] → (fonction) → n'ayant pas d'influence sur la similarité des graphes.

2.3.2. Les graphes conceptuels agrégés

Parmi les quatre constituants agrégés formant les UTs agrégées, les graphes conceptuels agrégés sont ceux présentant le plus grand intérêt, du fait à la fois de leur richesse et de leur rôle central dans la définition des UTs. Comme le montre la figure 6.4, un graphe conceptuel agrégé est d'abord un graphe conceptuel, c'est-à-dire un graphe bi-partite formé de concepts et de relations. Les graphes agrégés se veulent d'ailleurs une extension des graphes conceptuels de Sowa telle que les opérations associées à ces derniers restent valident lorsqu'elles s'appliquent à des graphes agrégés. Leur particularité essentielle réside dans le fait que chacun de leurs concepts et de leurs relations, appelés respectivement concepts agrégés et relations agrégées, résulte de la généralisation d'un ensemble de concepts et de relations jouant le même rôle dans des graphes ayant été jugés similaires. Le rôle d'un concept ou d'une relation dans un graphe est défini en l'occurrence par l'ensemble des relations qu'il entretient avec les autres éléments du graphe. Deux éléments, concepts ou relations, appartenant à deux graphes différents jouent donc le même rôle s'ils entretiennent les mêmes relations dans leurs graphes respectifs avec des éléments comparables.

Plus formellement, soient E1, un concept ou une relation appartenant à un graphe G1 et E2, un concept ou une relation appartenant à un graphe G2. On dit que E1 et E2 jouent

le même rôle dans leurs graphes respectifs s'il existe une généralisation commune[1] G à G1 et G2 incluant un concept ou une relation E tel que la projection de G dans G1 fasse correspondre E1 à E et la projection de G dans G2 mette en correspondance E avec E2.

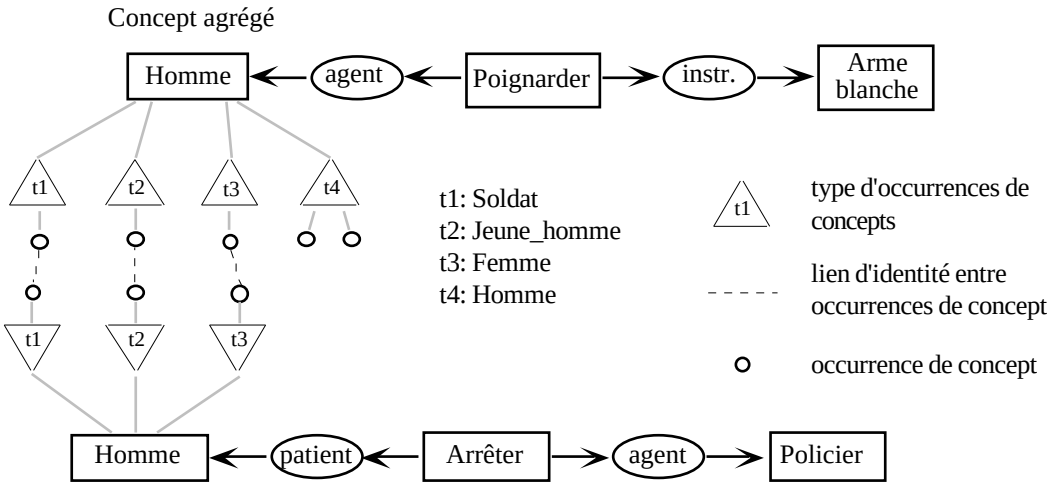


Fig. 6.4 - Liens entre graphes conceptuels agrégés par le biais des instances

Les concepts agrégés et les relations agrégées rendent compte du processus d'accumulation au niveau des graphes conceptuels. De même que l'agrégation des graphes n'intervient que dans le cadre d'UTs similaires, l'agrégation des concepts et des relations n'intervient que dans le cadre de graphes conceptuels similaires.

La généralisation qu'ils caractérisent est très simple. Le type d'un concept agrégé est donné par le sur-type commun minimal de tous les types des concepts qu'il regroupe, à condition toutefois que ce sur-type soit inférieur strictement au type du concept jouant le même rôle au niveau du graphe canonique associé au type du concept considéré comme le prédicat du graphe. Dans le graphe agrégé (a) de la figure 6.3.b, **Poignarder** a **Arme_blanche** pour **instrument** parce que le type *Arme_blanche* est le résultat de la généralisation des types plus spécifiques *Baïonnette*, *Cran_d'arrêt*, *Couteau*, *Épée* et *Couteau_de_chasse*. De plus, *Arme_blanche* est inférieur au type *Objet_pointu* qui se trouve dans le graphe canonique de *Poignarder* pour occuper le rôle d'instrument.

Lorsqu'il n'existe pas de sur-type commun minimal inférieur au type présent dans le graphe canonique, le type retenu est le sur-type regroupant le plus grand nombre d'instances. C'est le cas par exemple du type **Policier** pour l'**agent** du graphe (b) de la figure 6.3.b. Le graphe canonique du type de concept *Arrêter* impose que le type de son agent soit inférieur ou égal au type *Humain.* Or, il n'existe pas de sur-type commun aux types *Policier* et *Homme*, rencontrés au niveau des instances, qui soit inférieur à *Humain.*

[1] cf. chapitre 4 pour la notion de généralisation de deux graphes conceptuels

On retient donc comme généralisation le type *Policier*, qui est apparu, au moins à ce moment-là de l'évolution de la mémoire épisodique, plus souvent que le type *Homme* dans le contexte de cette UT agrégée.

Lorsqu'il n'est pas possible de dégager un sur-type regroupant un plus grand nombre d'instances que les autres, on choisit le type apparaissant au niveau du graphe canonique du prédicat associé. La présence de cette généralisation ne fait alors que garantir la bonne formation des graphes puisqu'elle est strictement non informative. Une telle situation est illustrée par l'**objet** du graphe (a) de la figure 6.3.b. Le type *Bras* se retrouve dans le même nombre d'instances que le type *Ventre* et tous les deux sont immédiatement inférieurs au type *Partie_du_corps*, présent au niveau du graphe canonique de *Poignarder*. C'est donc le type *Partie_du_corps* qui est utilisé comme généralisation.

La limite supérieure fixée pour la généralisation permet de respecter les contraintes sémantiques portées par les graphes canoniques. Elle attribue par là même un rôle particulier aux prédicats des graphes. Ceux-ci ne peuvent pas être généralisés puisqu'il en résulterait dans ce cas un changement du cadre de généralisation des concepts qui leur sont associés. Les connaissances sémantiques définies par l'intermédiaire du formalisme des graphes conceptuels donnent les moyens de réaliser des généralisations au travers du treillis des types de concept mais ne fournissent en revanche que peu de moyens de les contrôler. Les graphes canoniques constituent un instrument pour opérer un tel contrôle mais celui-ci n'intervient que de façon relative : on fixe un plafond pour la généralisation d'un concept relativement à un rôle qu'il joue vis-à-vis d'un autre concept, en l'occurrence celui dont le type porte le graphe canonique. En revanche, il n'existe pas au niveau de chaque type du treillis une information exprimant dans quelle mesure il est possible de le généraliser. Une telle information s'apparenterait à la notion de typicalité mise en évidence par Rosch dans le cadre de la psychologie cognitive [Rosch 1977]. En son absence, on est donc obligé de s'en tenir aux généralisations permises par les graphes canoniques et de ne pas généraliser les prédicats des graphes.

Les concepts agrégés et les relations agrégées ont une double nature et les généralisations que nous avons évoquées ci-dessus n'en constituent qu'une des deux faces. L'autre est incarnée par la conservation de l'ensemble des instances ayant servi à leur construction comme on peut le voir sur la figure 6.4. Ces instances sont le support indispensable afin de pouvoir reconstituer les graphes originellement présents au sein des épisodes et donc, les épisodes tout entier. En dehors de la généralisation des concepts et des relations communs aux différents graphes agrégés, l'opération d'agrégation a en effet pour conséquence de cumuler toutes les différences de ces graphes. Pour savoir quel morceau est associé à tel autre dans un épisode précis, il est donc nécessaire de conserver les informations renfermées par leurs instances au niveau de chaque concept et de chaque

relation agrégés. Cette conservation est également essentielle si l'on veut être capable de reconstituer les rôles des UTs originelles et donc savoir que le concept d'un graphe désigne la même entité que le concept d'un autre graphe de la même UT.

En pratique, on ne garde pas les concepts et les relations originellement présents dans les graphes des UTs car cela conduirait à retenir la plus grande partie des graphes des textes, ce qui n'est pas très économique du point de vue de l'encombrement mémoire. Nous avons préféré conserver pour chaque concept ou relation agrégé l'ensemble des types de leurs différentes instances en associant à chacun d'entre eux la liste des identifiants des épisodes dans lesquels le concept ou la relation est apparu avec ce type. Pour les concepts, on fait une distinction supplémentaire au sein d'un même type entre les concepts ayant un référent ensembliste et un référent individuel. Ces informations permettent à elles seules de reconstituer les graphes des UTs originelles et conjuguées à celles apportées par les rôles d'UTs agrégés, de retrouver les rôles de ces UTs.

Il est ainsi possible de déterminer que dans l'épisode 1 de l'UT Tentative de meurtre de la figure 6.3, un soldat poignarde un chef d'état avec une baïonnette alors que dans l'épisode 2 de cette même UT agrégée, c'est un jeune homme qui en poignarde un autre avec un couteau à cran d'arrêt.

Les graphes conceptuels agrégés ainsi que les concepts et les relations agrégés n'échappent pas à l'autre dimension du processus d'accumulation de la mémoire épisodique que constitue leur pondération en fonction de leur degré de récurrence. Nous ne nous attarderons pas sur leur poids absolu puisque les principes de son calcul sont les mêmes que ceux exposés au §2.2. Le poids relatif d'un graphe conceptuel agrégé est calculé quant à lui en s'appuyant sur sa relation de dépendance du type composant/composé vis-à-vis de l'UT agrégée qui le contient. Il est donc égal au nombre d'occurrences du graphe rapporté au nombre d'occurrences de cette UT agrégée. Il traduit ainsi l'importance du graphe par rapport à l'UT.

Le poids relatif des concepts et des relations agrégés est évalué de façon similaire puisqu'il s'appuie lui aussi sur une relation de dépendance du type composant/composé, cette fois-ci vis-à-vis du graphe conceptuel agrégé qui les contient. Il est donc égal au nombre d'occurrences du concept ou de la relation rapporté au nombre d'occurrences de ce graphe et traduit l'importance du concept ou de la relation par rapport au graphe. Dans le cas des concepts, s'il n'existe pas de sur-type commun minimal aux types des différentes occurrences, on prend comme nombre d'occurrences du concept le nombre d'occurrences du sur-type rassemblant le plus grand nombre d'instances. Lorsqu'un sur-type unique n'existe pas (cas où le type du concept agrégé est le type du concept équivalent dans le graphe canonique associé au type du prédicat), c'est que l'on a

plusieurs types rassemblant le même nombre d'occurrences. Le nombre d'occurrences du concept agrégé est alors égal à cette valeur.

2.3.3. Rôles et relations intra-UTs agrégés

Relations intra-UTs agrégées

Étant donné que les graphes d'une UT peuvent entretenir des relations, les graphes conceptuels d'une UT agrégée en entretiennent eux aussi. Ces relations intra-UTs sont bien entendu des relations agrégées, comme le sont les relations de suivi thématique entre les épisodes agrégés. La figure 6.3 en fait apparaître trois de nature causale prenant place entre des graphes appartenant à deux attributs différents. Bien entendu, elles peuvent exister entre graphes d'un même attribut et caractériser des dépendances temporelles.

Pour qu'une relation venant d'une UT s'agrège à une relation d'une UT agrégée, il faut qu'elle remplisse trois conditions :

- le graphe source de la relation textuelle doit s'agréger avec le graphe source de la relation agrégée;
- le graphe destination de la relation textuelle doit s'agréger le graphe destination de la relation agrégée;
- le sur-type commun au type de la relation textuelle et au type de la relation agrégée doit être inférieur ou égal au premier niveau des types de relation intra-UT.

La dernière condition est imposée par notre volonté de permettre la définition d'une hiérarchie de relations intra-UTs afin de pouvoir s'adapter de la façon la plus étroite possible à la nature des dépendances mises en évidence dans les textes. Au sommet de cette hiérarchie, on trouve les grandes catégories de relation distinguées. En l'occurrence, elles se limitent à deux : les relations causales et les relations temporelles. Conformément à la troisième condition posée ci-dessus, l'agrégation de deux relations appartenant à des catégories différentes est interdite. Chaque catégorie peut être raffinée avec une précision arbitraire. Selon que les moyens disponibles pour l'analyse sont importants ou faibles, on pourra ainsi mettre en évidence des relations plus ou moins fines.

Comme dans le cas des concepts et des relations casuelles agrégés, les relations intra-UTs agrégées ne conservent pas toutes les relations textuelles qu'elles regroupent. Elles se contentent de maintenir, par type de relation, une liste des identifiants des épisodes dans lesquels elles ont eu une occurrence.

Le poids relatif d'une relation intra-UT agrégée se calcule quant à lui suivant le même principe que le poids relatif d'une relation de suivi thématique agrégée. Il rend compte du nombre de fois où cette relation est présente, i.e. son poids absolu, par rapport au nombre de fois où les deux graphes qu'elle relie sont présents de façon simultanée dans un même épisode.

Rôles d'UTs agrégés

Le principe qui préside à la formation des rôles d'UTs agrégés est le même que celui qui prévaut pour les rôles d'épisodes agrégés. Chaque rôle d'UT est décrit par un ensemble de couples (*graphe de l'UT, concept du graphe*). L'agrégation de deux rôles d'UTs r1 et r2 consiste à fusionner les deux ensembles qui les caractérisent en s'appuyant sur l'une des deux règles suivantes pour chacun de leurs couples. Soit c1, un couple de r1 :

- si le graphe de c1 (g1) s'agrège avec le graphe de c2 (g2), un des couples de r2, et que le concept de c1 (cpt1) s'agrège avec le concept de c2 (cpt2), alors c1 et c2 peuvent être fusionnés. Le couple résultant est alors composé du graphe agrégé comportant g1 et g2 et du concept agrégé de ce graphe comprenant cpt1 et cpt2;
- si le graphe de c1 ne s'agrège avec aucun des graphes présents dans un couple de r2, il est directement ajouté au rôle d'UT agrégé rassemblant les deux rôles. La seule transformation consiste à remplacer le graphe et le concept de c1 par le graphe et le concept auxquels ils ont été respectivement agrégés.

On obtient ainsi pour chaque rôle d'UT agrégé un ensemble de couples (*grapheagrégé de l'UT, concept agrégé du graphe*), appelés unités de rôle.

Comme tout élément de la mémoire, un rôle d'UT est caractérisé par un poids absolu et un poids relatif. Son poids absolu rend compte du nombre d'occurrences qu'il rassemble tandis que son poids relatif, exprimé par le rapport de son poids absolu sur le poids absolu de l'UT à laquelle il appartient, définit son importance vis-à-vis de celle-ci. Chaque unité de rôle possède également ces deux poids, le poids relatif étant déterminé dans ce cas en prenant le rôle comme référence. Ces principes de pondération s'appliquent de façon parfaitement similaire aux rôles d'épisode.

2.3.4. Caractéristiques des UTs agrégées

Au delà de leurs structures, les UTs agrégées se singularisent par un profil d'évolution spécifique, dont découle une constitution particulière. L'un comme l'autre résultent des principes adoptés quant au mode de formation de la mémoire épisodique. En dépit de son

manque de maturité – elle ne rassemble que cinq UTs – on peut déjà observer ces tendances au niveau de l'UT agrégée de la figure 6.3.

La plus générale d'entre elles concerne le relief global créé par les poids des graphes agrégés. On est à cet égard face à une situation assez binaire dans laquelle un petit nombre de graphes ont un poids élevé tandis que la majorité d'entre eux sont dotés d'une pondération assez faible. Sur un total de 22 graphes, on n'en compte ainsi que 3 possédant un poids supérieur ou égal à 0,6, c'est-à-dire que l'on retrouve dans une majorité des épisodes.

Une analyse plus fine montre que cette tendance globale se différencie en se plaçant à l'échelle des attributs. L'attribut *Circonstances* rassemble un nombre assez important de graphes agrégés mais chacun d'entre eux n'est le plus souvent le produit que d'un seul graphe conceptuel textuel c'est-à-dire issu d'un texte. Ce trait rejoint assez bien l'intuition selon laquelle une situation donnée peut intervenir dans une assez grande variété de circonstances différentes. Ici, une même situation de tentative de meurtre au couteau prend place dans cinq cadres assez éloignés les uns des autres : la visite d'un chef d'état, une querelle entre deux jeunes hommes pour un problème d'argent, la suppression d'une menace politique, une défaite militaire et enfin, un crime de maniaque. Il est donc plus difficile de trouver des recouvrements au niveau de cet attribut. Cela n'empêche pas de trouver quelques proximités entre les graphes pouvant être les prémisses de caractères plus marqués, moyennant le traitement d'un plus grand ensemble de textes concernant la même situation. Les graphes[1] *SeQuereller* et *Menacer*, bien que sans lien direct, nous placent ainsi dans le contexte d'un événement de nature violente, comme l'est indéniablement une tentative de meurtre. Les graphes *Être_localisé* ou *Habiter* nous renvoient quant à eux à une tendance plus générale des circonstances consistant à contextualiser la situation considérée, notamment sur le plan géographique.

L'attribut *Description* comporte lui aussi un nombre élevé de graphes mais à la différence de l'attribut *Circonstances*, il s'en dégage nettement un noyau significatif, incarné ici par les graphes *Poignarder* et *Arrêter*. Le nombre élevé des graphes est comme dans le cas des circonstances le résultat de la multiplicité des contextes mais on note que ceux rendant compte de la violence de la situation, tels que *Attaquer* ou *Frapper*[2], en plus de *Poignarder* et *Arrêter*, sont globalement plus nombreux et possèdent en moyenne plus de poids que les autres. Bien entendu, on trouve également des graphes tels que *SeBaigner* que l'on peut qualifier de purement anecdotiques par rapport à la situation et

[1] Dans ce qui suit, on fera référence aux graphes par l'intermédiaire de leur prédicat, ce qui est à la fois non ambigu de par le statut du prédicat dans les graphes agrégés, et plus significatif que leur numéro.

[2] Les graphes *Attacher* et *Déchirer* pourraient dans une certaine mesure être rattachés à ce contexte mais moins de façon intrinsèque que du fait des arguments auxquels s'appliquent leurs prédicat.

qui ne seront sans doute jamais renforcés par la suite, ce qui permettra de les éliminer lors de l'abstraction.

L'attribut *ÉtatsIncidents* est à la fois celui comportant le moins de graphes et celui dont les graphes sont les plus significatifs vis-à-vis de la situation représentée. Cette représentativité se manifeste aussi bien sur le plan quantitatif – la moyenne de leurs poids est plus élevée que celle des graphes appartenant aux autres attributs – que sur le plan qualitatif – aucun de ces graphes n'est sans rapport thématique avec la situation considérée. Ces propriétés trouvent leur explication dans la définition même de cet attribut. Les états incidents d'une UT représentent en effet les conséquences du déroulement de la situation caractérisée par cette UT. Ils sont donc liés aux événements formant le moteur de la situation. Les résultats des actions dont le rôle n'est pas central ne sont pas explicités en ce qui les concerne et ne viennent donc pas bruiter l'attribut *ÉtatsIncidents*. Outre leur taux d'agrégation important, la faiblesse du nombre des graphes de cet attribut s'explique par le fait que les conséquences des actions caractérisant une situation ne sont pas toujours explicitées, même lorsque ces actions sont importantes. Le contenu de l'attribut *ÉtatsIncidents* dépend donc en partie de la politique retenue vis-à-vis du développement des inférences sémantiques (cf. §3.3.2 du chapitre 5), qui permettent de mettre en évidence les conséquences les plus immédiates des actions. Cette dépendance ne remet cependant pas en question les traits de cet attribut évoqués précédemment.

Une caractérisation très générale des UTs agrégées conduirait donc à dire qu'elles comportent un grand nombre de circonstances, pour la plupart non spécifiques, un petit nombre d'actions décrivant le corps de la situation, entourées d'un ensemble plus large d'actions plus lointaines bien que significatives et enfin, un petit nombre d'états incidents fortement associés à la situation. Toutes les UTs agrégées ne sont pas supposées suivre ce schéma à la lettre mais les tests effectués au delà d'un exemple comme celui de la figure 6.3 tendent à montrer qu'il est au moins représentatif.

3. Rappel des connaissances au sein de la mémoire épisodique

Nous avons rappelé au début de ce chapitre que les fonctions de stockage et de rappel forment le corps fonctionnel d'un modèle de mémoire. Nous exposons dans ce paragraphe la manière dont la fonction de rappel est mise en œuvre dans notre modèle. La fonction de stockage sera abordée au paragraphe 4.

3.1. Contraintes pesant sur le mécanisme de rappel

Parmi les principes généraux de la mémoire épisodique exposés au paragraphe 1 de ce chapitre, certains posent des contraintes sur les propriétés du mécanisme de rappel des connaissances contenues dans cette mémoire. La première de ces contraintes, au nombre de trois, a trait à la façon dont la mémoire épisodique est utilisée tandis que les deux autres résultent de ses propriétés intrinsèques :

- le rappel des connaissances doit se faire de manière associative. Le traitement d'un texte nécessite de rechercher les épisodes et les UTs agrégées les plus en rapport avec les situations évoquées par ce texte. Cette recherche s'effectue sur la base des concepts exprimés par le texte pour évoquer les situations en question, concepts qui forment pour partie les épisodes et les UTs agrégées à retrouver. Ces concepts servent d'indices pour le rappel de ces structures mais constituent également un premier faisceau de contraintes permettant de réaliser un appariement de surface entre les connaissances trouvées et les éléments en provenance du texte.
  En outre, l'associativité de la mémoire doit s'exercer non seulement par rapport au contexte courant, en l'occurrence les concepts composant la proposition considérée à un moment donné, mais également vis-à-vis du contexte que représentent l'ensemble des propositions du texte déjà traitées, en particulier les plus récentes d'entre elles. Cette propriété doit assurer la levée d'éventuelles ambiguïtés découlant d'une insuffisance de spécificité du contexte courant;

- le processus de rappel doit opérer dans le cadre d'une mémoire ne possédant qu'une structure "plate". Cette propriété de la mémoire épisodique va à l'encontre de beaucoup de modèles de mémoire au sein desquels les connaissances, ou tout du moins les index pointant vers ces connaissances, sont organisés de façon hiérarchique[1], organisation sur laquelle prend alors largement appui le mécanisme de rappel. À défaut d'une structuration hiérarchique des connaissances, les modèles de mémoire se caractérisent souvent par la présence d'index, qui possèdent dans ce cas une structuration de ce type. Le cas de la mémoire épisodique est à cet égard particulier puisque tout concept apparaissant dans les UTs agrégées sert implicitement d'index, avec la caractéristique supplémentaire qu'un poids définit l'importance d'un concept vis-à-vis de chaque UT agrégée dans laquelle il est impliqué;

[1] Que ce soit sous la forme d'une hiérarchie de traits dans laquelle chaque nœud de l'arbre regroupe les entités possédant un même ensemble de traits ou bien d'un réseau de discrimination, structure spécifiquement dédiée à l'indexation dans laquelle chaque nœud représente une condition portant sur les traits des entités. La confrontation des critères de recherche avec cette condition détermine dans quelle sous-arbre est susceptible de se trouver l'entité recherchée.

- le processus de rappel doit opérer dans le cadre d'une mémoire dont la structure évolue au fil du temps. Cette évolution est de plus continue. Là encore, c'est un point assez spécifique de la mémoire épisodique. Les modèles de mémoire possédant une structure évolutive ne sont en effet pas les plus répandus. Le cas général s'incarne plutôt dans ce que l'on trouve majoritairement dans le domaine du CBR : l'évolution de la mémoire prend simplement la forme de l'ajout de nouveaux cas venant s'insérer dans un cadre de connaissances fixe[1]. Dans les rares modèles où ce cadre peut évoluer, cette évolution s'effectue plutôt de façon discrète, par des généralisations ponctuelles intervenant lorsque suffisamment de cas similaires ont été rencontrés.

3.2. Quelques éléments de solution

Le domaine dans lequel les modèles de mémoire et le problème de l'accès au contenu de ces mémoires en fonction d'un contexte donné ont été le plus étudiés est très certainement celui du Case-Based Reasoning. Dans son ouvrage de référence sur ce sujet [Kolodner 1993], Kolodner utilise deux critères pour caractériser les différents modèles de mémoire existant : mémoire à plat ("flat memory") ou bien hiérarchique; recherche séquentielle ou bien utilisant le parallélisme. Le premier critère nous intéresse tout particulièrement puisque notre modèle de mémoire possède une structure à plat. Kolodner souligne que ce type d'organisation, même s'il présente l'avantage d'une mise à jour facile[2], est sans doute le moins optimal de tous parce qu'étant le plus coûteux du point de vue de l'accès : il oblige en effet à parcourir l'intégralité de la base de cas pour trouver celui ou ceux qui sont intéressants pour le problème considéré.

Ainsi que nous l'avons évoqué au chapitre 1 (§2.2.1), Kolodner propose trois pistes pour améliorer les performances d'une mémoire à plat : l'utilisation du parallélisme, l'indexation de surface et la partition de la mémoire. Compte tenu de la contrainte d'associativité de la mémoire, il nous semble que l'indexation de surface est le point le plus important. Une telle indexation consiste simplement à indexer chaque cas par tout ou partie de ses constituants élémentaires. On obtient un index à plat qui permet tout de même, disposant d'un ensemble de ces constituants, de retrouver rapidement l'ensemble des cas les possédant. Dans le cas de la mémoire épisodique, les constituants sont les

---

[1] Même le problème de l'ajout des cas au sein d'une mémoire de cas est un problème en définitive peu abordé dans la littérature sur le CBR. À titre d'illustration de ce fait, on peut remarquer qu'un ouvrage de référence dans le domaine comme [Kolodner 1993] ne consacre à ce problème que quelques pages sur un total de plus de 650 pages.

[2] Ce n'est d'ailleurs pas notre cas ici puisque la mémorisation d'une représentation de texte est beaucoup plus complexe qu'un simple rangement dans un tableau.

types de concept formant les graphes conceptuels agrégés et les structures indexées sont les UTs agrégées (on peut retrouver les épisodes agrégés à partir des UTs agrégées).

La partition de la mémoire nous semble dans le cas présent difficilement applicable de façon statique compte tenu du caractère évolutif de la structure de la mémoire épisodique. La seule structure au-dessus des UTs agrégés est incarnée par les épisodes agrégés mais leur degré de granularité est trop faible pour être le support d'une partition intéressante de la mémoire. Tout autre partition demanderait un travail important d'évaluation de la similarité entre les UTs agrégées de façon à regrouper celles qui sont les plus proches.

Par ailleurs, cette partition devrait répondre aux deux impératifs suivants. Tout d'abord, un procédé de choix du (ou des) regroupement(s) d'UTs agrégées le(s) plus en adéquation avec une configuration de concepts venant d'un texte devrait être conçu. Ensuite, il est indispensable que le mécanisme régissant la partition de la mémoire soit étroitement intégré au mécanisme d'agrégation des représentations de texte présidant à la construction de la mémoire. L'évolution constante de la structure de la mémoire impose une telle intrication. Il est assez évident que la prise en compte de cette dimension est nécessaire pour disposer d'un modèle de mémoire viable sur le long terme – nous reviendrons sur ce point à la fin de ce chapitre – mais elle impliquerait une modification suffisamment substantielle des procédures de construction de la mémoire pour que nous ne l'ayons pas envisagé pour le moment[1].

Dans le cadre actuel, il nous semble préférable d'opter pour une partition dynamique de la mémoire. L'objectif est de sélectionner les UTs agrégées les plus directement en liaison avec les configurations de concepts venant des textes, lorsque celles-ci sont présentées à la mémoire, sans avoir néanmoins à parcourir l'intégralité de la mémoire. Une des façons les plus évidentes de réaliser cette tâche est de s'appuyer sur une diffusion d'activation. Cette solution est favorisée dans le cas de la mémoire épisodique par le fait que les liens entre les concepts et les UTs agrégées sont pondérés. On peut dès lors moduler assez naturellement la diffusion de l'activation en fonction de l'importance des concepts traduite par ces poids vis-à-vis des UTs agrégées à sélectionner.

Mais l'utilisation de la propagation d'activation peut s'étendre au delà de ladélimitation initiale d'un espace de recherche. Ce point est en liaison avec le recours au parallélisme, qui est la troisième piste évoquée par Kolodner afin d'améliorer la recherche dans une mémoire à structure plate. Une part des travaux impliqués dans cette voie se sont intéressés, suite notamment à la mise à disposition de la Connection Machine [Hillis

[1] Il faut préciser qu'en poussant cette logique jusqu'au bout, on transforme la partition de la mémoire en une véritable organisation hiérarchique, ce qui fait disparaître de lui-même le problème de la structuration "à plat" de la mémoire.

1985], à l'exploitation de l'architecture SIMD. Il nous semble cependant que le parallélisme ne résout intrinsèquement pas les problèmes de performance. Les améliorations dans ce domaine résultant de l'utilisation de machines parallèles sont toujours à mettre en relation avec la taille de la mémoire considérée. Elles ne sont en effet effectives que si cette taille ne dépasse pas les capacités de parallélisation de la machine. Or, dans une tâche telle que l'apprentissage de connaissances pragmatiques, il semble impossible de fixer une limite supérieure à la taille de la mémoire lorsqu'on aborde le problème dans toute sa généralité.

D'autres travaux, comme [Domeshek 1991] par exemple, se sont situés à un niveau un peu plus générique en explorant l'apport possible des modèles connexionnistes. Cette voie nous a semblé plus intéressante dans le cas de la mémoire épisodique, notamment du fait de la possibilité d'exploiter les poids intrinsèquement présents entre ses constituants. Il ne s'agit pas de faire appel en l'occurrence aux techniques connexionnistes dans toute leur étendue et leur originalité : nous ne cherchons pas à faire construire par un réseau de neurones une représentation distribuée de nos connaissances. Nous nous situons plutôt dans le cadre des réseaux à représentations locales, dans lesquels chaque neurone figure une entité directement interprétable. On les nomme également "réseaux à propagation d'activité". Ils ont été popularisés notamment au travers du modèle de Waltz et Pollack [Waltz & Pollack 1985] et des travaux de McClelland et Rumelhart [Rumelhart & McClelland 1986].

Il est maintenant clairement établi qu'un des principaux problèmes posé par ce type de réseaux concerne la possibilité de les doter de procédures d'apprentissage véritablement sensibles à leur contenu [Jodouin 1993]. Un réseau de ce type n'est en effet pas maître de la représentation des données qu'il contient – celles-ci sont encodées dans sa propre structure – et il lui est donc difficile de la faire évoluer tout en conservant leur cohérence. Cet aspect ne constitue cependant pas une gêne dans notre cas puisqu'avec le mécanisme d'agrégation, la mémoire épisodique dispose d'un moyen externe à la dimension connexionniste de faire évoluer les poids sur lesquels se fonde la propagation d'activité. Celle-ci n'est utilisée ici que pour sélectionner des connaissances contenues dans la mémoire en accord avec un contexte donné.

Au delà de la problématique du raisonnement à base de cas, les systèmes utilisant la propagation d'activité sont nombreux, même en se limitant à ceux se présentant comme des modèles de mémoire. La propagation d'activité a connu en effet un certain succès dans le cadre de la modélisation psychologique [Kekenbosh & Denhière 1988, McClelland 1988] et en conséquence, les travaux soucieux d'une vraisemblance vis-à-vis de la cognition humaine en ont fait usage dans des contextes divers. C'est ainsi que dans

un système tel qu'ACT* [Anderson 1983], cette propagation s'effectue au sein d'un réseau sémantique modélisant la mémoire sémantique, un peu à la manière de ce que font Waltz et Pollack, alors que dans un système comme INFLUENCE [Cornuejols 1989], elle est réalisée au sein de la mémoire épisodique et concerne les structures complexes que ce sont des instanciations de schémas. Nous verrons dans ce qui suit que notre besoin en termes de propagation d'activité couvre conjointement les deux situations.

Nous avons choisi de présenter un peu plus en détail deux systèmes plus récents que ceux évoqués précédemment : REMIND, dont nous nous sommes inspiré notamment sur le plan technique pour concevoir notre propre mécanisme de sélection des connaissances au sein de la mémoire épisodique et MOORE, qui possède quelques principes similaires à ceux que nous avons essayé de mettre en œuvre.

3.2.1. REMIND

REMIND (Retrieval from Episodic Memory through INferencing and Disambiguation) [Lange & Wharton 1993] est un modèle de mémoire reposant sur l'idée que le rappel de cas et la réalisation d'une tâche faisant intervenir ces cas sont étroitement liés et doivent s'influencer mutuellement. La tâche est en l'occurrence une tâche de compréhension du langage naturel aux niveaux sémantique et pragmatique. Les cas, appelés ici épisodes, sont quant à eux les représentations du résultat de la compréhension de textes déjà analysés. REMIND abrite à la fois des connaissances générales, exprimées sous la forme de schémas comparables à ceux décrits au chapitre 4[1], et des cas, se présentant comme des instanciations de ces schémas. Ces deux formes de connaissances cohabitent au sein du même réseau et sont donc intimement associées. L'interdépendance entre le rappel des épisodes et la compréhension est assurée par le fait que ces deux processus utilisent la même mémoire et le font en ayant recours au même mécanisme de propagation d'activité.

Celui-ci, de même que la façon dont les connaissances sont représentées, sont repris du système ROBIN (ROle Binding and Inferencing Network) [Lange & Dyer 1989], lequel cherchait à montrer comment des inférences de haut niveau peuvent être réalisées dans un réseau connexionniste à représentations locales. La réalisation de telles inférences dans un réseau connexionniste se heurte en effet au problème de la représentation des variables ("variable binding problem"). Pour ce faire, le réseau proposé par ROBIN est composé de deux types d'unités.

Les schémas et les concepts sont représentés par un premier type d'unités, comparables à celles composant une grande partie des réseaux connexionnistes : chacune

[1] La principale différence avec notre approche concernant la représentation des connaissances est que dans REMIND, les connaissances sémantiques sont représentées avec le même formalisme que les schémas.

d'entre elles possède un niveau d'activité, appelée "evidential activation", qu'elle réévalue périodiquement en combinant les activités qui lui sont envoyées par les unités qui lui sont liées. Chaque schéma et chaque concept se traduit par une unité unique dans le réseau et les références faites entre schémas, entre concepts ou de schéma à concept se matérialisent par autant de liaisons entre les unités du réseau les représentant. L'activité transmise d'une unité à une autre est modulée par le poids associé à la liaison entre les deux unités. Ce poids est informellement comparable à une probabilité conditionnelle. Si un schéma S_i fait référence à un schéma S_t, on peut voir le poids entre ces deux entités comme la probabilité d'activer S_i compte tenu de l'activation de S_t.

Les unités relevant du second type distingué ci-dessus, appelées aussi unités de liaison ("binding nodes"), sont chargées quant à elles de mettre en œuvre la notion de variable. Chaque schéma possède un ensemble de facettes constituant autant de places remplies par des références faites à des concepts ou à d'autres schémas. Chacune de ces places est figurée dans le réseau par une unité de liaison. Ces unités ne véhiculent pas une activité comme dans le cas précédent mais des signatures. Une signature est l'identifiant d'une entité de la mémoire, schéma ou concept. Les liens entre les unités permettent la circulation des signatures et rendent compte des contraintes définies sur les entités susceptibles de remplir une place.

De cette façon, on peut s'assurer que des contraintes d'identité entre des personnages intervenant dans des schémas différents, mais liés, sont respectées. Ces unités ont à cet égard un rôle un peu similaire à celui que jouent dans les schémas de la mémoire pragmatique les copies des graphes d'en-tête associés aux références vers des schémas ainsi que les variables de coréférence (cf. chapitre 4, §2.2.1). Par un certain nombre de mécanismes que nous ne détaillerons pas ici car nous n'envisageons pas d'utiliser la propagation d'activité pour réaliser les tâches d'analyse textuelle, les unités de liaison influencent la propagation d'activité dans les unités du premier type et lui permettent ainsi de tenir compte[1] des contraintes d'instanciation pour l'activation des schémas et des concepts.

Sachant que le mécanisme de propagation d'activité réalise de façon intégrée l'ensemble des tâches fixées, le fonctionnement global de REMIND, et par voie de conséquence celui de ROBIN, est simple. Il suffit en effet de présenter en entrée les énoncés à interpréter et d'observer, lorsque le réseau est stabilisé, quelles sont les connaissances les plus activées. Celles-ci sont alors retenues comme interprétation des énoncés considérés. Dans le cas de REMIND, s'ajoutent aux connaissances générales les épisodes les plus liés aux énoncés, d'où la conjonction de la compréhension et du rappel

[1] Il s'agit essentiellement de bloquer certains flux un peu à la manière des transistors.

des cas. Plus précisément, les deux systèmes prennent en entrée le résultat d'une analyse syntaxique des énoncés. L'activation du réseau est initiée en fixant à un niveau important l'activité des concepts composant ces énoncés. Lorsqu'un mot est ambigu, on active les différents concepts auxquels il est susceptible de renvoyer.

Sur un plan plus proprement technique, REMIND et ROBIN se distinguent par la façon dont ils assurent le contrôle global de l'activité du réseau. La fonction d'activation associée à chaque unité U_j, en dehors des unités de liaison, se définit relativement classiquement par :

$$a_j(t+1) = \sum_i w_i \;\; o_i(t) + a_j(t) \;\; (1-\Theta)$$

avec

$a_j(t)$: activité de l'unité U_j au temps t

Θ : facteur d'atténuation de l'activité du cycle précédent

i : indice parcourant l'ensemble des unités auxquelles U_j est liée

w_i : poids de la liaison entre U_j et l'unité désignée par i

$$o_i(t) = \begin{cases} a_i(t) & \text{si } a_i(t) \geq \theta \\ 0 & \text{sinon} \end{cases}$$

θ étant la valeur d'activité minimale d'une unité pour que son activité puisse être propagée vers ses unités voisines

Cette fonction réalise la somme des activités provenant des unités voisines et l'ajoute à l'activité du cycle précédent de l'unité considérée. La valeur de cette activité est plus ou moins atténuée selon la valeur de Θ, ce qui permet d'introduire une dépendance plus ou moins forte vis-à-vis du contexte antérieur. Par ailleurs, l'existence d'un seuil d'activité (cœfficient θ) à atteindre nécessairement pour autoriser la diffusion de l'activation vers les unités voisines permet de mettre plus ou moins l'accent sur des interprétations secondaires possibles.

Les valeurs positives adoptées pour les paramètres de la fonction d'activation des unités font que celle-ci cumule l'activité de cycle en cycle sans que ce cumul ait la moindre de chance de converger vers une valeur limite puisque les poids modulant les entrées d'une unité sont toujours positifs. Ce problème ne survient pas dans un réseau tel que celui de Waltz et Pollack car celui-ci comporte des liens inhibiteurs (poids négatif). Néanmoins, à côté de la difficulté à justifier les liens de ce type[1], leur présence a tendance à faire adopter au réseau un comportement de type *winner-take-all*, ce qui ne permet pas les réinterprétations ultérieures ou l'exploration d'interprétations secondaires une fois la

[1] Comme nous le verrons au chapitre 9 avec le calcul de l'information mutuelle entre mots, il est difficile de mettre en évidence des corrélations négatives.

stabilité atteinte. Pour éviter que la valeur d'activité des unités ne croisse indéfiniment sans pour autant introduire des inhibitions entre unités, le réseau est doté d'un mécanisme global de régulation. À la fin de chaque cycle, la valeur d'activité de chaque unité est divisée par la valeur moyenne des valeurs d'activité de toutes les unités du réseau. On conserve ainsi les différences relatives entre les activités des différentes unités tout en les maintenant dans un intervalle restreint, ce qui permet entre autres d'interpréter leur évolution beaucoup plus facilement.

3.2.2. MOORE

Le modèle de mémoire MOORE (Memory Organization and Optimized Retrieval Engine) [Francis 1994] s'inscrit dans un contexte différent de celui de REMIND. En premier lieu, il n'est pas spécifiquement lié au raisonnement à base de cas. Ensuite, et surtout, il ne fait pas l'hypothèse que le rappel des connaissances en mémoire et la tâche utilisant ces connaissances doivent être intriqués l'un dans l'autre. MOORE se présente au contraire comme un modèle de mémoire doté d'une certaine autonomie et ayant vocation à être intégré dans différents systèmes. Cette autonomie se manifeste en particulier par le fait que MOORE est une mémoire asynchrone. Autrement dit, elle reçoit des requêtes sous la forme de messages qu'elle exécute à son rythme, indépendamment du ou des processus qui ont envoyé ces requêtes.

Cette caractéristique centrale de MOORE est le produit du contexte dans lequel cette mémoire a été développée. L'objectif de Francis est en effet d'étudier le raisonnement opportuniste et plus particulièrement de fonder celui-ci sur un modèle de mémoire répondant aux exigences de ce type de raisonnement. Cette étude s'est matérialisée par un environnement générique de développement de ce type de systèmes appelé NICOLE [Francis 1995], environnement dont MOORE est l'un des principaux composants. Le raisonnement opportuniste s'intéresse de façon générale à la situation dans laquelle un agent engagé dans l'accomplissement d'un plan est confronté à la survenue d'événements imprévus mais entrant dans son champ d'intérêt. Ces événements peuvent être aussi bien de nature exogène, c'est-à-dire en provenance du monde dans lequel l'agent est plongé, qu'endogène, autrement dit en provenance de l'agent lui-même. La problématique de ce type de raisonnement n'est en pratique pas très éloignée de celle des tableaux noirs. Dans un tel contexte, les accès à la mémoire peuvent intervenir à tout moment. En particulier, une demande d'accès peut tout à fait survenir alors que le traitement d'une demande précédente n'est pas achevé, ce qui explique la nécessité de disposer d'une mémoire fonctionnant de manière asynchrone.

Sur le plan structurel, MOORE est plus spécifiquement formé de deux mémoires : une mémoire de travail et une mémoire à long terme. Cette distinction reprend celle opérée traditionnellement en psychologie cognitive. Sans prétendre être un modèle psychologique, MOORE a été en effet conçu avec le souci d'une certaine plausibilité psychologique.

La mémoire de travail est mise en œuvre par un système de type tableau noir. Celui-ci intègre les processus chargés de gérer les relations avec la mémoire à long terme ainsi qu'un ensemble de buffers spécialisés sur lesquels s'appuient ces processus. Ces buffers permettent à la fois d'assumer la tâche de mémoire à court terme en conservant un lien vers les parties de la mémoire à long terme les plus activées et de gérer les entrées/sorties avec les composants externes à la mémoire en servant de file d'attente pour les demandes d'accès et les réponses à ces demandes, permettant ainsi de mettre en œuvre la dimension asynchrone de MOORE. Dans le but de gérer une éventuelle pluralité des modalités (non testée pour le moment), ces buffers sont supposés spécifiques d'une modalité.

La mémoire à long terme est formée quant à elle d'un réseau sémantique comparable au réseau KODIAK décrit dans [Wilensky et alii 1988]. Il est classiquement formé de concepts reliés entre eux par des relations typées. Il comporte par ailleurs des relations d'indexation entre les concepts assimilables à des raccourcis d'enchaînementssignificatifs de relations. Ces index sont en particulier exploités par la propagation d'activité utilisée pour répondre aux demandes de rappel adressées à la mémoire. Enfin, la mémoire à long terme comporte également des unités permettant de rendre compte des regroupements prototypiques de concepts comparables par exemple aux MOPs de Schank. Ces unités sont liées aux concepts par des index et influencent donc également la propagation d'activité.

Compte tenu de cette structuration, le fonctionnement global de MOORE lors d'une demande de rappel (qui est le point nous intéressant ici) est le suivant. Le processus débute avec la réception de la demande de rappel, placée dans la file d'attente des requêtes à traiter. Une telle demande se compose de deux parties : d'une part un ensemble d'indices permettant de spécifier ce qui est recherché, et d'autre part un ensemble de critères chargés de déterminer si ce qui a été trouvé correspond à ce qui était recherché. Lorsque la requête de rappel arrive en phase de traitement, une propagation d'activité est déclenchée dans le réseau de la mémoire à long terme à partir des indices présents dans la requête. La requête passe alors en état d'attente de réponse. Il faut préciser qu'à tout moment après la réception d'une requête de rappel, MOORE peut recevoir une requête de mise-à-jour de cette requête initiale. Cette nouvelle requête se traduit alors par une nouvelle propagation d'activité dans la mémoire à long terme.

À chaque cycle, les unités les plus activées de cette mémoire à long terme sont sélectionnées et confrontées aux requêtes en attente de réponse. Si ces unités répondent aux critères associés à l'une de ces requêtes, elles sont proposées au demandeur comme réponse à sa requête. Celui-ci peut accepter la réponse, la juger non-satisfaisante, auquel cas sa requête est replacée dans la file des requêtes en attente de réponse, ou bien il peut annuler définitivement sa requête. La possibilité de demander un approfondissement de la réponse est intéressante car elle rend la recherche à la fois incrémentale et ouverte à un retour de l'utilisateur sur son résultat[1]. Lorsque le traitement d'une requête est achevé, qu'une réponse satisfaisante lui ait été apportée ou bien qu'elle ait été annulée, elle est utilisée afin d'ajuster les index au sein de la mémoire à long terme.

Dans ce mécanisme de rappel, nous sommes plus particulièrement intéressé par deux principes mis en avant par Francis et sous-tendant la propagation d'activité. Le premier d'entre eux est appelé *context focusing*. Il est issu de la constatation qu'il est trop coûteux, ainsi que nous le soulignions en préambule, de faire agir la propagation d'activité dans toute la mémoire à long terme et qu'il est donc nécessaire de restreindre son action à un sous-ensemble de celle-ci en rapport avec le contexte de la recherche. Francis a repris pour MOORE l'idée développée dans un autre de ses systèmes d'une propagation à double flux : des marqueurs d'un premier type sont utilisés pour délimiter l'espace dans lequel s'effectue ensuite la propagation de marqueurs d'un second type, chargés eux de l'activation proprement dite de la mémoire.

La délimitation opérée par le premier flux s'appuie sur la diffusion d'une quantité initiale d'activité qui, en faisant abstraction des phénomènes dissipateurs, reste globalement constante tout au long du processus de délimitation. Cette diffusion prend comme point de départ le contenu de la mémoire de travail. À mesure que l'activité se diffuse dans la mémoire à long terme, son niveau pour chaque nouvelle unité touchée est de plus en plus faible : d'une part le nombre des unités concernées devient de plus en plus grand, et d'autre part le passage par chaque nœud du réseau entraîne une certaine déperdition d'activité. Cette propriété, conjuguée à l'application d'un seuil en dessous duquel l'activité n'est plus propagée, permet de réaliser le *context focusing* visé. Il faut préciser que l'algorithme de gestion de ce premier flux n'a pas été implémenté par Francis.

Le second principe est partagé avec REMIND et justifie l'usage de la propagation d'activité. Il s'agit du *broadband indexing*. Cette notion est à rapprocher de celle d'indexation de surface. L'idée du *broadband indexing* est qu'il est préférable, pour un rappel efficace, d'utiliser le plus grand nombre possible d'indices concernant les entités

[1] On peut même imaginer un phénomène de va-et-vient conduisant progressivement à un équilibre entre ce que peut fournir la mémoire et ce qui est demandé par le module qui l'interroge.

que l'on conserve en mémoire. Ces indices doivent en conséquence recouvrir aussi bien les traits de surface de ces entités que ceux caractérisant leurs propriétés plus profondes et la causalité qui les régit. De ce point de vue, la propagation d'activation présente l'avantage d'offrir un mécanisme unique capable de s'affranchir des disparités existant entre des traits de différents types et de les mêler afin de réaliser le plus efficacement possible les recherches en mémoire. De cette façon, on ne présume pas de la présence ou de l'absence de certaines informations parmi les indices mis à disposition. La recherche peut s'effectuer en tout état de cause à partir des éléments disponibles. Il est bien entendu évident que plus ceux-ci sont nombreux et significatifs et plus le rappel sera rapide et son résultat intéressant.

3.3. Description du processus de rappel

Ainsi que pouvait le laisser supposer l'analyse réalisée au paragraphe précédent, le processus de rappel que nous avons associé à la mémoire épisodique se fonde sur un mécanisme de propagation d'activité. Dans ses grandes lignes, ce dernier est comparable au mécanisme existant dans REMIND pour gérer l' "evidential activation", en particulier ce qui concerne la régulation globale de l'activité, en y ajoutant une étape de délimitation de l'espace de sélection comparable au *context focusing* décrit pour MOORE.

Le processus de rappel de la mémoire épisodique reprend donc le principe du *broadband indexing* commun à ces deux systèmes. Celui-ci s'impose tout particulièrement dans le cas présent étant donné que l'apprentissage incrémental caractérisant la mémoire épisodique produit des représentations dans lesquelles par essence, traits de surface et traits plus profonds cohabitent étroitement, sans d'ailleurs que le statut des uns comme des autres ne soit franchement bien établi, au moins pendant le temps assez long de la genèse de ces représentations.

Nous allons maintenant entrer plus avant dans le détail de la description de ce processus de rappel.

3.3.1. Principes généraux

Le processus de rappel est conçu pour sélectionner les UTs agrégées de la mémoire épisodique qui sont les plus proches, compte tenu de l'état dans lequel se trouve cette mémoire, d'une proposition d'un texte présentée en tant qu'indice de rappel. L'état de la mémoire épisodique est en l'occurrence le résultat du traitement des propositions précédentes du texte considéré et se caractérise par le niveau d'activité des UTs agrégées composant la mémoire. Du point de vue du processus de rappel, le traitement d'un texte consiste à présenter successivement à la mémoire les différentes propositions qui le

constituent en suivant l'ordre que possèdent ces propositions au niveau de la forme de surface du texte. Plus précisément, on soumet la représentation sémantique de chacune de ces propositions en tant qu'indice d'évocation utilisable par la mémoire. Le rappel des UTs agrégées s'effectue donc à partir d'une configuration de type de concept.

Pour chaque proposition, ce rappel est initié par l'activation des types de concept qu'elle regroupe. La propagation d'activité qui en résulte s'effectue en deux phases. La première phase a pour objectif de délimiter l'espace de la mémoire dans lequel la sélection sera plus précisément menée afin de ne pas mobiliser inutilement l'ensemble de la mémoire. C'est la mise en œuvre du principe de *context focusing* exposé précédemment. Elle est réalisée par la diffusion d'un flux d'activité constant à la fois depuis les types de concept de la proposition courante et les UTs agrégées les plus activées à l'issue du traitement de la proposition précédente. Cette diffusion s'arrête lorsque le niveau d'activité de tous les nœuds du réseau de propagation atteints par le flux mais n'ayant pas encore émis leur activité est trop faible pour autoriser cette émission. Il faut préciser que la diffusion s'effectue à la fois au sein de la mémoire épisodique et du réseau sémantique constitué par le treillis des types de concepts et les connaissances associées à ces types. L'espace délimité comprend donc à la fois une partie de la mémoire épisodique et une partie du réseau sémantique.

La seconde phase prend place dans le sous-réseau de propagation défini par la première phase tout en laissant de côté les résultats de cette dernière en termes de niveau d'activité. Pour les UTs agrégées déjà considérées lors du traitement de la proposition précédente, on reprend leur niveau d'activité à l'issue de ce cycle antérieur. Ces valeurs matérialisent le contexte établi par les propositions précédentes. Les concepts de la proposition reçoivent quant à eux une valeur d'activité fixée à priori. Enfin, les autres entités de l'espace de sélection se voient attribuer une valeur d'activité nulle. À partir de ces valeurs initiales, on laisse ensuite l'activité du réseau de propagation évoluer naturellement en fonction des règles de propagation définies au niveau de chacun des types d'entités composant ce réseau.

Cette évolution se poursuit jusqu'à ce qu'un niveau de stabilité de l'activité des UTs agrégées, qui sont les entités à sélectionner dans le cas présent, ait été atteint ou dépassé. Ce niveau s'exprime par un pourcentage fixé à priori, en l'occurrence égal à 95% dans les expérimentations que nous avons menées ici. Cette stabilité est évaluée en pratique par la moyenne de deux termes complémentaires :

- le changement en pourcentage de l'activité globale des UTs agrégées par rapport au cycle précédent, autrement dit de la somme des valeurs d'activité de toutes les UTs agrégées situées dans l'espace de sélection;

- le changement moyen, également en pourcentage, de l'activité individuelle des UTs agrégées. Ce terme est plus précisément obtenu en sommant les différences de rang, d'un cycle à l'autre, des UTs agrégées de l'espace de sélection relativement à leur niveau d'activité.

Chacun de ces deux termes apparaît en réalité dans la formule de calcul de la stabilité sous la forme de son complémentaire : 100 - valeur du terme.

Plus formellement, le processus global de rappel du contenu de la mémoire épisodique peut s'exprimer par l'algorithme suivant :

```
graphe_courant ← analysePropositionSuivante(texte)
Tantque non fin(texte) faire
    mémoire_activité_UTAs ← mémorisation de la valeur d'activité des UTs agrégées de
                            espace_sélection
    -- première phase de propagation
    espace_sélection ← définitionEspaceSélection(réseau_propagation, graphe_courant,
                       espace_sélection)
    -- seconde phase de propagation
    réseau_propagation ← sélection(réseau_propagation, espace_sélection, graphe_courant,
                         mémoire_activité_UTAs)
    -- traitement de la proposition courante à la lumière des connaissances sélectionnées
    traitement(graphe_courant, activité(réseau_propagation))
    graphe_courant ← analysePropositionSuivante(texte)
Fin_tantque
```

avec les fonctions secondaires :

```
définitionEspaceSélection(réseau_propagation, graphe_courant, espace_sélection)
    UTAs_contexte ← sélection des UTs agrégées les plus activées de espace_sélection (on retient un
        quart des UTs agrégées les plus activées)
    remise à zéro de l'activité de toutes les entités de espace_sélection en dehors du contenu de
        UTAs_contexte
    activation des types de concept présents dans graphe_courant
    sources_diffusion ← UTAs_contexte + types de concepts composant graphe_courant
    Tantque non estVide(sources_diffusion) faire
        sources_diffusion ← cycleDiffusion(réseau_propagation,sources_diffusion)
    Fin_tantque
    espace_sélection ← sélection de toutes les entités touchées par le flux d'activité
    renvoie espace_sélection

sélection(réseau_propagation, espace_sélection, graphe_courant, mémoire_activité_UTAs)
    remise à zéro de l'activité des unités faisant partie de espace_sélection
    restauration du contexte des cycles précédents, c'est-à-dire de l'activité des UTs agrégées de
        mémoire_activité_UTAs
    activation des types de concept présents dans graphe_courant
    Tantque non stabilité(réseau_propagation, espace_sélection) faire
        cycleActivation(réseau_propagation, espace_sélection)
    Fin_tantque
    renvoie réseau_propagation
```

3.3.2. Structure du réseau de propagation

Dans le paragraphe précédent, nous avons utilisé le terme de réseau de propagation sans en donner une définition plus précise que son interprétation littérale, c'est-à-dire le réseau dans lequel la propagation d'activité se fait. Les principales entités composant la mémoire épisodique sont reliées les unes aux autres par des liens multiples : les épisodes agrégés font référence aux UTs agrégées, ces dernières sont elles-mêmes liées les unes aux autres et font par ailleurs référence aux types de concept par le biais des graphes agrégés qu'elles contiennent. L'ensemble forme donc un réseau, qui n'est cependant pas nécessairement connexe (cf. par exemple le cas d'une représentation de texte mémorisée sans qu'aucune de ses UTs ne s'agrège avec une UT agrégée). En ajoutant à ce réseau les liens existant entre les types de concept du fait de leur appartenance au treillis des types de la mémoire conceptuelle, on garantit en revanche la connexité du réseau formé et de ce fait, la possibilité pour un flux d'activité d'atteindre tout point du réseau à partir d'un autre.

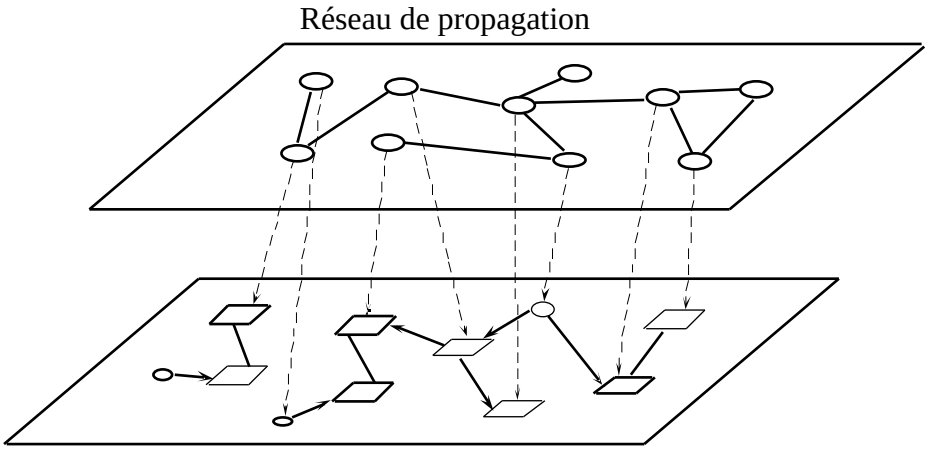


Fig. 6.5 - Relations entre le réseau de propagation d'activité et les mémoires épisodique et conceptuelle

Néanmoins, les structures de la mémoire épisodique, de même que celles de la mémoire conceptuelle, ne sont pas originellement conçues pour servir de support à une propagation d'activité. Ainsi que l'illustre la figure 6.5, il nous a semblé préférable de dissocier la représentation des entités constituant ces mémoires de l'exploitation du réseau qu'elles forment par le mécanisme de rappel des UTs agrégées. Ce réseau a donc été matérialisé par une structure autonome. Cette solution présente l'avantage de rendre possible le développement de structures de données et d'outils génériques dédiés au traitement des problèmes de propagation. On pourra à cet égard se reporter à l'annexe E pour avoir une brève description de l'environnement de test MALCOM que nous avons

construit pour mettre en œuvre et étudier ce type de réseau. Cette solution permet en outre de s'affranchir du problème de l'hétérogénéité de structure des différentes entités manipulées pour implanter le mécanisme de propagation d'activité, ce qui constitue aussi le gage d'une extensibilité aisée. La distinction entre les différents types de mémoire s'efface donc au sein du réseau de propagation.

Chaque entité pouvant jouer un rôle dans le rappel des UTs agrégées est représentée dans le réseau de propagation par un nœud unique. Le type de ce nœud est déterminé par la nature de l'entité représentée et définit son comportement, autrement dit la fonction d'activation utilisée par ce nœud pour évaluer son niveau d'activité en fonction de l'activité que lui envoient ses voisins. Chaque nœud du réseau de propagation conserve par ailleurs un lien de référence vers l'entité qu'il représente. Les connexions d'un nœud avec les autres nœuds du réseau sont le reflet des liens existant entre l'entité représentée par ce nœud et les entités, elles-mêmes représentées au niveau du réseau de propagation, avec lesquelles elle est liée au sein des mémoires épisodique et conceptuelle. Lorsqu'il

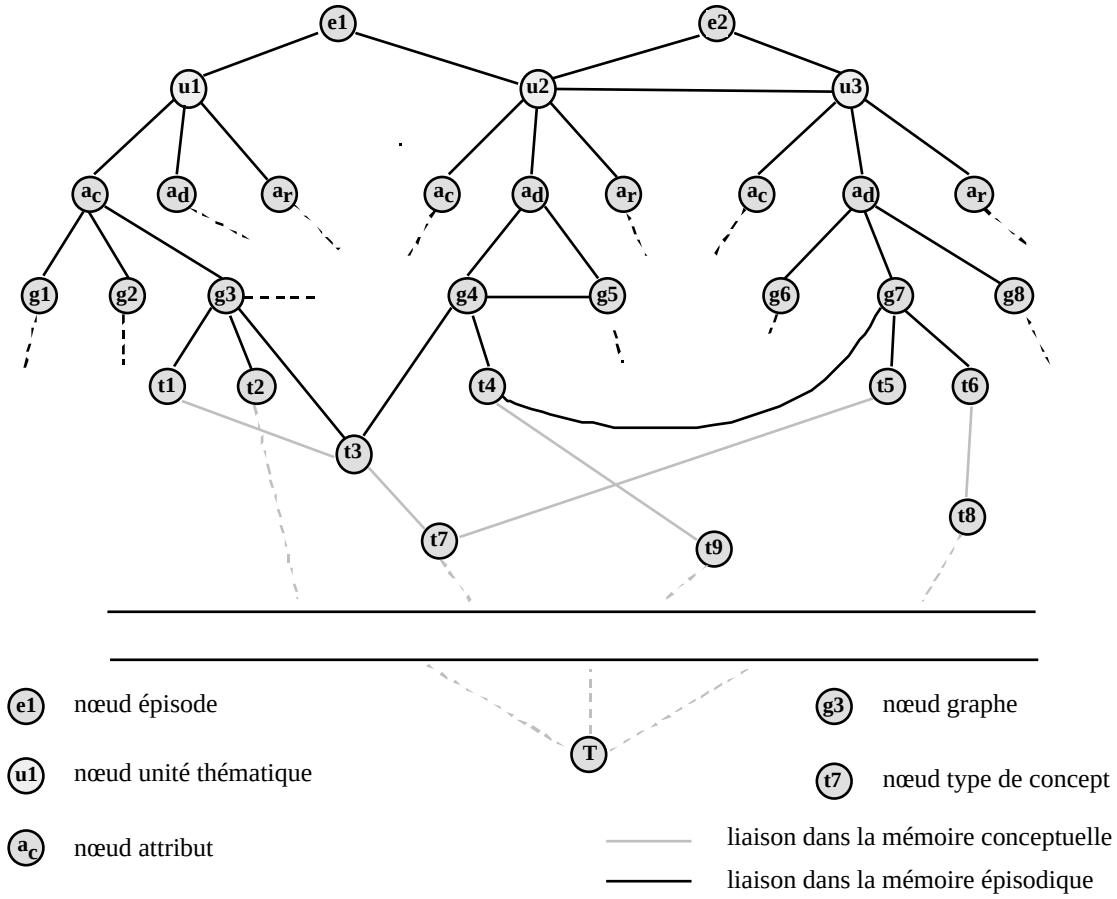


Fig. 6.6 - Aperçu de la structure du réseau de propagation

n'est pas fixé a priori comme dans le cas de la mémoire conceptuelle, le poids de ces connexions est établi sur la base du poids que possèdent les éléments de la mémoire

épisodique impliqués dans cette connexion. L'évolution continuelle de ces poids à mesure de la mémorisation de nouvelles représentations de texte oblige à ce que les poids des connexions ne soient pas calculés une fois pour toutes mais réévalués à la demande à partir des références que les nœuds du réseau de propagation conservent vers les éléments de la mémoire épisodique qu'ils représentent.

Les connexions sont d'autre part symétriques, ce qui signifie que le même poids est appliqué pour un sens de parcours d'une connexion comme pour l'autre sens. Le mode de calcul des poids au sein de la mémoire épisodique ne permet pas en effet de faire une distinction entre les deux points de vue possibles et nous ne souhaitions pas privilégier l'un plutôt que l'autre, c'est-à-dire avoir une approche plutôt montante (des concepts vers les UTs agrégées) ou au contraire plutôt descendante (des UTs agrégées vers les concepts). À notre avis, l'intérêt de la propagation d'activité réside justement dans sa capacité à mêler étroitement et naturellement les deux approches de façon à permettre l'établissement progressif d'un équilibre entre ces deux flux.

La figure 6.6 donne un aperçu de la structure de ce réseau de propagation en faisant plus précisément apparaître les différents types d'unités qui le forment ainsi que les liaisons permettant de véhiculer le flux d'activité entre ces unités.

Tout choix dans ce domaine nécessite au moins de faire figurer deux types d'entités : les UTs agrégées et les types de concept. Les premières sont en effet les entités à sélectionner tandis que les seconds sont les points de départ obligés de l'activité. À partir de cette base, on peut analyser la structure du réseau de propagation selon le point de vue des différentes relations possibles entre ces deux types d'entités :

- relations entre les types de concept et les UTs agrégées.

Si l'on excepte les rôles, laissés de côté ici[1], il n'existe pas de relation directe entre les UTs agrégées et les types de concepts au niveau de la mémoire épisodique. Il faut en effet passer par les deux structures intermédiaires que sont les attributs et les graphes conceptuels agrégés pour établir cette liaison. Comme on peut le voir sur la figure 6.6, nous avons choisi de faire figurer ces deux niveaux intermédiaires dans le réseau de propagation. Il aurait été plus simple, et pas fondamentalement différent du point de vue de la sélection, de lier directement les unités représentant les UTs agrégées et celles représentant les types de concept. La présence de ces deux niveaux rend toutefois le processus de sélection plus sensible à la signification des structures qu'il parcourt en offrant la possibilité de lui spécifier que tous les attributs d'une UT n'ont pas

[1] On a choisi de ne pas représenter les rôles dans le réseau de propagation en estimant qu'ils n'apportent pas d'informations nouvelles du point de vue de la sélection des UTs agrégées par rapport aux autres constituants de celles-ci. Les types de concepts qu'ils font apparaître sont en effet déjà présents au niveau des graphes de ces mêmes UTs.

nécessairement la même importance ou bien encore que l'élément central d'un graphe est son prédicat.

On obtient donc la structuration suivante : un nœud type de concept peut être lié à un nœud graphe, lequel est nécessairement relié à un nœud attribut, lui-même obligatoirement lié à un nœud UT agrégée. Un nœud type de concept ainsi relié de façon indirecte à un nœud UT agrégée fait référence au type d'un concept agrégé appartenant à un graphe de l'UT agrégée considérée. Rappelons que chacun de ces types généralise les types des différentes instances rassemblées par le concept agrégé. On ne fait pas figurer ceux-ci dans le réseau dans la mesure où ils pourront être activés par l'intermédiaire des liens issus du treillis des types de concept.

Au sein de cette ossature verticale liant les nœuds type de concept aux nœuds UT agrégée, il faut ajouter de possibles liens transversaux entre les nœuds graphe. Ces liens représentent les relations temporelles et/ou causales récurrentes intervenant entre les événements d'une UT agrégée. Précisons que ces relations constituent par la même occasion des relations transversales entre nœuds attribut puisque les graphes reliés n'appartiennent pas nécessairement au même attribut. En outre, dans le cas particulier des relations de causalité, le passage de l'activité dans le sens de la relation peut être interprété comme une déduction alors que son passage dans l'autre sens correspond d'une certaine façon à une abduction.

Le poids associé à la connexion entre un nœud type de concept et un nœud graphe est donné par le poids du concept agrégé correspondant par rapport au graphe dont il fait partie. Il est à noter que contrairement au cas général, la référence vers le concept agrégé n'est pas faite à partir du nœud type de concept mais à partir du lien entre celui-ci et le nœud graphe. Du point de vue du réseau de propagation, on ne représente en effet le concept agrégé que par son type, qui est une entité générale, non spécifiquement liée à l'UT agrégée à laquelle appartient le concept. Dès lors, la seule façon de conserver une référence vers le concept agrégé, donc vers son poids, est de faire porter celle-ci par la connexion entre le nœud graphe représentant le graphe d'appartenance du concept agrégé et le nœud type de concept référençant son type.

Le poids associé à la connexion entre un nœud graphe et un nœud attribut est donné quant à lui par le poids que possède le graphe représenté vis-à-vis de l'UT agrégée comprenant ce graphe. Pour sa part, le poids de la connexion entre un nœud attribut et un nœud UT agrégée est fixé a priori et ne varie pas ultérieurement, au contraire des autres poids impliquant des éléments de la mémoire épisodique. Ce poids rend compte de l'importance relative des trois attributs d'une UT agrégée. Dans le cas présent, nous avons décidé d'attribuer un poids de 1[1] à la connexion faisant intervenir les attributs

[1] Au sein de la mémoire épisodique, les poids sont normalisés entre 0 et 1. Compte tenu des relations entre ces poids et ceux de réseau de propagation, ces derniers reprennent le même intervalle de valeurs.

Description et *ÉtatsIncidents* tandis que ce poids n'est égal qu'à 0,75 pour celle concernant l'attribut *Circonstances*, traduisant ainsi sa moindre sélectivité par rapport aux deux autres.

Enfin, le poids de la connexion entre deux nœuds graphe obéit à une logique un peu différente de celle des cas précédents. Cette connexion représente en effet une relation matérialisée par une entité au sein de la mémoire épisodique. Le poids de la connexion n'est de ce fait pas établi en exploitant les références maintenues par les nœuds reliés mais en utilisant une référence, conservée au niveau de la connexion, vers la relation que cette connexion représente. Le poids de la connexion est égal à celui de la relation.

À l'exception des poids entre les nœuds attribut et les nœuds UT agrégée qui sont uniquement des modulateurs, il faut noter que les poids des connexions formant l'ossature verticale entre les nœuds type de concept et les nœuds UT agrégée peuvent être interprétés au même titre que dans REMIND comme des probabilités conditionnelles. Dans cette optique, le poids d'un graphe par rapport à une UT agrégée est vu comme la probabilité de sélectionner cette UT étant donné un certain degré d'activité de l'événement représenté par le graphe. Cependant, comme dans REMIND, la symétrie des relations remet en cause ce schéma d'interprétation dans la mesure où aucune raison théorique ne justifie l'égalité entre P(A/B) et P(B/A).

En revanche, les poids portés par les connexions entre nœuds graphe ne sont pas interprétables en ces termes. Une interprétation possible en ce qui les concerne consiste à les assimiler à des cœfficients d'incertitude associés à des règles d'inférence.

- relations entre les types de concept.

Les connexions entre les nœuds type de concept ont deux origines : le treillis des types de concepts et les graphes de définition associés à chacun de ces types. Dans les deux cas, l'intérêt de la présence de ces connexions dans le réseau de propagation réside dans la possibilité de mettre plus facilement en relation les concepts exprimés dans les textes avec ceux présents dans les UTs agrégées. Grâce à ces connexions en effet, lorsque le flux d'activité touche un type de concept, il active en même temps un ensemble de types de concept proches composés à la fois de ses sur-types, de ses sous-types ainsi que des types de concept servir à le définir. Les chances de pouvoir sélectionner les UTs agrégées en rapport avec un passage d'un texte en cours de traitement sont alors plus grandes, surtout lorsque la façon dont la situation est évoquée par le texte est assez éloignée de celles rencontrées précédemment à propos de la même situation. Globalement, les connexions entre les nœuds type de concept ont donc un rôle de passerelle et introduisent un degré de liberté dans l'appariement entre le contenu de la mémoire et le contenu des textes.

Dans le cadre de cette exploitation des connaissances sémantiques, il aurait été possible d'ajouter aux connexions provenant du treillis et des graphes de définition des connexions s'appuyant sur les graphes canoniques. Deux raisons nous ont conduit à juger que cet ajout n'était pas nécessaire. Tout d'abord, nous nous plaçons dans un contexte, ainsi que nous l'avons vu au chapitre 4, dans lequel les graphes canoniques expriment des contraintes assez générales, au contraire de systèmes tels que les systèmes d'extraction d'information par exemple, dans lesquels les structures équivalentes intègrent des connaissances très spécifiques par rapport au domaine abordé. Les types de concept présents dans ces graphes canoniques étant généraux, ils ne sont pas intéressants pour la sélection des UTs agrégées car ils ne sont pas discriminants. La seconde raison tient quant à elle à ce que les graphes présents dans les UTs agrégées dérivent nécessairement des graphes canoniques. Compte tenu des connexions issues du treillis des types de concept permettant de propager en direction de concepts plus généraux, les graphes canoniques n'apportent donc pas d'informations supplémentaires permettant d'étendre le champ de la sélection.

Contrairement au cas des relations entre types de concept et UTs agrégées, la matérialisation des relations entre types de concept ne donne pas lieu à des structures intermédiaires. En ce qui concerne le treillis, chaque nœud type de concept est connecté dans le réseau de propagation à chacun des nœuds représentant un de ses sous-types ou sur-types directs. La connexion est symétrique et possède un poids fixe de valeur 1. Bien entendu, il ne s'agit pas d'activer systématiquement tout le treillis à partir d'un seul type de concept. Nous verrons au paragraphe 3.3.4 comment la fonction d'activation associée aux nœuds type de concept permet de traiter ce problème. Le cas des graphes de définition est traité de manière similaire. Un nœud type de concept est en effet connecté à chacun des nœuds représentant les types de concepts apparaissant dans le graphe de définition du type de concept qu'il référence. Cette connexion est comme précédemment symétrique et dotée d'un poids fixe de 1. Il faut préciser que si plusieurs concepts de la définition possèdent le même type – il s'agit souvent de types assez généraux, donc pas particulièrement significatifs – celui-ci est traité comme s'il n'était présent qu'une seule fois. Le seul type de la définition laissé de côté est le type par différenciation duquel celui que l'on considère est défini. Les deux nœuds correspondant dans le réseau de propagation sont en effet déjà liés par l'intermédiaire des relations issues du treillis.

- relations entre les UTs agrégées.

Ces relations sont de deux types. On distingue les relations directes, incarnées par les relations de suivi thématique, et les relations indirectes, passant par la médiation des épisodes agrégés. Dans ce dernier cas, on considère en fait qu'il existe des relations

implicites entre toutes les UTs agrégées appartenant à un même épisode agrégé. Du point de vue de la propagation d'activité, les relations entre les UTs agrégées, quel que soit leur type, offrent la possibilité d'étendre le champ de la sélection au sein même de la mémoire épisodique, ce qui est a priori plus sélectif que de réaliser la même chose par l'intermédiaire de la mémoire conceptuelle[1].

En ce qui concerne les relations thématiques, nous avons choisi de ne faire figurer dans le réseau de propagation que les relations de type déviation. Les relations de type changement de thème marquent en effet une rupture thématique qu'il ne semble pas logique d'exploiter pour activer une UT agrégée à partir d'une autre. Les relations de type déviation sont représentées comme une simple connexion entre les deux nœuds représentant les UTs agrégées présentes à ses deux extrémités. Comme dans le cas des connexions entre graphes, ces relations sont des entités à part entière au sein de la mémoire épisodique. La connexion fait donc référence directement à la relation qu'elle représente et son poids, égal à celui de la relation, est recherché dynamiquement en passant directement par cette référence et non par les entités reliées. Il est à noter enfin que du fait de l'autonomie des relations thématiques, ce poids ne s'interprète pas plus en termes de probabilités conditionnelles que celui entre nœuds graphe.

La représentation dans le réseau de propagation des relations d'appartenance d'une UT agrégée à un épisode agrégé présente pour sa part l'intérêt d'introduire un contexte de sélection plus large que celui des UTs agrégées, même si celles-ci demeurent au cœur du processus de rappel. L'activation d'un épisode agrégé permet d'activer un ensemble d'UTs agrégées apparaissant ensemble de façon récurrente, ce qui est un moyen par exemple de repérer plus facilement les évolutions thématiques, en particulier lorsqu'il s'agit de déviations. En pratique, les épisodes agrégés sont représentés dans le réseau de propagation par un type d'unité particulier, les nœuds épisodes agrégés, et l'appartenance d'une UT agrégée à un épisode agrégé est figurée comme précédemment pour ce type de relation par une connexion entre les nœuds du réseau référençant ces deux entités. Le poids de cette connexion est égal au poids de l'UT agrégée relativement à cet épisode agrégé et son actualisation dynamique est réalisée en y accédant par l'intermédiaire des références conservées au niveau des deux nœuds impliqués dans la connexion.

La structure que nous avons décrite ci-dessus spécifie la forme que doit avoir le réseau de propagation permettant d'exploiter une mémoire épisodique donnée. Mais au contraire de la mémoire conceptuelle, la mémoire épisodique évolue constamment à mesure qu'elle

[1] La plus grande sélectivité des relations au sein de la mémoire épisodique serait encore plus opérante si cette mémoire était plus structurée. L'argument est néanmoins valide d'un point de vue général dans la mesure où l'activité se propageant dans la mémoire conceptuelle ne fait qu'étendre, de manière aveugle par rapport aux UTs agrégées, le champ de la sélection en généralisant et en spécialisant les types de concept.

mémorise de nouvelles représentations de texte. Nous avons déjà partiellement pris en compte ce problème en ne faisant pas apparaître les poids des connexions du réseau sous la forme de valeurs fixes mais plutôt sous la forme de références à des poids associés aux constituants de la mémoire épisodique. Pour conserver un parallélisme parfait entre celle-ci et le réseau de propagation, il est nécessaire d'adjoindre à ce premier mécanisme une mise à jour automatique de la structure du réseau lorsqu'un changement de structure intervient au sein de la mémoire épisodique, c'est-à-dire dès qu'un nouvel épisode, une nouvelle UT, un nouveau graphe ou même une nouvelle relation thématique ou une relation inter-graphe est ajouté. Cette phase de mise à jour du réseau de propagation doit intervenir à la suite de chaque mémorisation d'une nouvelle représentation de texte.

3.3.3. Un exemple

Avant de présenter plus complètement la façon dont l'activité est propagée dans le réseau de propagation, nous allons illustrer sur un bref exemple l'intérêt de l'utilisation de la propagation d'activité pour le rappel associatif des constituants de la mémoire épisodique. Dans cet exemple, seule la seconde phase du rappel, c'est-à-dire la phase de sélection proprement dite, est mobilisée étant donné que la taille de la mémoire considérée ne nécessite pas la définition d'un contexte plus restreint. Par ailleurs, la structure du réseau de propagation est un peu plus simple que celle présentée ci-dessus : elle ne comprend pas de nœud épisode agrégé, donc pas de connexion entre les nœuds de ce type et les nœuds UT agrégée, ni de connexion entre les nœuds graphe.

La figure 6.7 montre comment les éléments d'une mémoire épisodique, en l'occurrence quatre graphes agrégés[1], sont activés en fonction du contenu des propositions d'un texte qui lui sont soumises et comment cette activité évolue à mesure que l'on progresse dans ce même texte. Cet exemple s'appuie sur le petit texte suivant :

> Hier, je suis allé faire des courses. J'ai acheté une pièce détachée pour réparer ma voiture. Mais finalement, je l'ai laissée à réparer au garagiste.

Les deux graphes agrégés Acheter_Objet et Faire_Courses font partie d'une UT agrégée appelée *Aller_Supermarché*. Réparer_Voiture et Réparer_Moteur font partie quant à eux de l'UT agrégée Réparer_Voiture[2]. La figure montre que, parallèlement au changement de thème présent dans l'histoire (on passe du thème des courses à celui de la réparation automobile), on observe un changement parmi les éléments les plus activés de

[1] Ces graphes ont été sélectionnés comme les plus représentatifs, du point de vue de leur activité, des deux principales UTs agrégées mises en jeu dans cet exemple.

[2] Les noms des UTs et des graphes agrégés considérés sont attribués sur la base d'une interprétation extérieure de leur contenu pour une plus grande commodité de désignation et ne sont le produit d'aucun processus automatique.

la mémoire. Après l'activation des types de concept de la première phrase, les éléments les plus activés font partie de l'UT *Aller_Supermarché*. À l'inverse, les graphes de l'UT agrégée *Réparer_Voiture* ne sont pas activés car rien n'évoque la situation d'une réparation automobile.

Lorsque la deuxième phrase est introduite (après stabilisation de l'activité du réseau, c'est-à-dire après la 22^{e} itération), les graphes de l'UT *Aller_Supermarché* restent fortement activés, ce qui est le résultat de deux courants d'influence. Tout d'abord, l'état d'activité antérieur de la mémoire crée un contexte qui contraint l'évolution de son activité. Ensuite, les types de concept venant de la nouvelle phrase (dans le cas présent, le type de concept Acheter) apportent un renforcement du thème actif. Nous pouvons en même temps observer que d'autres types de concept de cette deuxième phrase (en l'occurrence, les types de concept Pièce_détachée, Voiture et Réparer) font apparaître un nouveau thème. Cela explique pourquoi les graphes de l'UT agrégée *Réparer_Voiture*, ainsi que l'UT elle-même, commencent à être activés plus fortement.

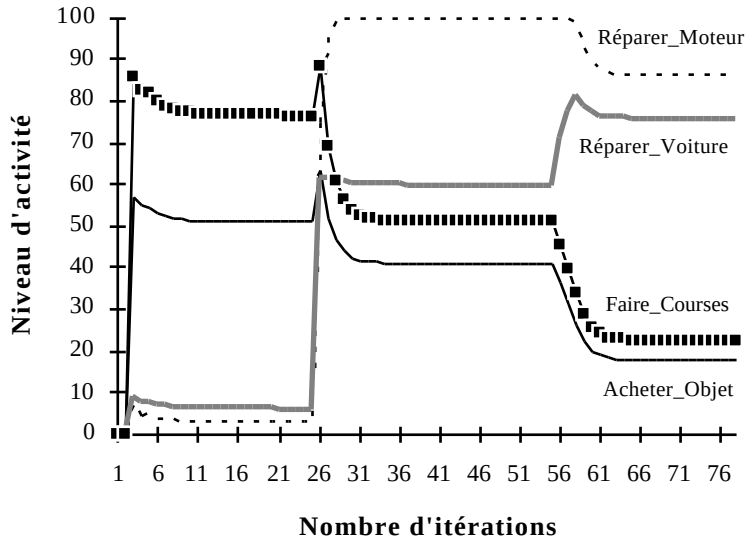


Fig. 6.7 - Activité de 4 graphes agrégés lors d'une phase de rappel

Après la prise en compte de la troisième phrase (après la 52^{e} itération), l'effet de contexte de la mémoire relatif au premier thème n'est plus significatif (la confirmation précédente n'ayant pas été très forte) et aucun élément de cette troisième phrase n'est particulièrement lié à ce thème. Les types de concept qui la constituent viennent au contraire renforcer le second thème. On observe ainsi que l'activité des éléments de l'UT *Aller_Supermarché* décroît de façon très importante tandis que l'effet inverse se produit pour les éléments de l'UT *Réparer_Voiture*, jusqu'à obtenir une stabilisation de la configuration d'activité à la 78^{e} itération. Dans cette configuration, la prédominance

finale de Réparer_Moteur par rapport à Réparer_Voiture s'explique sans doute par le fait qu'il est à la fois question de voiture et de pièce détachée (le moteur étant considéré comme une pièce détachée).

3.3.4. Description de la propagation d'activité

Phase de délimitation de l'espace de sélection

La délimitation d'un espace de sélection à partir des types de concept de la proposition courante et des UTs agrégées les plus activées de la mémoire épisodique consiste à rechercher les constituants du couple mémoire épisodique - mémoire conceptuelle les plus proches de ces éléments de départ. Ce couple possédant une structure de graphe (cf. §3.3.2), cette recherche est assimilable au parcours en largeur d'abord d'un graphe. Cette façon de faire repose sur l'hypothèse que deux entités du couple de mémoires mentionné ci-dessus sont d'autant plus proches qu'elles entretiennent davantage de relations, si possible assez directes. Cette hypothèse est acceptable à la fois pour la mémoire conceptuelle et pour la mémoire épisodique. Dans le cas de la première, le degré d'éloignement de deux types de concept est directement relié au nombre de relations qu'il faut parcourir dans le treillis des types pour les joindre. Les constituants de la seconde (épisodes, UTs, attributs, graphes) se ramènent quant à eux à des ensembles de types de concept dont le taux de recoupement, donc le nombre de liens, est corrélé avec leur degré de similarité.

Le parcours de graphe sous-tendant la délimitation d'un espace de sélection doit cependant être sensible aux poids associés aux relations entre les constituants des mémoires épisodique et conceptuelle afin de retenir prioritairement ceux qui sont les plus fortement liés aux éléments utilisés comme amorce. C'est l'objet du mécanisme de propagation adopté pour cette première phase du rappel.

Ce mécanisme est fondé, ainsi que nous l'avons esquissé au §3.3.1, sur la propagation d'un flux d'activité constant. Au niveau global, ce mode de fonctionnement se caractérise par le fait que la quantité d'activité circulant dans les connexions du réseau de propagation reste la même à tout moment. Pour ce faire, il est nécessaire que localement, la quantité d'activité entrant dans chacun des nœuds du réseau soit égale à la quantité d'activité sortant de ce même nœud. Lorsqu'un nœud est atteint par le flux d'activité, il réalise donc la sommation de ses entrées et redistribue cette quantitéd'activité vers ses sorties. Afin d'orienter préférentiellement l'activité vers les nœuds les plus fortement liés, cette redistribution est réalisée proportionnellement au poids des connexions jouant le rôle de sorties du nœud. Chaque nœud n'est traversé par le flux d'activité qu'une seule fois, ce qui garantit la convergence du processus global.

Autrement dit, un nœud ne réémet jamais d'activité en direction d'un nœud ayant déjà été touché par le flux d'activité. La valeur d'activité d'un nœud est ainsi égale à 0 tant qu'il n'a pas été atteint par le flux d'activité et reste égale à la valeur d'activité l'ayant traversé après le passage de ce dernier.

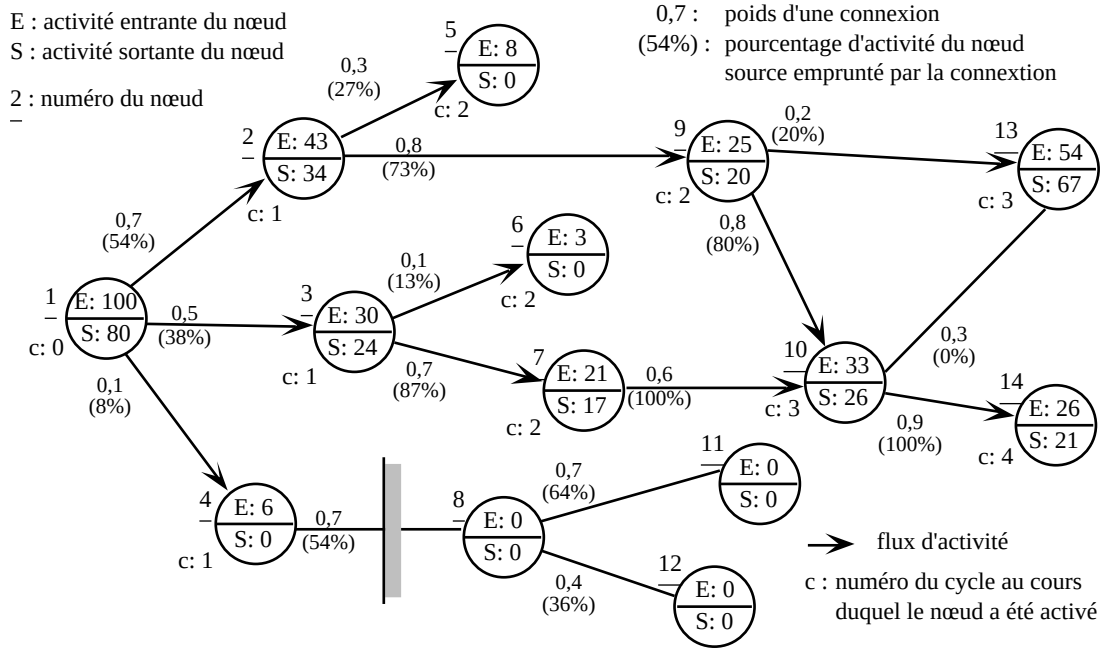


Fig. 6.8 - Principes de la propagation à flux constant

Ces principes sont illustrés par l'exemple de la figure 6.8 dans lequel une quantité d'activité égale à 100 est placée initialement au niveau du nœud 1 (cycle 0) et diffuse dans le réseau pour finalement toucher tous ses nœuds après 4 cycles. Chaque cycle correspond au franchissement d'une connexion pour l'ensemble des nœuds actifs. Un nœud actif est un nœud ayant reçu de l'activité au cycle précédent et qui réémet cette activité lors du cycle courant. À l'issue du cycle 1 par exemple, les nœuds actifs du réseau ci-dessus sont les nœuds 2, 3 et 4. L'activité en sortie du nœud 1 a été répartie lors de ce cycle entre ces trois nœuds suivant les poids figurant sur les connexions. Le poids de la connexion vers le nœud 2 correspond à 54% de la somme de ces trois poids et en conséquence, 54% de l'activité en sortie du nœud 1 est dirigée vers le nœud 2.

Les entrées et les sorties d'un nœud ne sont pas définies au niveau de la structure du réseau mais conditionnées par le sens de propagation de l'activité, celui-ci étant lui-même dépendant de l'endroit où se situent le ou les sources initiales de cette activité. Par exemple, les nœuds 1, 2, 3, 7, 9, 10 et 13 du réseau de la figure 6.8 verraient beaucoup de leurs connexions changer de rôle (passer du rôle d'entrée à celui de sortie et vice versa) si le point d'activation initial était le nœud 13 au lieu du nœud 1. Les entrées d'un nœud sont en fait constituées de ses connexions en provenance de nœuds actifs tandis que ses

sorties sont ses connexions restantes pointant vers des nœuds n'ayant pas encore été touché par le flux d'activité. Au cycle 3, les entrées du nœud 10 sont ainsi formées des nœuds 7 et 9 tandis que ses sorties, exploitées au cycle 4, ne comprennent que le nœud 14. Le nœud 13 n'est pas retenu comme sortie puisqu'ayant été atteint par le flux d'activité au cycle 3.

Le réseau de propagation étant une entité de taille finie, son activité ne peut pas en réalité rester constante et se trouve condamnée à diminuer progressivement pour finir par devenir nulle après avoir traversé tous les nœuds du réseau. En effet, une quantité d'activité parvenant à un nœud n'ayant plus de sortie est perdue et le nombre de ces nœuds augmente irrémédiablement à mesure que le nombre de nœuds visités par le flux d'activité augmente lui aussi. Les nœuds 13 et 14 du réseau ci-dessus sont une illustration de ce cas de figure[1].

Même si cette propriété garantit que la propagation d'activité considérée est un processus parvenant à un état stable en un temps borné, cet état stable n'est pas satisfaisant pour nous dans la mesure où il conduit à l'activation de tous les nœuds du réseau alors que notre but est de ne délimiter qu'une petite partie de celui-ci, sans avoir à en explorer la totalité. C'est pourquoi nous avons adjoint au processus décrit précédemment deux mécanismes complémentaires allant dans ce sens et transformant la propagation à flux constant en propagation dissipatrice à flux constant.

Le premier de ces deux mécanismes consiste à fixer un seuil d'activité en dessous duquel l'activité d'un nœud est jugée trop faible pour être redistribuée vers les nœuds de sortie. Ce seuil est équivalent au facteur θ présent dans la fonction d'activation de REMIND. À mesure que l'activité se diffuse dans le réseau de propagation, sa valeur moyenne par nœud devient de plus en plus faible étant donné que le nombre de nœuds touchés devient de plus en plus important et qu'au mieux, la quantité d'activité globale reste constante. Un nœud est en effet lié à un ensemble d'autres nœuds, eux-mêmes liés chacun à un ensemble d'autres nœuds et ainsi de suite. En fixant un seuil d'activité en dessous duquel on stoppe la propagation, on peut s'attendre à ce que le flux d'activité ne parcourt qu'une zone limitée autour des points initiaux de diffusion. Dans le cas du réseau de la figure 6.8, ce seuil, fixé à 10, est à l'origine de l'arrêt de la propagation en sortie du nœud 4 et explique que les nœuds 8, 11 et 12 ne soient pas atteints par l'activité. Les nœuds 5 et 6 sont également touchés par ce passage en dessous du seuil mais sans conséquence pour d'autres nœuds puisqu'ils n'ont plus de sortie possible.

[1] C'est le cas également des nœuds 5 et 6 mais dans ces deux cas, un autre phénomène que nous détaillerons un peu plus tard intervient également.

À lui tout seul, ce mécanisme est cependant insuffisant. Tout d'abord, la garantie de ne pas activer la plus grande partie, voire l'intégralité du réseau n'est pas assez forte. De fait, l'hypothèse de l'augmentation constante du nombre de nouveaux nœuds traversés par le flux d'activation lors d'un cycle, et donc, de la division en conséquence de l'activité, est exacte pour une structure d'arbre mais elle est déjà plus discutable dans le cas d'un graphe, notamment lorsque la densité de liens entre ses nœuds est importante. Une forte densité de liens favorise par ailleurs les configurations telles que celle du nœud 10, dans lesquelles des flux d'activité s'additionnent (flux venant des nœuds 7 et 9), ce qui permet à cette activité d'aller plus loin.

La seconde raison de cette insuffisance est la relative incertitude concernant le contrôle de l'extension de l'espace de sélection délimité. Les deux paramètres disponibles pour ce faire, à savoir la quantité initiale d'activité injectée et le seuil en dessous duquel l'activité n'est plus propagée, ne rendent pas en effet très facile l'appréciation du nombre maximum de nœuds d'un chemin (qui est équivalent au nombre de cycles avant stabilisation) et surtout, ne fournissent pas de garantie quant à ce nombre.

Afin d'obtenir un contrôle plus sûr du *context focusing*, nous avons choisi d'introduire un facteur de dissipation de l'activité au niveau des nœuds du réseau. Dans le cas du réseau de la figure 6.8, l'activité qui sort d'un nœud est ainsi 20%[1] plus faible que celle qui y pénètre. De cette façon, on est certain qu'au bout de n cycles, l'activité globale est au plus égale au produit $A_0 \cdot 0{,}8^n$, avec A_0, la valeur initiale de l'activité globale. Cela permet d'estimer mais également de contrôler plus finement le nombre de cycles nécessaires à la stabilisation, et donc l'extension de la zone de sélection délimitée.

Aux principes de base décrits ci-dessus, nous avons ajouté la possibilité de bloquer la diffusion de l'activité au niveau de certains nœuds afin de tenir compte de contraintes liées à la signification des structures traversées. Ce type de mécanisme prend partiellement en charge ce qui est traité dans REMIND par le réseau de propagation des signatures. Nous considérons ainsi qu'un graphe ne peut pas être activé à partir des types de concept qui le composent si aucune activité n'est issue de son prédicat. De même, un épisode agrégé ne doit être activé que si le flux d'activité provient de l'une au moins de ses UTs principales.

Les nœuds de type graphe et de type épisode agrégé intègrent donc une fonction d'activation plus complexe que les autres, chargée de vérifier ces contraintes. Au niveau d'un nœud épisode agrégé, elle se contente de vérifier que parmi toutes les entrées du nœud, l'une au moins provient d'un nœud représentant l'une de ses UTs principales. Le cas des nœuds graphe est un peu plus complexe car ceux-ci peuvent être activés par des nœuds type de concept mais également par d'autres nœuds graphe ou par un noeud

[1] Le chiffre de 20% est donné de façon purement indicative.

attribut. La fonction d'activation ne vérifie en fait la contrainte portant sur le prédicat que lorsque les entrées ne proviennent que de nœuds type de concept. Dans le cas des graphes comme dans celui des épisodes agrégés, si la contrainte n'est pas vérifiée, le nœud n'est pas activé et l'activité initialement orientée dans sa direction est redistribuée vers les autres connexions de sortie des nœuds qui en sont la source. Un tel blocage ne signifie pas toutefois que le nœud graphe ou le nœud épisode agrégé concerné ne pourra pas être activé lors d'un cycle ultérieur.

La dernière particularité à évoquer concernant la propagation d'activité est relative à la mémoire conceptuelle. La possibilité d'effectuer des généralisations ou des spécialisations en propageant de l'activité dans le treillis de types de concept est particulièrement intéressante afin de ne pas être prisonnier de similarités très formelles et restreintes. Cependant, elle doit aussi être maniée avec précaution, notamment pour éviter des sur-généralisations injustifiées conduisant à sélectionner un ensemble de connaissances trop important. Pour éviter un tel phénomène, nous préconisons d'utiliser la dissipation d'activité réalisée au niveau des nœuds comme moyen de contrôle de la propagation d'activité dans le treillis des types de concept.

Plus précisément, le treillis est arbitrairement découpé en trois zones d'égales hauteurs, caractérisées chacune par un taux de dissipation d'activité spécifique et d'autant plus grand que la zone en question contient des types de concept plus généraux. Le premier tiers regroupe ainsi les types de concept les plus généraux au sein desquels la propagation doit être très limitée (de l'ordre d'une connexion) : en propageant vers un sur-type, on perd en effet beaucoup de spécificité, donc d'information, et en propageant vers les sous-types, on risque d'activer un ensemble de types beaucoup trop important. Le deuxième et le troisième tiers du treillis contiennent des types de concepts de plus en plus spécifiques pour lesquels on autorise une propagation d'activité de plus en plus étendue, à la fois vers les sur-types et vers les sous-types (de l'ordre de deux connexions pour la deuxième zone et de trois connexions pour la troisième).

Cette différenciation du taux de dissipation de l'activité complique un peu la fonction d'activation des nœuds type de concept. Celle-ci doit d'abord différencier les entrées provenant d'autres nœuds type de concept de celles issues de nœuds graphe. Dans le premier cas, elle doit en outre déterminer à quelle partie du treillis appartient le type correspondant afin d'adopter le taux de dissipation adéquat. À titre d'indication, ce taux pourrait être de l'ordre de 80% dans le premier tiers, de 40% pour le deuxième tiers et de 20% pour la partie la plus spécifique du treillis, c'est-à-dire un taux comparable à celui utilisé dans l'exemple de la figure 6.8 pour le reste du réseau de propagation.

Une fois que le réseau de propagation a atteint un état stable, il reste encore à délimiter exactement le futur espace de sélection à partir du résultat de la propagation d'activité. Parmi toutes les entités touchées par le flux d'activité, nous retenons plus précisément :

- toutes les UTs agrégées activées à l'exception de celles dont ni l'attribut *Description* ni l'attribut *ÉtatsIncidents* n'ont été activés ou suffisamment activés. Un attribut est jugé comme insuffisamment activé lorsqu'aucun de ses graphes n'a lui-même été activé. On s'assure ainsi qu'une UT agrégée a été activée avec suffisamment de force et sur la base de ses éléments les plus significatifs. Lorsqu'une UT agrégée est ainsi retenue pour figurer dans l'espace de sélection, tous les éléments qui la composent sont évidemment sélectionnés avec elle;

- tous les épisodes agrégés à condition qu'ils soient reliés à au moins une UT agrégée sélectionnée. Au contraire de celles-ci, seul le nœud représentant l'épisode est retenu. La sélection des UTs agrégées qui en forment le contenu est indépendante;

- tous les types de concept activés, dans la mesure où ils sont liés à au moins une des UTs agrégées retenues. Ce lien peut être direct ou bien passer par l'intermédiaire d'un ou plusieurs types de concept activés.

On obtient ainsi un espace de sélection dont le réseau de propagation, bien que non nécessairement connexe du fait de la sélection de parties seulement du treillis des types, s'avère l'être dans une grande majorité des cas.

Sélection [1]

La seconde phase du processus de rappel, appelée phase de sélection, a pour but de valuer les UTs agrégées faisant partie de l'espace de sélection en fonction de leur degré d'adéquation vis-à-vis à la fois du contexte courant, représenté par les types de concept d'une proposition soumise en entrée de la mémoire épisodique, et du contexte établi par les propositions déjà traitées, incarné par le niveau d'activité que les UTs agrégées de l'espace de sélection possédaient préalablement à la nouvelle session de rappel. Cette phase est logiquement suivie d'une sélection par le processus demandeur du rappel des UTs agrégées les plus activées, l'ampleur de cette sélection étant définie par le processus en question.

Cette seconde phase est fondée comme la première sur un mécanisme de propagation d'activité. Les principes de cette propagation sont néanmoins assez différents de ceux de

[1] Nous tenons à remercier Bruce Bardou pour le travail d'expérimentation qu'il a mené sur ce problème dans le cadre de son stage de DEA. Nos propositions reprennent pour l'essentiel celles qu'il a formulées dans [Bardou 1995].

la première phase et s'apparentent plutôt à ceux que REMIND met en œuvre afin de gérer l' "evidential activation". À chaque cycle, l'activité de l'ensemble des nœuds du réseau de propagation de l'espace de sélection est réévaluée. Chaque nœud ne possède ni entrées, ni sorties, ou plus exactement ses connexions, symétriques, prennent alternativement le statut d'entrées lorsqu'il s'agit d'évaluer l'activité du nœud et de sorties lorsqu'il s'agit de recalculer l'activité de ses voisins. Le calcul de l'activité des nœuds du réseau s'effectue de façon synchrone : au sein d'un cycle, chaque nœud réévalue son niveau d'activité en fonction de l'activité qu'avaient ces voisins au cycle précédent.

Les caractéristiques de la propagation d'activité associée à cette seconde phase font du réseau de propagation un réseau récurrent, c'est-à-dire un type de réseau dont la dynamique dans le temps, relevant de la théorie des systèmes dynamiques, est beaucoup plus variée et complexe que celle des réseaux à couches, comme celui utilisé lors de la première phase du rappel[1]. En particulier, il est souvent difficile de préciser les conditions permettant d'aboutir à un état stable. Cette tâche est d'autant plus ardue pour des réseaux tels que ceux que nous manipulons qu'ils ne possèdent pas véritablement de structure régulière. Dans le cas de REMIND, les conditions de convergence ne sont d'ailleurs pas explicitées mais il est probable que le réseau de propagation des signatures et le mécanisme de blocage des flux qui l'accompagne jouent un rôle non négligeable dans l'atteinte d'un état stable.

Pour faire face à ce problème de convergence, nous sommes intervenu à deux niveaux. Concernant d'abord l'activité globale du réseau, nous avons repris le procédé de régulation adopté dans REMIND afin d'éviter une inflation constante de cette activité due à l'absence de liens inhibiteurs dans le réseau de propagation. Ce procédé consiste à diviser à chaque fin de cycle l'activité de chaque nœud par la moyenne des activités de tous les nœuds du réseau. Outre le fait d'éviter l'inflation mentionnée ci-dessus, cette méthode présente l'avantage de ramener toutes les valeurs d'activité sur une base commune significative[2].

Le second choix effectué en vue de garantir une stabilisation de l'état du réseau se situe au niveau de la fonction d'activation associée à ses nœuds. Celle-ci est en effet le fondement d'une dynamique d'évolution de l'activité du réseau en deux temps, avec un passage très progressif de l'un à l'autre. Le réseau dispose en premier lieu d'une grande

[1] Le réseau de la première phase est un réseau à couches non du fait de sa structure, comme dans le cas d'un perceptron multi-couches par exemple, mais de par la façon dont l'activité le parcourt. C'est en effet la propagation d'activité à flux constant qui définit dynamiquement les différentes couches en fonction des points d'injection de l'activité et de la topologie du réseau.

[2] Précisons que dans l'exemple du §3.3.2, les valeurs d'activité des graphes agrégés sont données avant division par la moyenne des activités.

liberté d'évolution qui lui permet de procéder à la valuation proprement dite des UTs agrégées en explorant l'espace des états qui lui sont accessibles. Cette liberté est ensuite de plus en plus contrainte afin de figer les choix opérés en faisant converger le réseau vers l'état représentatif de ces choix.

Cette dynamique est comparable dans une certaine mesure à celle d'une machine de Boltzmann. La différence entre les deux tient à la manière dont la liberté est de plus en plus contrainte : dans le cas d'une machine de Boltzmann, on diminue progressivement la probabilité d'accéder à un état augmentant l'énergie du réseau tandis que dans notre cas, ce figement est obtenu par la diminution progressive de la prise en compte de l'activité venant de l'extérieur des nœuds.

En pratique, cette diminution est traduite dans la fonction d'activation des nœuds UT agrégée (cf. [1]) par l'introduction d'une division de la classique somme pondérée des entrées par le carré du facteur temps, représenté ici par le nombre de cycles écoulés depuis le lancement de la propagation d'activité :

$$A_i(t+1) = \frac{\sum_j w_{ij} \quad A_j(t)}{t^2} + A_i(t) \qquad [1]$$

avec $A_i(t)$: niveau d'activité du nœud *i* après *t* cycles,

w_{ij} : poids associé à la connexion entre le nœud *i* et le nœud *j*.

On notera également l'importance du second terme de cette fonction puisqu'il représente la mémoire de chaque nœud UT agrégée. À ce titre, il met en œuvre la capacité du réseau à tenir compte, pour la valuation des UTs agrégées, du contexte établi par les phases de propagation précédentes. C'est également lui qui rend possible, au sein d'une phase de propagation, le processus de figement progressif de l'activité d'un nœud en conservant le produit des cycles déjà écoulés.

Jusqu'à présent, nous n'avons explicité que le mode de fonctionnement des nœuds UT agrégée. En fait, nous avons même opéré un raccourci en identifiant l'état du réseau de propagation tout entier à l'état de ses nœuds UT agrégée. Ce raccourci est motivé par le fait qu'en final, seule l'activité des nœuds UT agrégée nous intéresse. Dans le processus de valuation de ceux-ci, les autres nœuds jouent essentiellement un rôle de passerelle destiné à faire circuler l'activité entre les nœuds UT agrégée. La stabilisation de leur niveau d'activité est d'une certaine façon une conséquence indirecte de la stabilisation de l'état des nœuds UT agrégée. Ces derniers ayant une position centrale dans la structure du réseau de propagation, la stabilisation de leur état entraîne en effet la stabilisation de l'état des autres nœuds en présence. Le rôle de passerelle dévolu aux nœuds épisode agrégé, graphe, attribut et type de concept se traduit par une fonction d'activation se contentant de

restituer l'activité reçue en tenant compte de la modulation résultant du passage par les connexions :

$$A_i(t+1) = \sum_j w_{ij} A_j(t)$$

Il faut ajouter à ce principe général une particularité touchant l'activation des types de concept faisant partie de la proposition courante. On considère en effet que quelle que soit l'évolution de l'activité du réseau, ces nœuds doivent conserver une activité élevée dans la mesure où ils représentent ce qui est explicite, donc ce qui est également incontournable. Une solution pour ce faire aurait pu être de fixer la valeur de leur activité une fois pour toutes lors de la phase de propagation en question. Cette solution présente néanmoins l'inconvénient d'interdire toute modulation de l'activité de ces types de concept par le réseau en fonction de leur importance véritable dans la situation évoquée par le texte. Nous avons donc préféré implémenter leur statut de type de concept explicitement mentionné en les munissant d'une connexion en provenance d'un nœud fictif représentant l'occurrence du type de concept dans le texte et possédant toujours le même niveau d'activité. Leur activité est donc le résultat de la synthèse d'un flux externe issu du texte et d'un flux interne provenant des autres nœuds du réseau.

Parmi les particularités de la propagation d'activité, il faut noter qu'à la différence de la première phase de propagation, on ne pose plus de contrainte sur l'activation des nœuds épisode et graphe agrégés. Ces contraintes avaient une justification dans un réseau reposant sur une propagation d'activité à flux constant : elles permettaient en effet de limiter la diffusion de l'activité en exploitant la signification des structures traversées. En revanche, elles n'ont plus de raison d'être dans un réseau récurrent qui ne vise pas à délimiter une zone mais à converger vers un état devant mettre en évidence les connaissances pertinentes pour le traitement d'un passage de texte. Dans le même esprit, la propagation au sein du treillis de types de concept n'est plus sujette à un quelconque amortissement.

Le dernier point que nous aborderons concernant cette seconde phase reprend un mécanisme présent dans REMIND pour éviter que les différences de connectivité des nœuds du réseau de propagation n'interfèrent avec leurs différences d'activité. Par exemple, si un attribut d'une UT agrégée possède beaucoup de graphes, il a a priori plus de chances d'être fortement activé qu'un autre attribut possédant un nombre moins important de graphes. Ce problème se pose plus spécifiquement pour cette seconde phase puisque l'activité de tous les nœuds sélectionnés est prise en compte à chaque cycle.

La solution adoptée pour éviter un tel phénomène consiste à moyenner l'activité parvenant à un nœud pour l'ensemble des connexions d'un même type. Ce pré-traitement est réalisé en pratique par des unités spécialisées associées aux nœuds. Les connexions d'un nœud en provenance d'un type de nœud donné sont ainsi toutes orientées vers une unité de pré-traitement dédiée, elle-même reliée au nœud considéré. Un nœud graphe possède ainsi une unité de pré-traitement pour les connexions venant des nœuds types de concept, une autre pour les nœuds graphe et une dernière pour le nœud attribut auquel il est lié. Cette dernière unité n'est présente que pour l'homogénéité des structures de données manipulées. C'est le cas également de l'unité de pré-traitement associée aux nœuds UT agrégée regroupant les connexions en provenance des nœuds attribut ou à l'inverse de l'unité associée aux nœuds attribut pour gérer les nœuds UT agrégée puisque dans toutes ces situations, le nombre de nœuds en connexion est connu a priori et ne change pas. Il faut préciser que lors de la première phase, ces unités sont en quelque sorte transparentes puisqu'elles se contentent de réaliser la somme pondérée de leurs entrées actives, ce qui est tout à fait en accord avec la propagation à flux constant.

3.4. Validation et discussion

Ainsi que nous l'avons mentionné au cours de la présentation ci-dessus, le mécanisme de rappel associé à la mémoire épisodique a été partiellement mis en œuvre et validé en utilisant l'environnement MALCOM, dédié à l'implantation et au test de réseaux à propagation d'activité en général et plus spécifiquement de réseaux structurés et hétérogènes. Cet environnement a été implanté en Smalltalk, à l'instar des autres composants logiciels développés ici. Il nous a permis en particulier de tester la seconde phase du rappel, qui est en l'occurrence la plus délicate, du fait de l'utilisation d'un réseau récurrent. L'exemple du paragraphe 3.3.3 donne une illustration du type de tests qui ont été pratiqués pour mettre au point cette seconde phase. La faiblesse du nombre d'UTs agrégées disponibles (cf. §5 de ce chapitre) conjuguée à la relative indépendance entre la mémoire épisodique et le réseau de propagation sur lequel s'appuie le rappel nous ont conduit, pour mener ces test, à construire manuellement un réseau de propagation de taille significative (256 nœuds) reproduisant la structure de la mémoire épisodique, à l'exception des différences, relativement minimes, détaillées au §3.3.3.

La seconde phase du mécanisme de rappel peut donc être considérée comme validée dans une certaine mesure. Bien entendu, des tests supplémentaires restent à mener. Le premier d'entre eux consistera à travailler avec un réseau construit automatiquement à partir d'une mémoire suffisamment vaste. Le processus de construction du réseau de propagation associé à un couple mémoire épisodique - mémoire sémantique reste

cependant à implanter, sachant qu'il devra en pratique prendre la forme d'une nouvelle phase ajoutée à la mémorisation d'une représentation de texte. Par ailleurs, il est évident que des tests avec plusieurs mémoires différentes devront être réalisés afin d'avoir une certitude un peu plus établie quant à la portée des résultats obtenus.

Le degré de validation de la première phase est beaucoup plus faible puisque celle-ci n'a pas été implémentée et n'a donc été testée que manuellement. La propagation à flux constant pose cependant moins de problèmes quant à l'appréhension a priori de ses résultats que l'activation d'un réseau récurrent. Seules quelques expérimentations devront être accomplies afin de régler précisément la valeur des paramètres déterminant l'étendue de la zone de sélection délimitée (quantité d'activité initialement injectée, taux de dissipation au niveau des nœuds et seuil de non propagation) en fonction des résultats désirés quant à ce point. Par ailleurs, la conception de notre mécanisme de propagation à flux constant s'est appuyée sur les résultats de l'application assez large de ce type de processus réalisée dans le cadre de la partie acquisition de concepts de MoHA [Gruselle 1997].

Chacune des deux phases du rappel ayant été validée, l'étape suivante sera bien évidemment de tester le rappel dans son ensemble, ce qui n'a pu être fait étant donné l'absence d'implantation de la première phase. Il faut d'ailleurs souligner les difficultés inhérentes à un tel test. Il n'est pas forcément évident en effet de déterminer quelles sont les UTs agrégées d'une mémoire épisodique les plus intéressantes pour traiter un passage de texte et donc, de constituer une référence servant de support aux tests. Une façon minimale de procéder est de vérifier que la présentation des éléments stockés dans la mémoire épisodique provoque le rappel de ces mêmes éléments, ou tout du moins des structures dans lesquels ils sont présents. On peut ainsi soumettre à la mémoire épisodique les textes ayant permis de la construire et vérifier qu'ils induisent le rappel des UTs agrégées contenant les UTs provenant de ces textes.

Une autre façon de faire est d'opérer une évaluation différentielle en reportant le problème de l'évaluation du mécanisme de rappel sur celle du traitement auquel il participe : on juge qu'un procédé de rappel est meilleur qu'un autre parce qu'il conduit à une meilleure performance de la tâche dont il fait partie. Cela suppose d'une part que la tâche en question puisse être évaluée, de préférence plus facilement que le rappel lui-même, et d'autre part que l'on dispose de plusieurs mécanismes de rappel afin de servir de points de comparaison. Une solution moins coûteuse, mais également moins fiable, consiste à s'appuyer sur l'interprétation humaine des UTs sélectionnées et du texte traité. C'est en l'occurrence la solution que nous avons adoptée pour le test de la seconde

phase du rappel. Pour compenser les effets de subjectivité, il est alors nécessaire de recouper le jugement de plusieurs sujets, ce que nous n'avons en revanche pas fait.

En supposant qu'une procédure de test ait été élaborée, elle permettrait, entre autres choses, de préciser quel doit être le degré de complexité de la structure du réseau de propagation. Celui que nous avons décrit ici est le reflet assez exact des structures de la mémoire épisodique, ceci afin de rendre possible la mise en place de mécanismes de contrôle de la propagation d'activité sensibles à la signification de ces structures (importance du prédicat dans les graphes, des UTs principales dans les épisodes, ...). Il n'est cependant pas du tout évident que les résultats obtenus avec un tel réseau soient plus intéressants que ceux que l'on obtiendrait avec un réseau plus simple, dans lequel les nœuds type de concept serait par exemple directement reliés aux nœuds UT agrégée, sans la présence de nœuds graphe et attribut.

Enfin, si l'on aborde le domaine des extensions plus large, il serait intéressant d'étudier comment les deux phases du rappel pourraient être regroupées en une seule. Cela permettrait notamment de rendre le *context focusing* plus souple en autorisant des rétro-actions et de pallier le fait que la propagation à flux constant n'est peut-être pas très bien adaptée aux réseaux structurés hétérogènes.

Dans le sens également de l'élaboration d'un mécanisme unique, mais sur un plan plus théorique, il serait également intéressant d'examiner les apports possibles d'un cadre théorique établi. Celui dont ce type de réseau semble être le plus proche est celui des réseaux bayésiens [Pearl 1988]. Nous avons vu en effet qu'un certain nombre de poids figurant dans le réseau de propagation peuvent être interprétés en termes de probabilités conditionnelles (poids associés aux connexions sous-tendant un lien d'appartenance, par exemple l'appartenance d'un graphe à une UT agrégée). Toutefois, nous avons vu également que d'autres poids ne sont pas interprétables en ces termes (connexions entre les nœuds types de concept ou connexions renvoyant à des relations thématiques) et que la symétrie des connexions n'est pas forcément compatible avec les principes des probabilités conditionnelles. Le rapprochement avec un cadre théorique comme celui des réseaux bayésiens demanderait donc un travail important de clarification et de mise en correspondance.

4. Construction de la mémoire épisodique

La mémoire épisodique se construit de façon progressive par les mémorisations successives des représentations de texte produites par l'analyse thématique. Les

mécanismes qui régissent cette construction sont donc également ceux régissant la mémorisation d'une représentation de texte. Ils sont exposés ci-après.

4.1. Principes de la mémorisation d'une représentation de texte [1]

La mémorisation d'une représentation de texte se déroule en trois étapes. La première étape consiste à sélectionner les éléments de la mémoire épisodique les plus en phase avec la représentation de texte considérée. Par élément, on entend ici les constituants de la mémoire dotés d'une certaine autonomie au sein de celle-ci, c'est-à-dire les épisodes et les UTs agrégés. On fait appel au cours de cette première étape au mécanisme de sélection de connaissances exposé au §3 de ce même chapitre. Cette étape n'est cependant pas clairement identifiable en tant que telle puisqu'elle se fond dans le cadre de l'analyse thématique. Ainsi que nous le verrons au chapitre 8, cette analyse s'appuie en effet sur la mémoire épisodique pour opérer et fait appel pour ce faire au mécanisme de rappel présenté au §3. Le résultat produit est principalement une représentation thématique des textes mais on peut également y ajouter la sélection de l'ensemble des éléments de la mémoire épisodique les plus en accointance avec le contenu des textes traités. Compte tenu du mécanisme de sélection utilisé, ces éléments sont en outre pondérés en fonction de la force de cette accointance.

La deuxième étape de cette mémorisation consiste à évaluer la similarité entre la représentation de texte à mémoriser et les éléments de la mémoire épisodique précédemment sélectionnés. Cette évaluation s'effectue suivant l'ordre décroissant de pertinence de ces éléments, établi lors de la première étape. Un ensemble de conditions permettent de déterminer si la représentation de texte présente une similarité suffisamment forte avec un élément de la mémoire pour y être agrégée. On retient le premier, parmi ceux des éléments de la mémoire issus de la sélection initiale, pour lequel cet ensemble de conditions est rempli.

La troisième et dernière étape est l'opération d'agrégation proprement dite. Elle réalise la mémorisation effective de la représentation de texte. Cette dernière étape ne peut intervenir néanmoins que dans la mesure où une similarité suffisamment forte a été trouvée entre la représentation de texte et l'un des éléments de la mémoire sélectionnés. Dans le cas contraire, la mémorisation se traduit par la construction d'une nouvelle structure agrégée au sein de la mémoire épisodique.

---

[1] Nous donnons ici les principes généraux, sans nous occuper comme nous le ferons après, du fait qu'une représentation de texte, en raison de sa décomposition en UTs, n'est pas mémorisée comme un seul bloc.

Globalement, la mémorisation d'une représentation de texte se traduit donc par l'algorithme suivant :

```
liste_éléments_sélectionnés ←  sélection des éléments de la mémoire épisodique proche de
                               représentation_de_texte
Répéter
    élément_mémoire ← élémentSuivant(liste_éléments_sélectionnés)
    sim ← similarité(représentation_de_texte,élément_mémoire)
Jusqua (sim = vrai) ou estVide(liste_éléments_sélectionnés)
Si (sim = vrai) alors
    agrégation(représentation_de_texte,élément_mémoire)
Sinon
    création d'une nouvelle structure agrégée à partir de représentation_de_texte
Fin_si
```

Il faut préciser que le schéma *évaluation de la similarité puis agrégation ou création d'une nouvelle structure agrégée suivant le résultat de la similarité* est général et s'applique à tous les niveaux, que ce soit pour les épisodes, les UTs, les graphes, les relations inter et intra-UTs ou les rôles.

4.2. Similarité entre représentations de texte et mémoire épisodique

4.2.1. La similarité au niveau des épisodes

Ainsi que nous l'avons vu au §2.1 de ce chapitre, la mémoire épisodique est beaucoup plus centrée sur la notion d'UT que sur celle d'épisode. La mémorisation d'une représentation de texte passe donc en premier lieu par la mémorisation des UTs qu'elle contient. Dès lors, il apparaît normal que la similarité entre représentation de texte et structures de la mémoire se définisse sur la base de la similarité entre UT et UT agrégée. À l'échelon le plus élevé, un épisode est jugé similaire à un épisode agrégé si son UT principale, ou l'une au moins de ses UTs principales s'il en possède plusieurs, est jugée similaire à l'une des UTs principales de l'épisode agrégé. Pour que deux épisodes s'agrègent, ils faut donc qu'ils possèdent au moins un thème principal en commun.

L'agrégation des UTs principales, donc des épisodes, n'impliquent pas une agrégation des UTs secondaires. Lors de l'agrégation d'un épisode avec un épisode agrégé, une UT secondaire de la représentation de texte peut être agrégée à une UT de cet épisode agrégé comme elle peut être agrégée à une UT agrégée n'ayant aucun rapport avec cet épisode agrégé si la similarité avec elle est jugée comme étant la plus forte. Cela signifie en particulier qu'une UT agrégée peut être référencée par plusieurs épisodes agrégés différents.

Dans le même esprit, l'absence de similarité des épisodes ne signifie pas que l'intégralité de la représentation de texte va donner lieu à de nouvelles structures agrégées. Elle n'implique une telle création que pour l'épisode. Du fait de l'autonomie des UTs agrégées au sein de la mémoire, des agrégations peuvent en effet intervenir au niveau des UTs constituant l'épisode sans qu'il y ait agrégation au niveau de celui-ci. Dans le cas le plus favorable, toute une partie d'une représentation de texte, consistant en une configuration d'UTs et de relations thématiques liant ces UTs, trouve un équivalent au niveau de la mémoire épisodique. Il s'agit alors d'un ensemble de situations apparaissant simultanément de façon récurrente avec des liens analogues. Le cas le plus simple et le plus fréquent est celui du triptyque UT – relation de déviation – UT développant un des événements de l'UT source de la déviation. C'est ce que l'on observe dans l'exemple de la figure 6.1 entre l'UT *Tentative_de_meurtre* et l'UT *Hôpital*.

La présence d'une telle similarité suppose que pour chaque relation de la configuration considérée, on trouve une relation agrégée similaire telle que les UTs agrégées qu'elle relie soient elles-mêmes similaires aux UTs reliées par la relation présente dans le texte. De plus, cette relation agrégée doit faire partie de la même grande catégorie de relations thématiques que la relation textuelle. Comme dans le cas des relations intra-UTs, on laisse en effet la possibilité de définir une hiérarchie de relations thématiques. Les types situés au premier niveau en dessous du sommet de cette hiérarchie, incarnée par un type générique tel que Relation_thématique, définissent chacun une grande catégorie de relations. On distingue pour ce qui nous concerne les relations de déviation thématique et les relations de changement de thème. La similarité entre deux relations thématiques ne peut avoir lieu que si leurs types respectifs appartiennent à la même catégorie.

À l'échelon le plus élémentaire de la similarité intervenant au niveau des épisodes, on se contente d'évaluer deux à deux la similarité des UT textuelles avec les UTs agrégées ayant été sélectionnées.

4.2.2. La similarité des Unités Thématiques

Les UTs étant des représentations structurées, leur similarité est déterminée en suivant la voie classique consistant à l'évaluer en fonction de la similarité de leurs constituants. En l'occurrence, leurs constituants de plus haut niveau sont leurs attributs *Circonstances*, *Description* et *États Incidents*. La similarité entre un attribut d'une UT et son homologue d'une UT agrégée est caractérisée par une valeur numérique synthétisant la similarité de leurs constituants respectifs, c'est-à-dire les graphes conceptuels qu'ils contiennent. L'évaluation de la similarité des UTs s'appuie sur une analyse qualitative de ces valeurs de similarité des attributs. Cette analyse consiste dans un premier temps à discrétiser les

valeurs de similarité des attributs en les comparant à trois intervalles de valeurs. Ces valeurs discrétisées sont ensuite exploitées par un petit ensemble de règles permettant d'établir la similarité globale des UTs.

La différence de méthode entre les attributs des UTs et les UTs elles-mêmes dans l'évaluation de la similarité tient à la différence de nature de leurs constituants. Les attributs formant les UTs sont des entités pré-définies dont on connaît précisément le rôle vis-à-vis de l'UT qui les contient et le comportement du point de vue du processus d'accumulation au sein de la mémoire épisodique. On peut de ce fait mettre en œuvre un ensemble de règles tenant compte de ce rôle et de ce comportement de façon fine pour évaluer la similarité au niveau des UTs. On sait par exemple qu'une similarité des attributs *Circonstances* de deux UTs n'a pas la même valeur, du point de vue de la similarité globale de ces deux UTs, que la similarité de leurs attributs *Description*. Il suffit pour s'en convaincre de se reporter à l'analyse faite sur les différents attributs d'une UT agrégée au §2.3.4 de ce chapitre.

Les graphes conceptuels composant les attributs présentent en revanche des caractéristiques strictement inverses. Leur nombre n'est pas fixé à l'avance et aucune contrainte n'encadre ni leur nature, ni leur contenu. En fait, l'objectif de l'apprentissage réalisé par le processus d'accumulation vise justement à déterminer quels sont ces graphes pour une situation donnée. En conséquence, on ne peut pas s'appuyer sur un cadre fixe et pré-déterminé pour établir des règles d'évaluation de la similarité des attributs. C'est pourquoi nous avons adopté une méthode plus robuste telle que le calcul d'un agrégat numérique.

Plus exactement, la similarité de deux attributs peut être déterminée suivant deux niveaux de précision en fonction des besoins. Au premier d'entre eux, le moins élevé, le calcul de la similarité repose sur un simple ratio entre le nombre de graphes similaires et le nombre total de graphes. En pratique, on calcule deux taux de similarité : l'un est le nombre de graphes similaires par rapport au nombre total de graphes contenus dans l'attribut de l'UT agrégée; l'autre est le nombre de graphes similaires par rapport au nombre total de graphes regroupés par l'attribut de l'UT textuelle.

Le calcul de deux taux, l'un prenant comme référence ce qui a déjà été mémorisé et l'autre, ce qui est nouvellement apporté, permet de limiter les différences inhérentes à l'hétérogénéité des représentations de texte, notamment quant au niveau de détail avec lequel les situations sont décrites. Une UT agrégée assez complète peut ne pas partager beaucoup de graphes, de son point de vue, avec une UT relative à la même situation et n'évoquant celle-ci que de manière superficielle. En revanche, du point de vue de l'UT textuelle, ce nombre apparemment faible de graphes communs peut représenter l'essentiel de son contenu. Il semblerait raisonnable dans un tel cas de pencher en faveur de la

similarité des deux UTs, ce qui n'est possible que si l'on n'axe pas l'évaluation uniquement sur l'UT agrégée. La situation inverse peut également se produire : une UT textuelle détaillée face à une UT agrégée assez générale. Cette prise en compte des deux points de vue contribue dans une certaine mesure à s'affranchir de l'ordre dans lequel les représentations de texte sont présentées à la mémoire. Le fait de présenter d'abord des descriptions détaillées d'une situation ou au contraire des descriptions assez générales revêt donc une moins grande importance.

Pour obtenir une évaluation qualitative de la similarité des attributs, les deux taux de similarité obtenus sont comparés à deux seuils fixés a priori, t1 et t2, avec t1 < t2. Si les deux taux sont inférieurs à t1, les attributs sont jugés *non similaires* : il ne partagent pas suffisamment de points communs par rapport à l'ensemble de leurs valeurs. Si l'un des deux taux est supérieur à t2, on a au contraire une *similarité forte* entre les deux attributs : la plupart des graphes, soit de l'attribut de l'UT textuelle, soit de l'attribut de l'UT agrégée, voire éventuellement des deux, trouvent à s'apparier.

Si aucune des conditions précédentes n'est remplie, on conclue à une *similarité normale*. Celle-ci fait alors l'objet d'une évaluation plus approfondie afin de juger de la qualité de la similarité des graphes similaires, en tenant compte de leur importance vis-à-vis de la situation. Cette évaluation se traduit par le calcul d'une valeur de similarité, détaillé au §4.2.3. Si cette valeur est supérieure à un troisième seuil, t3, on opte en faveur d'une *similarité forte* des deux attributs. Autrement, on reste sur le jugement de *similarité normale*. Deux attributs partageant un nombre moyen de graphes sont donc considérés comme proches si ces graphes sont fortement similaires et significatifs pour la situation considérée.

Le recours à deux niveaux de précision pour le calcul de la similarité entre attributs d'UT est lié au contexte de connaissances incertaines dans lequel nous nous plaçons. Lorsque l'on dispose d'une mémoire structurée sur la base des connaissances décrivant le domaine abordé, le processus de recherche en mémoire des cas pertinents vis-à-vis du traitement d'un texte peut être à la fois rapide et assez sûr quant à l'intérêt de ses résultats. Tel n'est pas le cas de la mémoire épisodique, qui se caractérise plutôt par une indexation de type indexation de surface, telle que nous l'avons évoquée au chapitre 1 (cf. §2.2.1). Dans cette perspective, le processus de sélection de connaissances présenté précédemment est chargé de fournir un ensemble assez large d'UTs agrégées en relation avec le contexte d'analyse courant. Cet ensemble doit ensuite être exploré en utilisant une mesure spécifique de similarité entre UTs destinée à sélectionner l'UT agrégée représentant la même situation que l'UT textuelle considérée. Cette mesure a donc pour vocation à être utilisée à de très nombreuses reprises, ce qui lui impose d'être efficace. Dans le même

temps, le jugement qu'elle rend doit être suffisamment fiable, donc reposer sur une analyse détaillée.

Le recours à une mesure de similarité à double niveau de précision permet de concilier efficacité et profondeur d'analyse. Le premier niveau, celui des deux ratios, permet de dégager les grandes tendances de manière rapide : similarité évidente, absence de similarité évidente ou cas nécessitant un examen plus attentif. La dernière option déclenche l'application du second niveau d'évaluation qui peut être plus précis, donc également plus coûteux, sans pénaliser l'ensemble des performances puisqu'il ne porte que sur un sous-ensemble des cas.

Les différents seuils évoqués ci-dessus constituent autant de paramètres de la mémoire épisodique. Dans les tests que nous avons effectués, nous avons retenu les valeurs suivantes : t1 = 0,5, t2 = 0,8 et t3 = 1,0. Les ratios pour lesquels t1 et t2 servent de référence sont compris dans l'intervalle [0,1]. Pour la similarité fine entre attributs, qui concerne t3, les valeurs ne sont pas supérieurement bornées mais la valeur de 1,0 peut s'interpréter comme une forme de moyenne (cf. §4.2.3). Ainsi que nous le verrons à la fin de ce chapitre, les valeurs qui sont données ici ne sont pas de "bonnes valeurs" dans l'absolu; elles contribuent simplement à modeler la mémoire épisodique d'une certaine manière.

Après que la similarité des attributs *Circonstances*, *Description* et *États Incidents* a été évaluée suivant les modalités exposées ci-dessus, il est possible de déterminer la similarité globale d'une UT agrégée et d'une UT textuelle à partir des règles qui suivent. Les deux UTs sont jugées similaires seulement si l'une au moins de ces règles s'applique :

R1 : similarité *forte* de l'attribut *Circonstances* et similarité *normale* de l'attribut *Description*;
R2 : similarité des attributs *Circonstances* et *États Incidents*, avec au moins une similarité *forte* pour l'un des deux;
R3 : similarité *normale* de l'attribut *Description* et similarité *normale* de l'attribut *États Incidents*;
R4 : similarité *forte* de l'attribut *Description*.

Les contraintes fixées par ces règles rejoignent assez naturellement les résultats de l'analyse faite au §2.3.4 à propos de la constitution des UTs. On constate ainsi qu'un rôle central est accordé à l'attribut *Description* : trois règles sur les quatre le font intervenir, dont une, R4, le fait intervenir seul, ce qui n'est le cas d'aucun des deux autres attributs. L'attribut *États Incidents* se voit quant à lui attribuer un rôle légèrement plus important que l'attribut *Circonstances* : dans des conditions similaires (R1 et R3), on impose une

similarité forte pour *Circonstances* et seulement une similarité normale pour *États Incidents*. Cette différence provient de la moindre fiabilité intrinsèque du premier par rapport au second. L'attribut *Circonstances* comporte en effet plus de bruit que l'attribut *États Incidents* et son lien avec les événements constituant le corps de la situation est souvent plus ténu. En conséquence, les risques de reconnaissance d'une similarité erronée sont plus forts pour cet attribut, ce qui justifie cette contrainte plus sévère, destinée à éviter une erreur de reconnaissance au niveau de la situation entière.

4.2.3. La similarité des attributs et des graphes

Principes de la similarité fine des attributs

Lorsque l'évaluation de la similarité des attributs fondée sur le seul ratio du nombre de graphes similaires sur le nombre total de graphes ne donne pas de résultat suffisamment net, il est nécessaire d'examiner plus finement les graphes similaires. Cet examen porte à la fois sur la qualité de la similarité de ces graphes et sur leur importance vis-à-vis de la situation. Il s'effectue en deux temps. On calcule d'abord une valeur de similarité pour chaque couple de graphes similaires (graphe de l'UT agrégée – graphe de l'UT textuelle). Ces valeurs sont ensuite agrégées en tenant compte pour chaque couple du poids relatif de son graphe agrégé par rapport à l'UT agrégée dont il fait partie.

Ce mode de calcul de la similarité est analogue à celui adopté dans des systèmes tels que [Kolodner & Simpson 1989] pour résoudre le problème de la similarité entre structures complexes. Il consiste à juger de la similarité des différents traits composant ces structures, pour laquelle une réponse immédiate peut être apportée, et de combiner ces similarités élémentaires par une opération de type moyenne. Afin de rendre compte des différences d'influence relatives existant entre ces traits, on associe à ces derniers un poids fixe caractérisant leur importance. La similarité de deux structures est donc donnée par la moyenne pondérée des valeurs de similarité de leurs traits.

L'originalité de la similarité des attributs d'UT par rapport à ce schéma réside dans la nature des poids associés aux traits. Généralement, ceux-ci sont fixés a priori et une fois pour toutes par le modélisateur. Ils encodent une forme spécifique de connaissances sur le domaine, complémentaire de celle exprimée par exemple au travers des graphes conceptuels. Dans notre cas au contraire, ces poids sont définis dynamiquement. Ils sont calculés à chaque sollicitation en fonction de l'état courant de la mémoire. Du fait de la mémorisation continue de nouvelles représentations de texte, cet état se modifie en permanence et les poids caractérisant en l'occurrence l'importance des graphes agrégés au sein des situations évoluent eux-mêmes en conséquence. Le principe d'accumulation sur

lequel la mémoire épisodique est fondée repose bien entendu sur l'hypothèse qu'une telle évolution est convergente.

Similarité des graphes

La similarité des graphes est évaluée, comme celle des attributs, en deux temps. Nous avons vu au §2.3.2 de ce chapitre que le prédicat d'un graphe est doté d'un statut particulier dans la mesure où il ne peut être généralisé. De cette contrainte, on tire immédiatement une règle de similarité entre graphes : un graphe agrégé et un graphe venant d'un texte ne peuvent être similaires que si les concepts qui sont désignés comme leur prédicat ont exactement le même type. La confrontation de deux attributs impose de réaliser un ensemble de comparaisons égal au produit cartésien des deux ensembles de graphes que forment les deux attributs. Un tel ensemble représente un nombre potentiellement très élevé d'opérations. La règle ci-dessus permet d'apporter une réponse directe à la plupart de ces comparaisons et donc de se concentrer sur les similarités véritablement pertinentes.

Elle explique par ailleurs que les graphes (c) et (e) de la figure 6.3.b n'aient pas été agrégés bien que leurs prédicats respectifs, **Attaquer** et **Frapper**, entretiennent une relation hiérarchique proche. En fait, au sein de la mémoire épisodique, on ne cherche pas à privilégier un niveau de description donné. Ce travail est laissé au processus de généralisation qui décidera, en fonction des relations explicatives trouvées et du niveau de spécificité des autres événements de l'UT, du niveau de description à choisir pour chaque schéma formé.

La contrainte d'égalité portant sur le type des prédicats est complétée par la nécessité d'avoir au moins un de leurs objets similaire : il doit exister dans les deux graphes au moins une relation casuelle équivalente reliant le prédicat à un concept similaire. On rappelle que la similarité entre concepts et entre relations repose dans ses grandes lignes sur l'existence d'un sur-type commun minimal à leurs types qui soit inférieur au type de l'entité jouant le même rôle dans le graphe canonique associé au type du prédicat. On se reportera au §2.3.2 pour la définition complète de cette similarité. L'association des deux contraintes précédentes définit le premier temps de l'appréciation de la similarité de deux graphes.

La second niveau d'évaluation de la similarité entre graphes reprend globalement le même principe que son équivalent au niveau des attributs. Il s'agit de calculer une moyenne pondérée des valeurs de similarité des constituants des graphes. En l'occurrence, on ne retient que les concepts dans la mesure où la similarité entre concepts fait intervenir leur rôle dans le graphe, donc par voie indirecte, ses relations aussi. Le

poids associé à chaque valeur de similarité correspond au poids relatif du concept concerné dans le graphe agrégé. La fonction de calcul de la similarité entre le graphe g et le graphe agrégé g´ contenant un prédicat de même type et au moins un concept similaire s'écrit donc comme :

$$SimGraphe(g, g') = \frac{\sum_{i=1}^{n} wc_i \ \ SimConcept(c_i, c'_i)}{\sum_{i=1}^{n} wc_i} \qquad [1]$$

avec

c_i, les concepts du graphe g en dehors du prédicat;

c'_i, les concepts agrégés du graphe g´ en dehors du prédicat;

SimConcept(c_i,c'_i), la similarité entre le concept c_i et le concept agrégé c'_i. Elle vaut 1 si les concepts sont similaires, 0 autrement. ci et c´i doivent jouer le même rôle dans leurs rôles respectifs;

n, le nombre de concepts dans le graphe agrégé g´;

wc_i, le poids relatif du concept agrégé c'_i dans le graphe g´.

Une autre façon d'exprimer le fait que deux graphes sont similaires s'il partagent au moins un objet similaire consiste à dire que deux graphes sont similaires si *SimGraphe*(g,g´) > 0.

Pour calculer cette mesure de similarité, il faut être capable de mettre en correspondance les deux graphes impliqués. Parmi les opérations existantes sur les graphes conceptuels, celle qui met en œuvre une telle capacité est la projection. Cependant, cette dernière n'est applicable que dans des conditions bien particulières : il faut que le graphe projeté soit plus général[1] que le graphe cible de la projection. Comme le montrent les graphes de la figure 6.9, que l'on considère ici comme similaires, ces conditions ne peuvent pas être respectées dans le cadre d'utilisation de la mesure de similarité entre graphes.

Nous avons donc étendu l'opération de projection afin de lui faire perdre son caractère dissymétrique. Un concept d'un des deux graphes peut ainsi être plus général que son équivalent dans le second graphe en même temps qu'un autre concept de ce second graphe est plus général que son équivalent dans le premier. Dans l'exemple de la figure 6.9, l'agent du graphe (1) est ainsi plus général que l'agent du graphe (2) alors que le patient du graphe (2) est plus général que le patient du graphe (1). Dans le même esprit, chacun des deux graphes peut posséder des parties n'ayant pas de correspondant

[1] Ses concepts doivent être plus généraux (type de concept supérieur et référent plus générique) que les concepts correspondant dans le graphe cible et il doit en outre représenter à concepts et relations égaux un sous-graphe de ce graphe cible.

dans l'autre graphe. Le graphe (1) fait apparaître une relation de type *Manière* pour qualifier le transport, celle-ci étant absente du graphe (2). De façon symétrique, ce même graphe (2) spécifie un moyen de transport ainsi qu'une fonction pour l'agent qui ne sont pas présents dans le graphe (1).

Cette opération de projection modifiée ne vise pas à construire un nouveau graphe, ce qui est du ressort de l'agrégation, mais simplement à effectuer une mise en correspondance des concepts et des relations équivalents afin d'appliquer la mesure de similarité [1]. La contrainte d'égalité portant sur les types des prédicats respectifs en fait par ailleurs une opération dont le résultat est le plus souvent unique[1], ce qui n'est pas le cas de la projection traditionnelle. On peut parler à cet égard de projection dirigée.

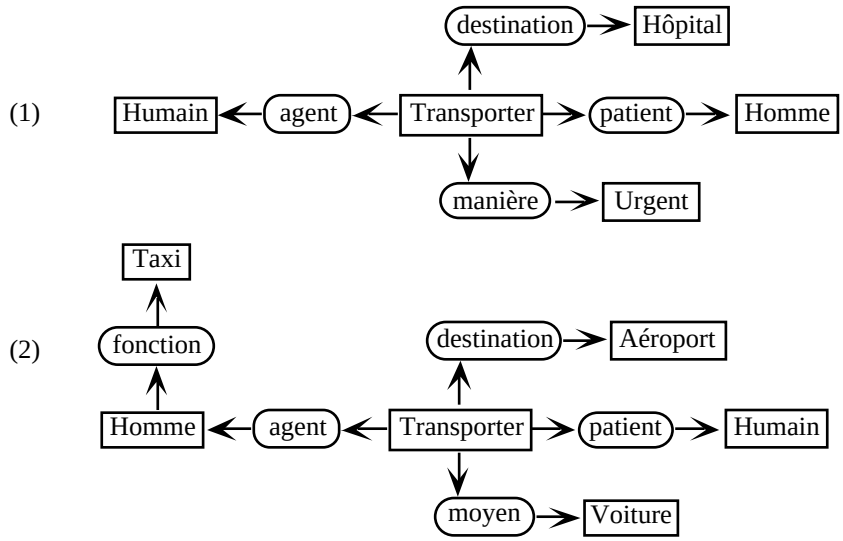


Fig. 6.9 - Deux graphes similaires

En constatant la façon dont est jugée la similarité de deux graphes, il apparaît nécessaire d'ajouter une contrainte supplémentaire sur l'analyse sémantique par rapport à celles déjà explicitées au chapitre 5. Cette contrainte stipule précisément que l'analyse sémantique ne doit pas produire de graphes trop complexes. Il est en effet vraisemblable que plus un graphe regroupera d'informations et plus il sera spécifique. En conséquence, il aura d'autant moins de chances de s'apparier fortement avec un autre graphe présentant les mêmes caractéristiques. Néanmoins, il est raisonnable de penser que cette contrainte est déjà pour une bonne part assumée par le choix que nous avons fait de choisir la proposition en tant qu'unité de représentation des textes. Compte tenu de ce choix, seule la présence de caractéristiques issues d'adjectifs ou d'adverbes peut introduire des informations susceptibles de varier de façon significative d'un graphe similaire à un autre.

[1] Il n'y a pas de garantie théorique que le résultat de cette opération soit unique mais la forme habituelle des représentations sémantiques des propositions conjuguée à cette contrainte sur l'égalité du type des prédicats font de l'unicité du résultat le cas le plus largement général.

Ces informations peuvent dégrader la similarité fine des graphes mais elle ne remet pas en cause en l'occurrence leur appariement.

En ce qui concerne les concepts, nous avons vu au §2.3.2 que lorsqu'il n'existe pas de sur-type commun minimal aux types en présence, la généralisation caractérisant un concept agrégé prend la forme du sur-type regroupant le plus grand nombre d'instances ou bien, dans le cas où un tel sur-type ne peut être dégagé, du type du concept équivalent au niveau du graphe canonique associé au type du prédicat. Du point de vue de la similarité, la première situation se traite exactement comme le cas général : il y a similarité s'il existe un sur-type commun minimal au type du concept agrégé et au type du concept venant de la représentation de texte, ce sur-type étant inférieur au type présent dans le graphe canonique du prédicat.

Dans le second cas au contraire, il ne peut jamais y avoir similarité car le type virtuel du concept agrégé est justement le type présent dans le graphe canonique. Cela n'empêche pas néanmoins l'émergence possible d'un type plus spécifique dans le cas où une agrégation serait tout de même réalisée avec ce concept du fait d'une agrégation plus générale avec le graphe qui le contient. Une similarité avec ce concept agrégé redeviendrait alors envisageable.

Pour être complet à propos de la similarité des concepts, il faut préciser que son évaluation ne prend pas en compte le type de référent, individuel ou ensembliste, des concepts. On considère en effet qu'un référent ensembliste est seulement plus général qu'un référent individuel et que du point du vue de la formation des connaissances sur les situations, cette distinction n'est pas suffisante pour bloquer une éventuelle agrégation.

Détails de la similarité fine des attributs

Après avoir défini la similarité entre graphes, nous pouvons maintenant définir de façon complète le second niveau de similarité entre un attribut *attr* d'une UT d'un texte et un attribut de même intitulé *attr´* d'une UT agrégée de la mémoire grâce à la formule suivante :

$$SimAttribut(attr, attr') = \frac{\sum_{i=1}^{s} wg_i \;\; SimGraphe(g_i, g'_i)}{\overline{wg} \;\; s}$$

avec

g_i, un graphe similaire venant de l'UT du texte; donc $SimGraphe(g_i, g'_i) > 0$;

g_i', un graphe similaire venant de l'UT agrégée;

wg_i, le poids du graphe agrégé g_i' de l'attribut attr´;

s, le nombre de graphes similaires;

et $\overline{wg} = \frac{\sum_{i=1}^{n} wg_i}{n}$, n étant le nombre de graphes contenus dans attr´. $\overline{wg}$ est donc la moyenne des poids relatifs des graphes appartenant à l'attribut attr´ de l'UT agrégée.

Contrairement à la mesure de similarité définie pour les graphes, celle-ci ne possède pas de borne supérieure. La différence entre les deux mesures se situe au niveau de la référence que l'on adopte pour juger de la représentativité des entités similaires par rapport à l'ensemble de celles composant les structures dont on évalue la similarité. En pratique, cette référence s'incarne dans chacun des termes formant le dénominateur des deux mesures. Dans le cas des graphes, le nombre total de concepts présents dans un graphe agrégé n'est jamais beaucoup plus important que le nombre de concepts présents dans un graphe similaire venant d'une représentation de texte. On peut donc adopter comme référence le total des poids des concepts agrégés sans craindre qu'un ensemble de concepts significatifs, donc ayant un fort poids relatif, voient leur influence "écrasée" par un large ensemble de concepts tout à fait contingents. Ce choix permet en particulier d'obtenir une mesure de similarité bornée à la fois inférieurement et supérieurement.

Dans le cas des attributs, la situation est différente dans la mesure où le rapport du nombre de graphes significatifs sur le nombre de graphes contingents est souvent assez faible ainsi que nous l'avons vu au §2.3.4. Or, dans le cadre qui est le nôtre, le fait essentiel n'est pas que la somme totale des poids des graphes similaires représente une part importante du poids total des graphes agrégés. Il faut surtout que les graphes similaires soient des graphes importants vis-à-vis de la situation. Deux attributs partageant un grand nombre de graphes ayant un très faible poids relatif ne doivent ainsi pas être jugés similaires, même si le poids total de ces graphes similaires est en proportion élevé par rapport à la somme totale des poids des graphes agrégés.

La référence que nous avons adoptée pour juger de la représentativité des graphes similaires est donc le produit de la moyenne des poids des graphes agrégés de l'attribut considéré par le nombre de graphes similaires. Ce point de comparaison représente le cas où tous les graphes similaires auraient un poids égal à la moyenne des poids des graphes de l'attribut. Compte tenu de la référence retenue, nous considérons que la similarité entre deux attributs est véritablement marquante lorsque *SimAttribut*(attr,attr´) est supérieur à 1, ce qui justifie la valeur attribuée à t3 au §4.2.2.

Le mode de calcul de la similarité entre attributs exposé ci-dessus s'applique aussi bien aux attributs de type *Circonstances*, *Description* qu'*États Incidents*. En cela, il ne tient pas compte de particularités propres à un type d'attribut, comme le fait que les graphes de *Description* entretiennent souvent des relations de précédence temporelle. Dans ce cas

précis, nous ne voulons pas privilégier un ordre des événements par rapport à un autre car l'attitude inverse conduirait à un calcul beaucoup trop restrictif. Si dans une UT, l'événement E1 intervient avant l'événement E2 et que dans une autre UT jugée similaire sur la seule nature des événements, on observe l'ordre inverse, on pourra sans doute en conclure que l'ordre d'apparition de ces deux événements n'est pas significatif. Cette constatation sera faite au niveau de l'UT agrégée rassemblant les deux UTs précédentes et sera prise en compte lors de la phase d'abstraction de l'UT agrégée visant à créer un nouveau schéma. En revanche, si l'ordre est pris en compte dans la similarité, on risque de conclure à une non-similarité, donc de ne pas réaliser l'agrégation entre les deux UTs et en final, de ne jamais faire la constatation de cette indifférence par rapport à l'ordre.

Plus généralement, et pour les mêmes raisons, nous ne tenons pas compte, lors de l'évaluation de la similarité de deux UTs, des relations existant entre les graphes composant ces UTs, que ces relations prennent place entre des graphes d'un même attribut ou entre des graphes appartenant à des attributs différents. Cette mise à l'écart volontaire ne signifie pas que la similarité entre les relations intra-UTs n'est pas évaluée. Elle joue en effet un rôle lors de l'agrégation en permettant de déterminer si deux relations intra-UTs peuvent ou non être agrégées. La façon dont cette similarité est jugée est équivalente à celle présentée au §4.2.1 pour les relations de suivi thématique.

Si l'on ne tient pas compte de la structuration interne des UTs pour calculer leur similarité, en dehors de celle, systématique, représentée par les attributs, il est intéressant en revanche de prendre en considération la structuration des épisodes. En fonction de la finesse de l'analyse effectuée sur les textes, des textes eux-mêmes et de la façon dont ils décrivent les situations, des graphes analogues peuvent ne pas se retrouver dans des UTs équivalentes. La figure 6.10 offre un exemple d'un tel phénomène. Dans l'UT du texte, g1 et g2 figurent au niveau des *Circonstances* tandis que dans l'UT1 de la mémoire, similaire à l'UT du texte, seul l'équivalent agrégé de g1 se trouve dans les *Circonstances.* En revanche, l'équivalent agrégé de g2 est présent dans les *États Incidents* d'une UT plus spécialisée venant détailler l'événement évoqué par g1. Dans un tel cas, il nous semble justifié de considérer qu'une similarité existe également concernant le graphe g2, en dépit de la différence de structure des épisodes.

Pour tenir compte de ce type de particularités, le processus d'évaluation de deux attributs effectue une recherche des graphes similaires en deux temps. Il commence, comme nous l'avons vu précédemment, par inventorier les graphes similaires entre les deux attributs en question. Il poursuit ensuite sa quête en explorant les attributs pertinents d'éventuelles UTs pouvant détailler aussi bien l'UT du texte que l'UT de la mémoire.

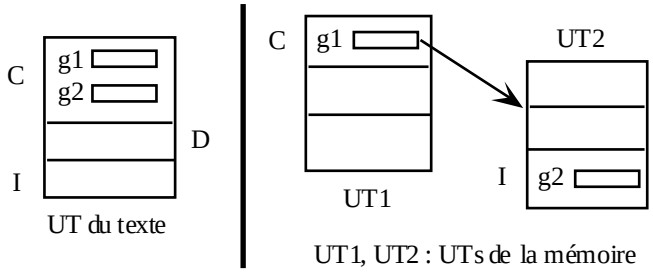


Fig. 6.10 - Comparaison d'UTs ayant une répartition différente des mêmes graphes

Cette exploration s'effectue selon les règles suivantes :

(1) graphes complémentaires pour l'attribut *Circonstances* :
   (a) déviation à partir d'un graphe des *Circonstances* : graphes des attributs *Circonstances* et *États Incidents* de l'UT déviation;
   (b) déviation à partir d'un graphe de *Description* : graphes de l'attribut *Circonstances* de l'UT déviation;
   (c) déviation à partir d'un graphe de *États Incidents* : graphes de l'attribut *Circonstances* de l'UT déviation.

(2) graphes complémentaires pour l'attribut *Description* :
   (a) déviation à partir d'un graphe des *Circonstances* : graphes de l'attribut *Description* de l'UT déviation;
   (b) déviation à partir d'un graphe de *Description* : graphes de l'attribut *Description* de l'UT déviation;
   (c) déviation à partir d'un graphe de *États Incidents* : graphes de l'attribut *Description* de l'UT déviation.

(3) graphes complémentaires pour l'attribut *États Incidents* :
   (a) déviation à partir d'un graphe des *Circonstances* : graphes de l'attribut *Circonstances* de l'UT déviation;
   (b) déviation à partir d'un graphe de *Description* : graphes de l'attribut *Circonstances* de l'UT déviation;
   (c) déviation à partir d'un graphe de *États Incidents* : graphes des attributs *Circonstances* et *États Incidents* de l'UT déviation.

Le cas illustré par la figure 6.10 correspond à la règle 1.a. La recherche s'effectue aussi bien au niveau des UTs venant préciser l'UT du texte que l'UT de la mémoire. Toutefois, on ne permet pas la comparaison de deux graphes tels que l'un appartient à une UT déviation de l'UT du texte et l'autre à une UT déviation de l'UT de la mémoire. Il faut

toujours que l'un des deux graphes comparés appartienne à l'une des UTs originellement comparées.

Lorsque des graphes trouvés selon les règles ci-dessus sont des graphes agrégés, il faut leur attribuer un poids si l'on veut les faire intervenir dans la mesure de similarité fine des attributs. Celui-ci ne peut être le poids qu'ils possèdent au sein de leur UT d'appartenance puisque des UTs sans lien avec l'UT qui est source de la déviation peuvent avoir servi à sa construction. Pour déterminer le poids absolu de ces graphes, nous avons choisi de nous servir, comme pour le poids des relations, des informations conservées au sein de la mémoire sur les épisodes dans lesquels les différents éléments des UTs ont été présents.

De fait, le poids absolu d'un graphe agrégé d'une UT déviation, lorsque ce graphe est ramené au niveau de l'UT source de la déviation, est donné par le nombre d'occurrences de la relation de déviation auquel on soustrait le nombre d'épisodes, parmi ceux dans lesquels la relation est présente, où le graphe visé n'était pas présent dans l'UT cible de la déviation. Le poids relatif d'un tel graphe suit le principe habituel : il est égal à son poids absolu divisé par le poids absolu de l'UT source de la déviation, puisque c'est la similarité d'un attribut de cette dernière que l'on évalue.

4.2.4. Un exemple

Pour illustrer l'intégralité du processus de calcul de similarité entre une UT d'un texte et une UT agrégée, nous avons choisi d'évaluer la similarité entre l'UT *TentativeAssassinat* de la figure 5.5 et l'UT agrégée *Tentative de Meurtre* de la figure 6.3. La figure 6.11 rappelle les grandes lignes du contenu de l'UT *Tentative-*

Circonstances	
[*Être_fou*] (source) → [Femme]	
Description	
(*) [*Poignarder*] (agent) → [Femme]	**[*Transporter*]** (destination) → [Hôpital]
États Incidents	
[*Être_dans*] (lieu) → [Hôpital]	**(*) [*Être_blessé*]** (source) → [Homme]

Fig. 6.11 - Un résumé de l'UT *TentativeAssassinat* [1]

Assassinat et précise ceux de ses graphes qui sont similaires à des graphes de l'UT *Tentative de Meurtre*.

[1] Pour chaque graphe de l'UT, on ne fait figurer que le prédicat et le concept le plus significatif. Ceux qui sont similaires à des graphes de l'UT agrégée *Tentative de meurtre* sont précédés de (*).

L'évaluation de la similarité entre ces deux UTs s'effectue selon le processus décrit précédemment en commençant par l'évaluation de la similarité de leurs attributs puis en essayant d'appliquer au résultat de cette évaluation les règles de similarité des UTs.

Similarité des attributs *Circonstances*

pas de graphes similaires

- similarité 1er niveau :

$Taux_{txt}$: nombre de graphes similaires / nombre de graphes de l'attribut de l'UT du texte

$Taux_{mem}$: nombre de graphes similaires / nombre de graphes de l'attribut de l'UT agrégée

$Taux_{txt} = 0 / 1 = 0$; $Taux_{mem} = 0 / 8 = 0$

$Taux_{txt}$ et $Taux_{mem} < t1$ pas de similarité

Similarité des attributs *Description*

1 graphe similaire : graphe ayant comme prédicat *Poignarder*

- similarité 1er niveau :

$Taux_{txt} = 1 / 2 = 0,5$; $Taux_{mem} = 1 / 10 = 0,1$

$Taux_{txt} \geq t1$ similarité normale

- similarité 2nd niveau :

Pour chaque paire de concepts (**concept du graphe agrégé**, concept du graphe du texte) occupant le même rôle dans leurs graphes respectifs, on donne le sur-type commun minimal de leurs deux types, on situe ce sur-type par rapport au type du concept équivalent dans le graphe canonique associé au prédicat et on en déduit la similarité entre les concepts.

(agent) : sur-typeComMin(**Homme**,Femme) = Humain égal à Humain
SimConcept(**Homme**,Femme) = 0

(destinataire) : sur-typeComMin(**Homme**,Homme) = Homme < Humain
SimConcept(**Homme**,Homme) = 1

(objet) : sur-typeComMin(**(Partie_du_corps)**,Poitrine) = Partie_du_corps égal à Partie_du_corps
SimConcept(**(Partie_du_corps)**,Poitrine) = 0

(partie_de) : sur-typeComMin(**Homme**,Homme) = Homme < Humain
SimConcept(**Homme**,Homme) = 1

(instrument) : sur-typeComMin(**Arme_blanche**,Coupe_papier) = Arme_blanche < Objet_pointu
SimConcept(**Arme_blanche**,Coupe_papier) = 1

(manière) : sur-typeComMin(**Sauvage**,Brutal) = Violent < Qualificatif_Action
SimConcept(**Sauvage**,Brutal) = 1

SimGraphe(**Poignarder**,Poignarder) =
(0,8 * SimConcept(**Homme**,Femme)
+ 1,0 * SimConcept(**Homme**,Homme)
+ 0,4 * SimConcept(**(Partie_du_corps)**,Poitrine)
+ 0,4 * SimConcept(**Homme**,Homme)
+ 1,0 * SimConcept(**Arme_blanche**,Coupe_papier)
+ 0,2 * SimConcept(**Sauvage**,Brutal)) / 3,8 =
(0,8 * 0 + 1,0 * 1 + 0,4 * 0 + 0,4 * 1 + 1,0 * 1 + 0,2 * 1) / 3,8 = 2,6 / 3,8 = 0,68

SimAttribut(**Description**,Description) = (1,0 * 0,68) / (1 * 0,34) ≈ 2,0 > 1,0, 0,34 étant la moyenne des poids des graphes de l'attribut *Description* de l'UT agrégée similarité forte

Similarité des attributs *États Incidents*

1 graphe similaire : graphe comme prédicat *Être_blessé*

- similarité 1er niveau :

$Taux_{txt}$ = 1 / 2 = 0,5; $Taux_{mem}$ = 1 / 4 = 0,25

$Taux_{txt} \geq t1$ similarité normale

- similarité 2nd niveau :

(source) : sur-typeComMin(**Humain**,Homme) = Humain < Être_vivant
SimConcept(**Humain**,Homme) = 1

SimGraphe(**Poignarder**,Poignarder) =
(1,0 * SimConcept(**Humain**,Homme) / 1,0 = 1,0

SimAttribut(**ÉtatsIncidents**,ÉtatsIncidents) = (0,4 * 1,0) / (1 * 0,4) = 1 ≤ 1,0, 0,4 étant la moyenne des poids des graphes de l'attribut *ÉtatsIncidents* de l'UT agrégée similarité normale

Similarité des deux UTs

Compte tenu des résultats concernant la similarité des attributs, il est possible d'appliquer la règle R4, de même d'ailleurs que la règle R3, puisque la condition sur la similarité *normale* requise pour l'attribut *Description* peut être considérée comme remplie si la similarité est *forte.* Les deux UTs mise en présence ici sont donc similaires et peuvent être agrégées.

4.2.5. La similarité des rôles

Que ce soit au niveau des épisodes ou des UTs, les rôles représentent, du point de vue de la similarité et de l'agrégation, un cas un peu particulier vis-à-vis des autres éléments

des représentations de texte et de la mémoire puisqu'ils n'interviennent pas dans le jugement de similarité portant sur les UTs ou les épisodes qui les abritent. Ainsi que nous l'avons déjà mentionné à propos des relations intra-UTs, notre objectif n'est pas en effet de définir des conditions de similarité très restrictives, ce qui donnerait peu de chances à deux UTs ou deux épisodes d'être similaires. Or, poser des conditions de similarité sur les rôles peut être qualifié de restrictif puisque l'on exigerait alors non seulement une similarité des éléments deux à deux mais que l'on imposerait au delà la similarité de configurations d'éléments. Compte tenu de ce principe, l'évaluation de la similarité des rôles n'est dirigée que par le seul l'objectif consistant à fournir des précurseurs pour les rôles des schémas. Elle détermine en particulier quand il est nécessaire de créer de nouveaux rôles agrégés.

La détermination de la similarité des rôles ne prend place qu'à la suite de la phase d'agrégation des autres constituants des représentations de texte. Les rôles sont chargés de conserver un lien d'identité entre différentes entités : entre concepts pour les rôles d'UT, entre rôles d'UT pour les rôles d'épisode. Il faut de ce fait attendre de savoir à quelles entités de la mémoire se sont agrégées les constituants d'une représentation de texte pour avoir une référence commune entre la mémoire et cette représentation. On peut alors juger si deux entités faisant référence au même objet se sont agrégées à deux entités de la mémoire faisant elles-mêmes référence à un même objet. Ainsi, pour déterminer si un rôle d'UT pointant vers deux concepts d'un épisode peut s'apparier à un rôle d'UT agrégé, il faut connaître les concepts auxquels se sont agrégés les deux concepts en question, de manière à savoir si ce sont les mêmes que ceux référencés par le rôle d'UT agrégé.

De façon générale, un rôle se compose d'un ensemble d'unités de même type, composées chacune d'une référence à une entité ainsi que d'une référence à la structure dont elle fait partie. Une unité de rôle d'UT est ainsi constituée d'un concept et du graphe dans lequel ce concept apparaît tandis qu'une unité de rôle d'épisode comporte un rôle d'UT ainsi que l'UT abritant ce rôle. La détermination de la similarité d'un rôle passe par celle de la similarité de ses unités. Une unité d'un rôle d'une représentation de texte (*entitéT,structureT*) est jugée similaire à une unité d'un rôle équivalent de la mémoire (*entitéM,structureM*) si *entitéT* s'agrège à *entitéM* et si *structureT* s'agrège à *structureM*.

Pour déterminer la similarité globale de deux rôles, on reprend les principes adoptés pour les attributs. On retrouve en effet la même caractéristique d'un nombre faible d'unités de rôle significatifs un peu noyés au milieu d'un nombre important d'unités de faible poids. La similarité est donc donnée par le rapport entre la somme des poids relatifs des unités similaires et le produit de la moyenne des poids relatifs des unités du rôle

agrégé par le nombre d'unités similaires. Pour que la similarité puisse être décidée, il faut que ce rapport soit supérieur ou égal à 1,0. Comme pour les attributs, une condition supplémentaire est posée concernant le nombre d'unités similaires, en plus de celle portant sur leur poids. Il faut en effet que les unités similaires représentent au moins la moitié des unités du rôle agrégé ou du rôle venant de la représentation de texte.

4.3. Mémorisation d'une représentation de texte : l'opération d'agrégation

4.3.1. Principes généraux de l'agrégation

Quel que soit le niveau considéré, l'agrégation suit dans ses grandes lignes les mêmes principes :

lorsqu'un élément E_{txt} d'une représentation de texte appartenant à une structure S_{txt} est trouvé similaire à un élément E_m de la mémoire appartenant à une structure S_m, le premier est fusionné avec le second si S_{txt} et S_m sont elles-mêmes agrégées[1]. Le poids absolu de E_m augmente alors d'une unité tandis que son poids relatif reste stable. Du fait de l'augmentation du poids absolu de la structure S_m englobant cet élément agrégé les autres éléments de cette structure qui n'ont pas d'équivalent au sein de la représentation de texte voient leur poids relatif diminuer par un effet naturel de son mode de calcul. Les éléments de la mémoire s'agrégeant comme Em avec des éléments de la représentation de texte se trouvent donc renforcés par rapport aux autres du fait du simple maintien de leur poids. C'est ainsi que l'on fait émerger les traits récurrents des situations par le traitement d'un ensemble important de textes.

Lorsqu'un élément d'une représentation de texte n'a pas d'équivalent au niveau de la mémoire, il est simplement ajouté à la structure de la mémoire à laquelle sa structure englobante est agrégée. Son poids absolu est alors de 1. Le processus est identique si cette structure englobante est elle-même ajoutée à la mémoire et non agrégée à l'un des éléments déjà existants. L'ajout s'effectue alors dans la nouvelle structure créée en mémoire.

Plus formellement, on peut décrire l'agrégation de la façon suivante. Soient

$A = \{a_1, \ldots, a_n\}$, une structure de la mémoire composée des éléments a1 à an;

$T = \{t_1, \ldots, t_m\}$, une structure d'une représentation de texte composée des éléments t_1 à t_m;

$A' = \{a'_1, \ldots, a'_p\}$, la structure A à la suite de l'agrégation de T à A.

[1] En dehors du cas des UTs, deux éléments ne peuvent être agrégés que si les structures auxquelles ils appartiennent s'agrègent elles aussi. Cela s'applique en particulier aux graphes et à leurs constituants.

On définit alors l'agrégation de T à A par l'algorithme donné ci-après. Cet algorithme se veut très général. En particulier, il ne tient pas compte du fait que dans la pratique, les couples d'éléments similaires sont déjà formés à la suite de la phase d'évaluation de la similarité.

```
Pour i = 1 a m faire
    indSim ← faux
    j ← 1
    Tantque non indSim et (j ≤ n) faire
        Si similarité(ti,aj) alors
            a´j ← fusionner(ti,aj)
            indSim ← vrai
        Sinon
            j ← j + 1
        Fin_si
    Fin_tantque
    Si non indSim alors
        ajouter(ti,A´)
    Fin_si
Fin_pour
```

La fonction *similarité(t_i,a_j)* fait référence à une similarité déjà évaluée auparavant. La fonction *fusionner(t_i,a_j)* renvoie quant à elle à l'opération d'agrégation lorsqu'il s'agit d'une structure non élémentaire et à une fusion en tant que telle lorsqu'il s'agit d'éléments terminaux comme les concepts ou les relations. Enfin, la fonction *ajouter(t_i,A´)* réalise la création d'un nouvel élément agrégé dans A´ à partir de t_i.

4.3.2. L'agrégation des épisodes

Du fait de l'autonomie des UTs dans la mémoire, l'agrégation des épisodes se démarque du processus général d'agrégation présenté au §4.3.1 sur deux points. Tout d'abord, l'absence de similarité entre une représentation de texte et au moins un épisode agrégé de la mémoire n'entraîne pas nécessairement la création d'un nouvel élément agrégé pour chacun de ses constituants, en l'occurrence les UTs. Une UT textuelle peut en effet s'agréger avec une UT de la mémoire sans que les épisodes dont elles font partie soient similaires. Le second point est en quelque sorte une extension du premier. Il stipule en effet que lors de l'agrégation d'un épisode textuel avec un épisode agrégé, une UT du premier ne trouvant pas à s'apparier dans le second ne donne pas forcément lieu à la création d'une nouvelle UT agrégée au sein de celui-ci mais peut tout à fait s'agréger avec une UT de la mémoire située en dehors de cet épisode si les deux UTs sont similaires.

La mémorisation d'une représentation de texte suit donc le processus suivant. On commence par procéder à l'agrégation de toutes les UTs qu'elle contient, indépendamment de leur relation avec un éventuel épisode agrégé similaire à la

représentation de texte. On agrège ainsi chacune de ses UTs avec l'UT de la mémoire qui lui est le plus similaire. Si aucune n'a satisfait les critères de similarité, l'UT textuelle est ajoutée à la mémoire en tant que nouvelle UT agrégée.

La seconde étape consiste à agréger les relations de suivi thématique du nouvel épisode avec celles déjà présentes en mémoire lorsqu'elles sont similaires. Autrement, elles donnent lieu à la création de nouvelles relations thématiques agrégées. Dans le cas où ces relations sont plus finement différenciées que la simple distinction faite ici entre déviation thématique et changement de thème, le type de la relation agrégée peut être abstrait si le sur-type commun minimal des types des deux relations, la relation agrégée et la relation textuelle, est supérieur au type actuel de la relation agrégée.

Les deux étapes précédentes se déroulent systématiquement, que la nouvelle représentation de texte soit similaire ou non à un épisode agrégé de la mémoire. Si une telle similarité a été trouvée, elles sont complétées par l'agrégation au niveau des épisodes. Celle-ci prend corps au travers de trois opérations : l'ajout d'une référence vers les UTs agrégées nouvellement créées à partir de celles de la représentation de texte, la mise à jour du poids absolu par rapport à l'épisode considéré des UTs agrégées ayant intégré une nouvelle UT et enfin, l'évaluation de la similarité des rôles d'épisode entre épisode textuel et épisode agrégé, suivie de leur agrégation (cf. 4.3.5). Si aucun épisode agrégé similaire n'a été trouvé, on crée un nouvel épisode agrégé qui référence toutes les UTs agrégées intégrant une UT de l'épisode textuel. Les rôles de ce nouvel épisode sont alors issus de la transformation directe des rôles de l'épisode textuel en rôles d'épisode agrégé.

4.3.3. L'agrégation des Unités Thématiques

L'agrégation des UTs présente comme celle des épisodes une spécificité par rapport au processus général décrit au §4.3.1. Les constituants les plus directs des UTs sont leurs attributs. Bien qu'une mesure de similarité existe les concernant, la détection d'une absence de similarité entre deux attributs n'entraîne cependant pas leur mémorisation en tant que deux entités distinctes. Un attribut d'une UT textuelle est ainsi systématiquement fusionné avec son homologue de l'UT agrégée à laquelle cette UT textuelle est agrégée, que les deux attributs soient ou non similaires. On impose uniquement à ces deux attributs d'être de même type : on ne peut pas agréger un attribut *Circonstances* et un attribut *Description* par exemple, contrainte déjà prise en compte lors de l'évaluation de la similarité des UTs.

Cette fusion systématique des attributs de même type provient de leur nature. Ils structurent les composants véritables des UTs que sont les graphes représentant les

propositions mais ne sont pas eux-mêmes de réels constituants. Au sein d'un même attribut, les graphes suivent en revanche le processus d'agrégation général. Les graphes issus du texte similaires à des graphes de l'UT de la mémoire sont agrégés à ces graphes agrégés tandis que les graphes textuels sans équivalent au niveau de l'UT agrégée y sont ajoutés en tant que nouveaux graphes agrégés. C'est ainsi qu'une UT de la mémoire peut à la fois s'enrichir et renforcer ses traits caractéristiques de façon progressive.

Plusieurs spécificités sont à noter à propos de l'agrégation des graphes. La première que nous évoquerons est le cas dans lequel l'attribut X d'une UT A comporte plusieurs graphes dotés de prédicats de même type, i.e. ayant le même type de concept, appelé ici P. En pratique, ces graphes sont rarement plus de deux. Lorsqu'une UT B, similaire à l'UT A, contient dans son propre attribut X un, voire plusieurs graphes ayant un prédicat de type P, se pose le problème de déterminer quel graphe apparier avec quel autre graphe. Pour opérer un choix, on calcule la similarité fine pour tous les couples formés par le produit cartésien des graphes concernés et l'on effectue l'appariement en fonction des plus fortes valeurs obtenues pour cette similarité.

Toujours à propos de l'appariement des graphes possédant des prédicats de même type, il faut citer le cas des graphes ayant été jugés non similaires dans le cadre d'UTs qui, au contraire, sont similaires. Ces graphes ont des prédicats équivalents mais aucun de leurs objets n'est similaire à un objet équivalent de l'autre graphe. La solution la plus évidente consisterait à suivre la logique générale et à créer pour chacun d'entre eux un nouveau graphe agrégé. Néanmoins, du point de vue de la réalité de la situation, la présence de deux événements de même type n'est pas équivalente à la présence d'un seul. Même si la vocation de la mémoire épisodique est davantage de faire émerger de grandes tendances plutôt que de fines distinctions, il nous semble en l'occurrence préférable, notamment pour la simplicité des opérations de généralisation, de ne faire apparaître dans un attribut deux graphes ayant des prédicats de même type que si ces deux graphes étaient présents de manière similaire dans une des représentations de texte ayant servi à construire l'UT concernée.

Ce principe a pour conséquence que deux graphes non similaires ayant des prédicats de même type doivent tout de même être agrégés s'ils appartiennent à deux attributs identiques de deux UTs que l'on agrège. Cette agrégation s'effectue alors sans renforcement, c'est-à-dire sans modification du poids absolu du graphe agrégé, ni du poids absolu des concepts et des relations qu'il contient.

Le dernier point touchant à l'agrégation des graphes vue du niveau des UTs a trait aux graphes similaires n'appartenant pas à des UTs similaires. Nous avons vu que le calcul de la similarité fine des attributs est capable de s'affranchir de certaines différences de

structure entre les UTs, en allant chercher éventuellement des graphes similaires dans des UTs déviation. La différence gommée par la mesure de similarité est en revanche conservée par l'agrégation. On se repose en effet sur le processus d'accumulation de la mémoire pour déterminer quel est le niveau adéquat de description à retenir lors de la généralisation de l'UT source de la déviation. Si les graphes recherchés dans une UT déviation sont en fait présents la plupart du temps dans cette UT source, ils apparaîtront directement dans le schéma qui résultera de sa généralisation. Dans le cas contraire, ils feront partie d'une généralisation de l'UT déviation.

Le déroulement de l'agrégation de deux UTs suit globalement le même schéma que celui des épisodes. Après que les constituants ont été agrégés, en l'occurrence les graphes, on peut procéder à l'agrégation des relations intervenant entre ces constituants. Ce sont ici les relations intra-UTs, rendant compte des liens de causalité et d'ordonnancement temporel. Le principe de leur agrégation est exactement le même que celui de l'agrégation des relations de suivi thématique, auquel nous renvoyons le lecteur (cf. §4.3.2). Le fait que certaines ne prennent place qu'entre graphes d'un même attribut ou entre graphes d'attributs différents n'a aucune influence. L'étape suivante de l'agrégation des UTs est l'évaluation de la similarité des rôles d'UT et leur agrégation. Le processus se termine par la mise à jour du poids absolu de l'UT agrégée qui est la cible de l'agrégation. On trouvera au §4.3.6 le résultat de l'agrégation de l'UT *TentativeAssassinat* et de l'UT agrégée *Tentative de meurtre* dont on a évalué la similarité au §4.2.4.

4.3.4. L'agrégation des graphes

La figure 6.12 donne le résultat de l'agrégation des graphes de la figure 6.9[1]. On y relève les deux traits essentiels de cette opération. Tout d'abord, l'agrégation de deux concepts entraîne la généralisation de leur type. Cette agrégation n'intervient que lorsque des concepts jouent le même rôle dans les deux graphes et qu'ils sont similaires. Dans l'exemple donné, une telle généralisation entraîne un changement de type, par rapport au type présent dans l'un au moins des deux graphes, pour l'*agent* du prédicat (le concept *Transporter*) ainsi que pour les concepts occupant le rôle du *patient* et celui de la *destination*. Dans les deux premiers cas, *Homme* et *Humain* sont généralisés en *Humain* tandis que dans le dernier cas, *Aéroport* et *Hôpital* sont généralisés en *Lieu_Public*.

À chaque fois, on retient le sur-type commun minimal des deux types s'il est inférieur au type du concept équivalent dans le graphe canonique associé au type du prédicat des

[1] En pratique, on réalise l'agrégation d'un graphe textuel à un graphe agrégé. Néanmoins, comme il s'agit globalement d'un appariement non orienté, au contraire par exemple de la projection, il n'est pas nécessaire de préciser si le graphe agrégé est le graphe (1) ou le graphe (2).

graphes. On se reportera au §2.3.2 sur la description des graphes agrégés pour les détails de la généralisation lorsqu'un tel sur-type n'existe pas. Il faut préciser en outre que les mêmes principes s'appliquent pour les relations casuelles entre concepts dans la mesure où il existe une hiérarchie de types de relations. Autrement, on se contente d'une fusion des relations équivalentes.

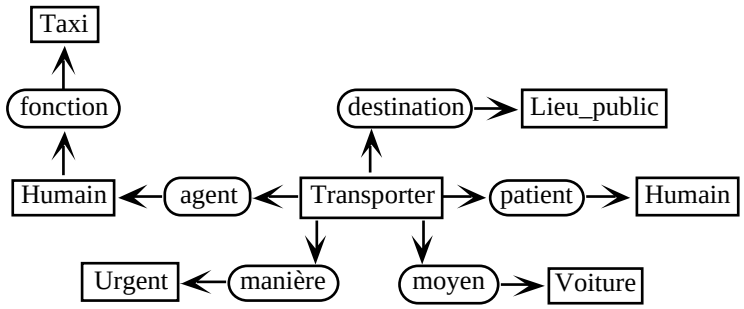


Fig. 6.12 - Graphe résultat de l'agrégation des graphes de la figure 6.9

Le second trait de l'agrégation des graphes consiste comme pour les épisodes et les UTs à ajouter les éléments nouveaux en tant que nouvelles entités agrégées. Ces éléments sont ici des morceaux de graphes, formés au moins d'une relation et d'un concept lui étant lié. Le graphe agrégé de la figure 6.12 est ainsi formé d'un sous-graphe commun aux deux graphes de la figure 6.9 :

[Transporter]
{ → (agent) → [Humain],
→ (patient) → [Humain],
→ (destination) → [Lieu_public]
}

et de trois morceaux de graphes. Les morceaux → (fonction) → [Taxi] et → (moyen) → [Voiture] viennent du graphe (2) tandis que le morceau → (manière) → [Urgent] est issu du graphe (1). Les concepts et les relations de ces morceaux deviennent de nouveaux concepts et relations agrégés, ne rassemblant qu'une seule occurrence.

Plus formellement, l'opération d'agrégation, si on la conjugue à l'évaluation de la similarité des graphes, s'apparente à la jointure maximale que nous avons définie au chapitre 4. Deux différences avec cette dernière opération sont à noter. La plus importante réside dans le fait que les types des concepts et des relations sont généralisés le plus possible dans le cadre de l'agrégation alors qu'ils sont au contraire spécialisés lors d'une jointure maximale. Pour chaque couple de concepts ou de relations équivalents, cette dernière retient en effet le type le plus spécifique.

La seconde différence vient de ce que la conjugaison de l'évaluation de la similarité des graphes et de leur agrégation produit moins de solutions que la jointure maximale. Dans la

très grande majorité des cas, la solution est même unique. Au contraire de la jointure maximale, l'opération conjuguant similarité et agrégation est en effet dirigée par un concept spécifique, le prédicat des deux graphes, ce qui limite les possibilités d'appariement. En dehors des spécificités de ces deux opérations, il faut ajouter que la forme des graphes constituant la représentation sémantique des propositions contribue également à l'unicité des solutions. Ces graphes ne comportent en particulier pas de cycles et ne développent pas de branches profondes. De plus, ils ont tous la même allure générale : les concepts sont arrangés en étoile autour du prédicat et possèdent eux-mêmes quelquefois une ou deux ramifications formées d'un seul concept.

Comme pour l'agrégation des UTs et des épisodes, l'agrégation de deux graphes se clôt par la mise à jour de son poids absolu ainsi que par celle du poids absolu de ses constituants, c'est-à-dire ses concepts et ses relations casuelles. Leur poids relatif étant calculé à partir des poids absolus, cette mise à jour touche donc également les poids relatifs. Nous avons vu au §4.3.3 que seule l'agrégation de deux graphes non similaires ayant un prédicat de même type n'entraîne pas cet accroissement des poids.

4.3.5. L'agrégation des rôles

Que ce soit pour les rôles d'épisode ou les rôles d'UT, l'agrégation se déroule exactement selon la procédure décrite par l'algorithme du §4.3.1. Les constituants sont dans ce cas formés par les unités de rôle. Le détail de l'application aux rôles de cet algorithme est plus spécifiquement présenté au §2.3.3, auquel le lecteur pourra se reporter. Le seul point à ajouter est l'étape finale de mise à jour des poids absolus des rôles et des unités qui les composent.

4.3.6. Un exemple

Les figures 6.13.a et 6.13.b nous montrent le résultat de l'agrégation de l'UT *TentativeAssassinat* de la figure 6.11 à l'UT agrégée *Tentative de meurtre* du §2.3.1. Globalement, cette agrégation vient confirmer certaines tendances déjà émergentes et apportent des éléments nouveaux dont le devenir sera fixé par des agrégations futures. Parmi les tendances confirmées, il faut citer le renforcement du graphe *Poignarder* dans

(a) *Être_localisé* (0.16)
(objet) (1.0) **Événement (1.0)**
(objet) [1] Événement [1]
(lieu) (1.0) **Aéroport (1.0)**
(lieu) [1] Aéroport [1]

(c) *Habiter* (0.33)
(agent) (1.0) **(Humain)**
(agent) [3,5] Homme_politique (0.5) [3], Femme (0.5) [5]
(lieu) (1.0) **Habitation (1.0)**
(lieu) [3,5] Appartement [3], Maison [5]

(e) *Menacer* (0.16)
(agent) (1.0) **Homme_politique (1.0)**
(agent) [3] [3] Homme_politique
(patient) (1.0) **Homme_politique: {*} (1.0)**
(patient) [3] Homme_politique: {*} [3]

(g) *Commander* (0.16)
(agent) (1.0) **Homme (1.0)**
(agent) [4] Homme [4]
(objet) (1.0) **Armée (1.0)**
(objet) [4] Armée [4]

(b) *SeQuereller* (0.16)
(agent) (1.0) **Jeune_homme (1.0)**
(agent) [2] Jeune_homme [2]
(objet) (1.0) **Argent (1.0)**
(objet) [2] Argent [2]
(co-agent) (1.0) **Jeune_homme (1.0)**
(co-agent) [2] Jeune_homme [2]

(d) *Croire* (0.16)
(agent) (1.0) **Homme_politique (1.0)**
(agent) [3] Homme_politique [3]
(objet) (1.0) **Idée: {*} (1.0)**
(objet) [3] Idée: {*} [3]

(f) *Soutenir* (0.16)
(agent) (1.0) **Homme_politique: {*} (1.0)**
(agent) [3] Homme_politique: {*} [3]
(objet) (1.0) **Idée: {*} (1.0)**
(objet) [3] Idée: {*} [4]

(h) *Dormir* (0.16)
(agent) (1.0) **Femme (1.0)**
(agent) [5] Femme [5]
(temps) (1.0) **Nuit (1.0)**
(temps) [5] Nuit [5]

(i) *Être_fou* (0.16)
(source) (1.0) **Femme (1.0)**
(agent) [6] Femme [6]

États Incidents

(a) *Être_emprisonné* (0.33)
(source) (1.0) **Homme (1.0)**
(source) [1,2] Soldat [1], Jeune_homme [2]

(c) *Être_mort* (0.5)
(source) (1.0) **Homme (1.0)**
(source) [2,3,4] Jeune_homme [2], Homme [4], Homme_politique [3]

(d) *Être_Guillotiné* (0.16)
(source) (1.0) **Femme (1.0)**
(source) [3] Femme [3]

(b) *Être_blessé* (0.5)
(source) (1.0) **Humain(1.0)**
(source) [1,5,6] Chef_d'état [1], Femme [5], Homme [6]
(manière) (0.33) **Léger (0.33)**
(manière) [1] Léger [1]

(e) *Être_dans* (0.16)
(lieu) (1.0) **Hôpital (1.0)**
(lieu) [6] Hôpital [6]
(source) (1.0) **Homme (1.0)**
(source) [6] Homme [6]

Relations intra-UTs agrégées :
D.b → I.a (1.0)
causalité [1, 2]
D.a → I.b (1.0)
causalité [1,5,6]
D.a → I.c (1.0)
causalité [2,3,4]

Fig. 6.13.a - UT *Tentative de meurtre* après agrégation avec l'UT *TentativeAssassinat* de la figure 6.11 (attributs *Circonstances* et *États Incidents*)

Description

(a) *Poignarder* (1.0)
(agent) (1.0) **Homme (0.66)**
(agent) [1,2,3,4,5,6] Soldat [1], Jeune_homme [2], Femme [3,6], Homme [4,5]
(patient) (1.0) **Homme: x (1.0)**
(patient) [1,2,3,4,5,6] Chef_d'état: x [1], Homme [4], Homme_politique [3], Femme [5], Jeune_homme: x [2]
(objet) (0.5) **(Partie_du_corps)**
(objet) [1,2,6] Bras (0.16) [1], Ventre (0.16) [2], Poitrine (0,16)
(partie_de) (0.5) **homme: x (0.5)**
(partie_de) [1,2,3] Chef_d'état: x [1], Jeune_homme: x [2] Homme: x [6]
(instrument) (1.0) **Arme_blanche (1.0)**
(instrument) [1,2,3,4,5,6] Baïonnette [1], Cran_d'arrêt [2], Couteau [3], Épée [4], Couteau_de_chasse [5], Coupe-papier [6]
(manière) (0.33) **Violent (0.33)**
(manière) [5] Sauvage [5], Brutal [6]

(c) *Attaquer* (0.16)
(agent) (1.0) **Soldat (1.0)**
(agent) [1] Soldat [1]
(patient) (1.0) **Chef_d'état (1.0)**
(patient) [1] Chef_d'état [1]
(manière) (1.0) **Soudain (1.0)**
(manière) [1] Soudain [1]

(g) *SeBaigner* (0.16)
(agent) (1.0) **Homme_politique (1.0)**
(agent) [3] [3] Homme_politique
(lieu) (1.0) **Baignoire (1.0)**
(lieu) [3] Baignoire [3]

(i) *Attacher* (0.16)
(agent) (1.0) **Homme (1.0)**
(agent) [5] Homme [5]
(patient) (1.0) **Femme (1.0)**
(patient) [5] Femme [5]

(b) *Arrêter* (0.5)
(agent) (1.0) **Policier (0.66)**
(agent) [1,2,3] Policier [1,3], Humain [2]
(patient) (1.0) **Homme (0.66)**
(patient) [1,2,3] Soldat [1], Jeune_homme [2] Femme [3]

(d) *Trébucher* (0.16)
(agent) (1.0) **Soldat (1.0)**
(agent) [1] Soldat [1]

(e) *Frapper* (0.16)
(agent) (1.0) **Jeune_homme (1.0)**
(agent) [2] Jeune_homme [2]
(patient) (1.0) **Jeune_homme (1.0)**
(patient) [2] Jeune_homme [2]

(f) *Pénétrer* (0.33)
(agent) (1.0) **Humain (1.0)**
(agent) [3,5] Femme [3], Homme [5]
(destination) (1.0) **Habitation (1.0)**
(destination) [3,5] Appartement [3], Maison [5]
(manière) (1.0) **(Illégal)**
(manière) [3,5] Clandestin (0.5) [3], Par_effraction (0.5) [5]

(h) *Perdre* (0.16)
(agent) (1.0) **Homme (1.0)**
(agent) [4] Homme [4]
(objet) (1.0) **Bataille (1.0)**
(objet) [4] Bataille [4]

(j) *Déchirer* (0.16)
(agent) (1.0) **Homme (1.0)**
(agent) [5] Homme [5]
(objet) (1.0) **Chemise_de_nuit (1.0)**
(objet) [5] Chemise_de_nuit [5]

(k) *Transporter* (0.16)
(agent) (1.0) **Humain (1.0)**
(agent) [6] Humain [6]
(patient) (1.0) **Homme (1.0)**
(patient) [6] Homme [6]
(destination) (1.0) **Hôpital (1.0)**
(destination) [6] Hôpital [6]
(localisation) (1.0) **Quartier (1.0)**
(localisation) [6] Quartier [6]
(manière) (1.0) **Urgent (1.0)**
(manière) [6] Urgent [6]

Fig. 6.13.b - UT *Tentative de meurtre* après agrégation avec l'UT *TentativeAssassinat* de la figure 6.11 (attribut *Description*)

l'attribut *Description* et celui de *Être_blessé* dans *États Incidents*. Par effet de contraste, le poids d'une grande partie des graphes de l'UT agrégée diminue. Cette baisse concerne aussi bien des graphes que l'on peut juger comme anecdotiques, tels que *SeBaigner* dans *Description* ou *Dormir* dans *Circonstances*, que des graphes tels que *Arrêter* dans *Description* ou *Être_mort* dans *États Incidents*, qu'un jugement extérieur tendrait à sélectionner du fait de leur cohérence vis-à-vis de la situation représentée.

L'hypothèse portée par le principe d'accumulation affirme que sur un grand nombre d'agrégations, ces graphes significatifs seront présents plus souvent que les autres, même s'ils ne figurent pas dans toutes les représentations de texte. Il ne faut donc pas s'arrêter à une tendance locale mais juger à partir de l'évolution globale des poids. Sous cet angle d'ailleurs, l'hypothèse semble être confirmée puisque les poids des graphes comme *Arrêter* ou *Être_mort* sont supérieurs aux poids des graphes purement anecdotiques.

L'apport d'éléments nouveaux concerne le graphe *Être_fou* dans l'attribut *Circonstances*, *Transporter* dans *Description* et *Être_dans* dans *États Incidents*. Le premier apparaît trop spécifique pour être confirmé par la suite, bien que tout dépende des textes qui seront traités, mais les deux suivants, parce qu'ils font intervenir le type de concept *Hôpital*, font partie des enrichissements candidats à un renforcement ultérieur.

L'agrégation de deux UTs conduit également à renforcer les relations intra-UTs comme ici la relation de causalité entre le graphe *Poignarder* de *Description* et le graphe *Être_blessé* de *États Incidents*. Ce renforcement ne se traduit en tant que tel qu'au niveau de son poids absolu, son poids relatif ayant plutôt pour vocation à rendre compte du degré de nécessité de sa présence.

L'analyse d'une agrégation particulière, impliquant une UT agrégée à un stade de développement déjà avancé comme celle de la figure 6.3, conduit à s'interroger sur l'évolution plus globale des UTs agrégées ainsi que sur l'impact de l'agrégation à cette échelle. Un des angles d'analyse intéressant à considérer à cette occasion est celui de l'ordre dans lequel les UTs sont considérées. Une étude a priori laisse en effet penser que cet ordre est particulièrement important dans les premières étapes de la constitution d'une UT agrégée dans la mesure où les contours de celle-ci se figent souvent assez rapidement et ce, ainsi que nous l'avons vu au §2.3.4, sur la base d'un sous-ensemble relativement restreint de ses constituants.

Une analyse globale des agrégations montre en premier lieu que l'attribut *Circonstances* contribue généralement assez peu à la similarité des UTs. Cela signifie en particulier que les règles R1 et R2 sur la similarité des UTs sont rarement appliquées. L'UT agrégée de la figure 6.3 est d'ailleurs un exemple représentatif de cet état de fait. Seul le graphe *Habiter*, rassemblant deux occurrences, est présent dans plus d'une seule

UT. Par ailleurs, il n'a pas eu d'influence dans la formation initiale de l'UT agrégée puisqu'il n'est apparu que dans la troisième et la cinquième UT parmi celles ayant été agrégées. En l'occurrence, un ordre différent de présentation des UTs textuelles, surtout en considérant les épisodes 3 et 5 à la suite l'un de l'autre et au début de la formation de l'UT agrégée, n'aurait pu que favoriser la similarité.

L'importance de l'attribut *Description* dans l'agrégation est toute autre. Dans la succession des agrégations ayant abouti à la formation de l'UT agrégée de la figure 6.3, la similarité pour cet attribut a toujours été au moins une similarité *normale*. D'autre part, l'ordre de présentation des UTs s'est avéré avoir dans ce cas une influence non négligeable. La présence simultanée des graphes *Poignarder* et *Arrêter* dans les trois premiers épisodes a en effet contribué de façon significative à l'orientation de l'UT agrégée. En supposant par exemple que le premier épisode ait été suivi par les épisodes 4 et 5, les trois n'ayant pas d'autre graphe en commun que *Poignarder*, on n'aurait pas eu de similarité entre l'attribut *Description* de l'UT agrégée formée à partir des épisodes 1 et 4 et l'attribut *Description* de l'UT appartenant à l'épisode 5. Le ratio $Taux_{txt}$ aurait en effet eu la valeur de 0,25 tandis que le ratio $Taux_{mem}$ aurait été égal à 0,2, soit deux valeurs assez nettement inférieures au seuil de similarité *normale*, fixé à 0,5.

Même en supposant qu'une similarité *normale* ait pu être détectée, il aurait été peu probable que le calcul de la similarité fine des attributs conclue à une similarité *forte.* En effet, après un petit nombre d'agrégation seulement, les poids relatifs des graphes ne sont pas encore très différenciés. Leur écart-type reste en conséquence faible. La similarité fine des attributs étant évaluée en prenant comme référence la moyenne des poids relatifs des graphes de l'attribut considéré, il est donc préférable d'avoir dans cette phase un nombre élevé de graphes similaires, surtout si la similarité des graphes en question n'est pas parfaite. Ce point est moins critique lorsque l'UT a acquis une certaine maturité, ainsi que le montre le calcul de la similarité avec l'épisode 6 au §4.2.4.

La succession d'agrégations envisagée ci-dessus entre des UTs des épisodes 1, 4 et 5 se serait en pratique arrêtée à la première d'entre elles dans la mesure où l'attribut *États Incidents* de l'UT venant de l'épisode 1 n'est pas similaire à celui de l'UT appartenant à l'épisode 4. Les deux attributs n'ont en effet pas de graphe commun. Ceci illustre l'importance que l'attribut *États Incidents* revêt lui aussi dans la constitution de la mémoire. Trois sur quatre des règles statuant sur la similarité des UTs nécessitent la similarité de deux attributs. Comme nous l'avons vu ci-dessus, la similarité des attributs *Circonstances* est assez rare. La seule règle ne faisant intervenir qu'un seul attribut impose une similarité *forte* de l'attribut *Description*. Or, l'étude de la similarité des attributs de ce type nous a montré qu'une similarité *forte* est difficile à obtenir dans la

phase initiale de formation d'une UT agrégée. On en conclut donc qu'au moins pendant cette période, la similarité des UTs repose essentiellement sur la règle R3, d'où l'importance de l'attribut *États Incidents*. À mesure qu'une UT agrégée devient plus stable, la règle R4 devient elle aussi applicable et rend par la même occasion la similarité de l'attribut *États Incidents* moins cruciale.

L'analyse des UTs agrégées réalisée au §2.3.4 a mis en évidence que l'attribut *États Incidents* se caractérise par un faible bruit, contrairement aux deux autres attributs. Cette spécificité se répercute de façon naturelle au niveau de la similarité. Dans la succession des agrégations ayant conduit à l'UT agrégée de la figure 6.3, la similarité relative à l'attribut *États Incidents* a ainsi toujours été jugée comme *forte* alors que celle relative à l'attribut *Description* n'a été parallèlement que *normale*.

Cette propriété ne met pas néanmoins cet attribut complètement à l'abri des effets de séquence mentionnés précédemment. C'est ce que l'on peut d'ailleurs observer lorsque l'on essaie d'agréger les UTs des épisodes 1 et 4. La variabilité concernant la façon dont les actions d'une situation sont exprimées dans les textes en est partiellement responsable. Nous avons vu au chapitre 5 que tantôt les actions apparaissent en tant que telles, tantôt sous la forme de leurs conséquences immédiates sur l'état du monde. Les deux formes sont plus rarement présentes simultanément puisqu'elles sont redondantes du point de vue du lecteur, qui possèdent les connaissances lui permettant de passer de l'une à l'autre. Cette variabilité d'expression se traduit par une variabilité parallèle du remplissage des attributs *Description* et *États Incidents*. Si l'on ne veut pas être tributaire de cette instabilité et de l'effet de séquence qui peut en résulter, il est nécessaire de recourir de façon systématique aux inférences reposant sur les connaissances sémantiques liées aux actions. Cela suppose bien entendu que les inférences requises par le texte ne soient pas de nature pragmatique.

La similarité des attributs de type *États Incidents* peut également être contrariée par la variété des conséquences d'une même situation. Si l'on prend, toujours à titre d'exemple, la situation représentée par l'UT agrégée de la figure 6.3, on constate que la victime peut n'être que blessée ou bien au contraire mourir tandis que l'auteur de l'attentat est simplement emprisonné ou bien plus radicalement exécuté. On obtient ainsi quatre configurations possibles. Sur les dix associations possibles de deux configurations (en incluant les associations de configurations identiques), une majorité d'entre elles, en l'occurrence huit, rassemblent deux configurations ayant au moins un élément en commun. On peut donc penser que les chances de détecter une similarité dans un tel cas de figure sont raisonnables et pas dépendantes de façon excessivement sensible de l'ordre dans lequel les épisodes sont considérés. Dans un cadre général, il faut cependant ajouter

aux 20% d'échecs le fait que la similarité des graphes communs n'est pas forcément bonne et enfin, que les attributs des UTs textuelles, et surtout ceux des UTs agrégées, comportent un nombre non négligeable, même au niveau de l'attribut *ÉtatsIncidents*, de graphes purement anecdotiques.

5. Validation et discussion

Dans cette partie, nous ne discuterons que de l'implémentation, la validation et les limites des structures de la mémoire épisodique ainsi que de ses mécanismes de construction. Pour un approfondissement de ces différents points concernant le rappel, on pourra se reporter au §3.4.

5.1. Implémentation

Les structures de la mémoire épisodique et la procédure de mémorisation des représentations de texte ont donné lieu à une implémentation complète. Celle-ci, réalisée en Smalltalk comme pour les autres parties de MLK, exploite particulièrement les possibilités offertes par ce langage à objets, notamment en termes d'héritage et de polymorphisme. Du fait de la proximité entre les constituants de la mémoire épisodique et ceux des représentations de texte, des structures de données communes ont en effet été développées et spécialisées ensuite par héritage pour définir d'un côté, les représentations de texte et de l'autre, les composants de la mémoire épisodique.

Dans le cas des graphes conceptuels agrégés, notre démarche, bien qu'également inspirée par le souci de réutiliser l'implantation des graphes conceptuels existante, a été légèrement différente. Cette implantation avait été réalisée indépendamment de notre travail et par conséquent, une abstraction des structures de données telle celle effectuée pour les autres constituants de la mémoire épisodique ne pouvait être faite.

Plutôt que d'utiliser l'héritage, qui ne donnait pas pleinement satisfaction en raison notamment de l'absence d'héritage multiple en Smalltalk, nous avons opté en faveur d'un mécanisme de redirection transparente : un graphe agrégé est un objet faisant référence à un graphe conceptuel "normal" et redirigeant vers ce graphe tous les envois de message qui le concernent[1]. Ce même objet est doté par ailleurs de tous les attributs d'une entité agrégée (nombre d'occurrences, références vers les épisodes d'origine, ...) nécessaires à la gestion de son appartenance à la mémoire épisodique. Le même mécanisme a été utilisé pour les concepts et les relations agrégées. De ce fait, le graphe conceptuel référencé par

[1] Cette redirection peut être réalisée de façon générique en Smalltalk en redéfinissant la méthode de traitement déclenchée pour traiter l'exception levée lorsqu'un message est envoyé à un objet sans correspondre à une des méthodes de ce dernier.

un graphe conceptuel agrégé pointe lui-même vers des concepts et des relations agrégées, et ce, de façon totalement transparente pour lui puisque ceux-ci se comportent de son point de vue comme des concepts et des relations “normaux”. L’intérêt d’un tel procédé réside en l’occurrence dans la possibilité de réutiliser tous les outils développés dans le cadre de la plate-forme de gestion de graphes conceptuels présentée au chapitre 4.

À côté de l’implémentation proprement dite des structures de la mémoire épisodique et de ses opérations de construction, nous avons conservé le souci, comme pour les autres types de représentation, de fournir des moyens de visualiser et de manipuler les structures de données complexes de cette mémoire épisodique afin de faciliter le plus possible les tests. On pourra trouver à l’annexe F une vue d’ensemble des interfaces développées dans ce but. Elles comprennent en particulier des inspecteurs spécialisés assurant une exploration plus aisée des structures imbriquées ainsi que des interfaces dédiées aux opérations de calcul de similarité et d’agrégation. Ces dernières ont été développées par Gaël de Chalendar dans le cadre de son stage de DEA.

Dans le même esprit de facilitation des tests, l’exécution des opérations de calcul de similarité et d’agrégation s’accompagnent, un peu à la manière des systèmes de maintien de la vérité (TMS), de la construction de structures de données spécifiques offrant la possibilité de reconstituer a posteriori le détail des opérations réalisées.

5.2. Validation et limites

La validation des principes sous-tendant la mémoire épisodique est un problème important dont nous avons déjà implicitement discuté au chapitre 5. Cette validation est en effet presque exclusivement tributaire des représentations de texte. Le problème central de la construction de ces représentations est de notre point de vue celui de l’analyse sémantique. Celui-ci devance même le problème de l’analyse thématique dans la mesure où le résultat de l’analyse sémantique forme la base de départ des autres processus considérés. L’analyse sémantique n’a pu être abordée durant le temps imparti à ce travail mais il est évident qu’une de ses extensions prioritaires devrait lui être consacrée.

À défaut de pouvoir travailler avec des représentations de texte obtenues automatiquement, nous avons testé le comportement de la mémoire épisodique grâce à des représentations de texte construites manuellement, comme dans l’exemple de la figure 6.13. Notre jeu de test est assez réduit puisqu’il se compose au total de 12 représentations de texte[1]. Ce faible nombre s’explique à la fois par la quantité importante de travail d’ingénierie des connaissances à produire pour chaque représentation et par

[1] Une partie de ce jeu de test a été constituée par Gaël de Chalendar lors de son stage de DEA.

notre volonté, compte tenu de l'absence de processus importants tels que l'analyse sémantique, de parvenir au moins à une spécification détaillée de toutes les dimensions d'ANTHAPSI, au détriment sans doute du degré d'approfondissement de chacune d'entre elles. La moitié de ces représentations de texte provient de courts articles journalistiques (à l'exception du texte de la figure 5.2 du chapitre 5, qui est un extrait de livre) et l'autre moitié se compose de textes que nous avons écrits afin de compléter le jeu de test sur un sous-domaine particulier (cf. UTs de l'annexe D).

Le domaine général abordé par la plupart de ces textes est celui du terrorisme et des tentatives d'assassinat. Outre son caractère assez classique dans les travaux portant sur l'apprentissage de connaissances pragmatiques, il présente l'avantage, comme plus généralement tout ce qui relève des faits divers, d'être illustré par un grand nombre de textes susceptibles de respecter les contraintes que nous avons posées. Ces textes, le plus souvent des dépêches d'agence de presse, sont en effet assez courts et rédigés avec un style narratif suffisamment marqué, ce que nous recherchons en priorité pour mener nos expérimentations.

L'analyse manuelle des 12 textes de notre jeu de test a donné lieu à la construction de 23 Unités Thématiques. Après mémorisation, on obtient en moyenne 13 UTs agrégées, ce chiffre variant un peu en fonction de l'ordre de présentation des représentations de texte. Nous avons vu au §4.3.6 que cet effet de séquence est inhérent à un mécanisme d'apprentissage incrémental tel que celui mis en œuvre ici. L'absence d'un jeu de test un peu étendu ne nous a cependant pas permis d'étudier son impact véritable sur la forme de la mémoire épisodique. Compte tenu des principes régissant cette dernière, cet impact ne doit en effet être jugé que sur le long terme. Le problème est similaire pour ce qui est de la détermination de la valeur des paramètres de la mesure de similarité (cf. seuils définissant le type de similarité entre les attributs des UTs agrégées). Le rôle de ces paramètres est central puisqu'en définissant le degré de sévérité de la mesure de similarité, ils constituent le seul moyen d'action sur la forme de la mémoire épisodique n'impliquant pas un changement des spécifications présentées ci-dessus. Toutefois, la valeur de ces paramètres est pour le moment davantage fondée sur l'intuition que sur le résultat d'expérimentations extensives.

Une validation plus complète de la mémoire épisodique passe donc par le développement d'un jeu de test beaucoup plus étendu. Pour que celui-ci soit représentatif, il faut au préalable que la systématisation des procédures de construction des représentations de texte soit plus poussée qu'elle ne l'est actuellement. Puisque nous n'avons pas opté en faveur d'une normalisation ramenant toutes les représentations de texte sur une même base (cf. chapitre 5), cet effort doit s'accompagner d'une mise en œuvre, par la mesure jugeant de la similarité de ces représentations, des inférences

sémantiques nécessaires au dépassement de certaines différences d'expression[1]. Une automatisation complète de ces points constituerait la solution idéale mais elle est peu réaliste à large échelle. Une solution plus réaliste est certainement de définir des guides beaucoup plus précis aussi bien pour la construction des représentations de texte que pour le jugement de leur similarité et de les faire implémenter sous la forme de processus automatiques pouvant interagir avec un opérateur lorsqu'ils ne peuvent opérer complètement seuls.

Un moyen de valider au moins les principes sous-tendant la mémoire épisodique, à défaut de valider directement celle-ci, est de les transposer dans un cadre où les pré-requis nécessaires à cette validation sont moins importants. C'est en partie la démarche dans laquelle s'inscrit le développement de ROSA. On pourra en particulier se reporter au chapitre 10 traitant de SEGAPSITH, que l'on peut considérer comme l'image de MLK à un niveau où les briques de base ne sont pas les concepts mais les mots.

5.3. Extensions possibles

Nous avons vu au paragraphe précédent que la forme de la mémoire épisodique est fortement conditionnée par la valeur des seuils associés à la mesure de similarité entre UTs. Suivant que ces seuils sont sévères ou au contraire lâches, on obtient une mémoire composée d'une multitude de petites UTs agrégées spécifiques regroupant chacune un faible nombre d'occurrences ou au contraire une mémoire constituée de quelques grandes UTs fortement agrégées mais assez hétérogènes.

Une première façon d'étendre le modèle de mémoire que nous proposons est de mettre en œuvre un mécanisme d'auto-adaptation du niveau de ces seuils. Pour cela, il est nécessaire de définir à la fois des critères servant de référence et des indicateurs permettant de juger si l'état de la mémoire est conforme à ces critères. Dans ce schéma de fonctionnement, la valeur des seuils est périodiquement reconsidérée et orientée afin de combler l'écart entre l'état constaté de la mémoire et l'état fixé comme objectif au travers des critères mis en avant. Parmi les critères possibles, on peut citer le fait de favoriser l'émergence d'UTs agrégées équilibrées, ce que l'on peut traduire par une distribution homogène du nombre d'agrégations des UTs agrégées, ou bien encore le fait de chercher à obtenir les UTs agrégées les plus homogènes possible, ce que l'on peut évaluer à partir du nombre moyen de graphes par UT agrégée.

L'auto-adaptation exposée ci-dessus s'exerce au niveau global mais sa transposition à un niveau plus local est également envisageable. On peut ainsi imaginer que les

[1] Ces inférences passent principalement par l'exploitation des graphes de définition, chose qui n'est pas du tout réalisée dans l'implantation actuelle de la mesure de similarité.

paramètres de la mesure de similarité n'aient pas nécessairement la même valeur tout au long de l'existence d'une UT agrégée. Lorsque celle-ci est au début de sa formation, son intérêt est de favoriser les agrégations, même un peu larges, de façon à faire naître une tendance significative. Durant cette période, la mesure de similarité devra donc être assez généreuse. Au contraire, lorsque l'UT agrégée commence à s'affirmer et à devenir stable, son intérêt est de rendre la similarité plus sélective afin de ne pas risquer une dérive progressive ou bien l'ajout de bruit inutile. Cette plus grande sélectivité est obtenue par un ajustement en conséquence des paramètres de la mesure de similarité.

La modification de ces paramètres simultanément aux niveaux local et global pose bien évidemment le problème de l'interaction entre ces deux niveaux. La solution la plus évidente à cet égard consiste à conserver au niveau de chaque UT agrégée une version propre de ces paramètres et de moduler ceux-ci en accord avec les changements de tendance préconisés globalement.

Une autre façon de s'occuper de la forme de la mémoire épisodique est de s'intéresser à sa structuration. Dans sa configuration actuelle, cette mémoire se présente essentiellement comme une collection d'UTs agrégées sans qu'aucune relation de proximité ou de hiérarchisation entre ces UTs n'existe. Cette situation n'est pas pleinement satisfaisante dans la mesure où elle risque de conduire à développer de façon indépendante des UTs agrégées représentant une même situation, ou tout du moins des situations très proches. Il suffit pour cela que ponctuellement, la similarité d'une UT agrégée et d'une UT textuelle ne soit pas détectée du fait d'une trop grande différence de forme, phénomène dont on peut penser qu'il est assez courant. Comme il n'existe pas de procédure de restructuration de la mémoire, la situation restera par la suite en l'état.

Une première approche pour remédier à ce problème pourrait consister à modifier les procédures de mesure de similarité et d'agrégation afin qu'elles se rapprochent de ce que l'on trouve dans un système de classification incrémentale tel qu'UNIMEM par exemple. Cette solution n'est cependant pas satisfaisante car elle ne fait que transformer le problème de non agrégation d'une UT en un problème de mauvais classement de cette UT. En fait, cette solution va à l'encontre d'un des principes de la mémoire épisodique stipulant que seule l'accumulation d'un ensemble de représentations d'une situation permet d'établir quels sont les traits significatifs de cette situation. Or, le bon classement d'une UT textuelle nécessite la connaissance de ces traits.

Une bonne solution pourrait donc prendre la forme d'une association entre les deux approches. La mémoire épisodique se présenterait alors sous la forme de deux espaces. Un des deux espaces correspondrait à la mémoire épisodique telle qu'elle existe actuellement tandis que le second contiendrait une hiérarchie d'UTs agrégées. Celles-ci

occuperaient les feuilles de l'arbre représentant la hiérarchie tandis que ses nœuds internes contiendraient les généralisations de ces UTs agrégées. La mémorisation d'une nouvelle UT textuelle s'opérerait alors de la façon suivante : la hiérarchie des UTs agrégées est d'abord parcourue sur la base du contenu de l'UT textuelle afin de déterminer si elle peut s'agréger à l'une des UTs agrégées de la hiérarchie. On retrouve à ce niveau l'avantage des index hiérarchiques. Dans l'affirmative, l'agrégation est réalisée et peut même conduire, comme dans UNIMEM, à l'éclatement de l'UT agrégée initiale en deux nouvelles UTs agrégées, toutes deux raccrochées à leur généralisation commune. C'est ainsi que la hiérarchie se développe[1]. Si le parcours de cette dernière s'avère négatif, l'UT textuelle est traitée dans le second espace mémoire de la même façon que dans la mémoire épisodique actuelle.

Ce second espace permet ainsi de faire émerger de nouvelles UTs agrégées. Lorsque celles-ci sont considérées comme suffisamment stables, on les incorpore à la hiérarchie du premier espace de la même façon qu'une nouvelle UT textuelle. La seule différence est que ces UTs agrégées stables peuvent donner lieu à la création de nouvelles feuilles, même en l'absence d'agrégation avec une UT agrégée de la hiérarchie. En procédant ainsi, il est possible à la fois d'ajouter la représentation de nouvelles situations à la hiérarchie et d'enrichir la représentation de situations déjà existantes (éventuellement en en créant des spécialisations) lorsque la nouvelle UT agrégée stable est jugée similaire à une UT agrégée déjà en place.

La hiérarchisation des UTs agrégées n'est cependant pas la seule dimension d'une structuration de la mémoire épisodique. Les relations de composition sont en effet tout aussi importantes que les relations de généralisation. Une telle relation de composition existe implicitement lorsqu'une UT agrégée est liée à une autre par une relation de déviation thématique. L'UT déviation constitue en effet le développement de l'un des événements de l'UT source de la déviation. Compte tenu du mode de constitution des UTs agrégées, on peut néanmoins s'attendre à ce que la séparation entre une UT source et une UT déviation ne soit pas forcément très tranchée et qu'un recoupement non négligeable existe entre les graphes de ces deux UTs. Lorsque les UTs en question deviennent suffisamment stables, il peut être intéressant, notamment dans la perspective de leur abstraction, de procéder à une restructuration visant à séparer plus clairement le contenu de l'une du contenu de l'autre.

Au delà, lorsque les capacités de l'analyse thématique ne sont pas très bonnes, on peut facilement imaginer que les besoins de restructuration sont encore plus grands. Une UT source et une UT déviation n'ayant pas été différenciées vont par exemple se retrouver fondues en une seule UT agrégée, qui risque d'être alors trop importante et trop

[1] À la différence d'UNIMEM, cette hiérarchie devra en réalité être une forêt car il est peu probable de faire émerger ses niveaux supérieurs de cette façon.

hétérogène pour s'agréger facilement avec d'autres UTs. Lorsque plusieurs de ces UTs de grande taille ont ainsi des sous-situations entières en commun, il peut être intéressant de les restructurer afin de rendre explicites ces sous-situations et donner ainsi davantage de viabilité à ces UTs. Les moyens de détecter l'opportunité de telles restructurations et de les réaliser restent cependant à étudier.

Parmi les extensions les plus directes à réaliser, figure la capacité à détecter de façon automatique le moment où une UT agrégée devient stable, autrement dit le moment où ses traits principaux ne changent plus. Ce besoin a été évoqué à plusieurs reprises dans ce qui précède : elle est nécessaire par exemple pour déclencher la modification des paramètres de similarité ou encore le classement d'une UT agrégée. Cette capacité est également très importante pour la mémoire pragmatique puisqu'elle conditionne le moment où l'abstraction d'une UT agrégée doit intervenir. Nous aborderons d'ailleurs plus en détail la façon de la mettre en œuvre au chapitre suivant. Précisons tout de même que ces différents besoins ne renvoient pas nécessairement aux mêmes mécanismes ou tout du moins, aux mêmes valeurs des paramètres d'un mécanisme commun, ce qui revient à différencier plusieurs degrés de stabilité en fonction de la tâche considérée. Du point de vue des structures de la mémoire épisodique, nous nous contenterons de souligner que cette détection peut nécessiter de s'appuyer sur un historique de l'évolution de l'UT agrégée dont on souhaite détecter la stabilité et que les informations correspondantes devront donc être ajoutées à la représentation des UTs agrégées.

La dernière extension que nous aborderons ici concerne les relations entre la mémoire pragmatique et la mémoire épisodique. Sous cette étiquette, apparaissent en fait tout un ensemble de problèmes s'articulant autour du devenir des UTs agrégées après leur abstraction. Si l'on conçoit les schémas comme un stade final de développement, il peut apparaître logique de faire disparaître les UTs agrégées dont ils sont issus. L'impact d'une telle disparition doit alors être examiné du point de vue du maintien de la cohérence de la mémoire épisodique. On peut également considérer, et ce serait a priori plutôt notre sentiment, que les schémas apportent une valeur ajoutée en permettant de dégager le certain et l'essentiel des UTs agrégées mais qu'ils ne les remplacent pas pour autant.

Dans ce cas, se pose naturellement le problème de déterminer de quelle manière le lien entre un schéma et l'UT agrégée, ou les UTs agrégées, dont il est issu est maintenu. La réponse à cette question est influencée par le devenir d'une UT agrégée ayant été abstraite : doit-elle rester en l'état ou peut-elle au contraire être encore l'objet de nouvelles agrégations? Il peut sembler en effet délicat de risquer qu'une UT agrégée change progressivement d'orientation par rapport à ce qu'elle était lors de son abstraction. Enfin, même en ne supprimant pas par principe les UTs agrégées abstraites, on ne peut échapper au problème d'un nécessaire processus d'oubli au sein de la mémoire épisodique dans la

perspective du très long terme. S'il ne sert pas à éliminer les UTs agrégées abstraites, cet oubli devrait au moins permettre de faire disparaître les UTs agrégées les plus atypiques, autrement dit les UTs dont on peut penser qu'elles n'atteindront jamais un niveau de stabilité et de représentativité suffisant pour être abstraite en schéma.

Récapitulatif

Dans ce chapitre, nous avons commencé par détailler les principes sous-tendant la mémoire épisodique pour examiner ensuite comment ils se concrétisent au travers de ses structures. Les principes fondateurs de ce modèle de mémoire sont au nombre de trois. Le premier d'entre eux stipule que la mémoire épisodique doit mettre en œuvre un mécanisme d'émergence par accumulation. En l'occurrence, on accumule des représentations de texte et l'on fait émerger des représentations de situations prototypiques. Ce processus d'accumulation opère par appariement des représentations de texte et pondération des éléments qui les composent en fonction de leur degré de récurrence. Le deuxième principe spécifie quant à lui qu'en dépit de ce processus d'accumulation, la mémoire épisodique doit conserver la capacité de reconstituer le contenu des représentations de texte individuelles. Enfin, en vertu du troisième et dernier de ces principes, le processus de rappel ne peut s'appuyer sur une structure d'indexation pré-établie mais doit être capable de s'adapter de lui-même aux évolutions constantes de la mémoire épisodique.

Du fait de son mode de constitution par appariement et fusion de représentations de texte – on parle également d'agrégation –, les structures de la mémoire épisodique sont proches des structures des représentations de texte. Au niveau global de la mémoire, on trouve ainsi des épisodes agrégés, produits de l'agrégation de plusieurs épisodes, mais aussi des Unités Thématiques agrégées, construites par l'agrégation de plusieurs UTs indépendamment de l'agrégation de leurs épisodes d'origine. Les principes d'accumulation s'appliquent de la même façon au sein des UTs agrégées puisque celles-ci contiennent des graphes et des rôles agrégés et que les graphes agrégés sont eux-mêmes formés de concepts et de relations agrégés. Ces structures agrégées sont de façon générale fondées sur le rassemblement et la fusion des éléments communs aux différentes structures qui sont agrégées, conjugués à la conservation de leurs différences ainsi que de l'origine de ces différences, ce qui permet en particulier de mettre en œuvre le deuxième principe énoncé plus haut.

Ce chapitre nous a également donné l'occasion de présenter en détail la façon dont les fonctions de mémorisation et de rappel, qui sont les deux opérations de base permettant

d'interagir avec une mémoire, sont instanciées au niveau de la mémoire épisodique. Conformément au troisième principe posé ci-dessus, la fonction de rappel de cette mémoire est adaptée à la fois à sa structure "plate" et à son caractère évolutif. Nous ajouterons qu'elle fonctionne sur un mode associatif et qu'elle est capable de tenir compte du contexte antérieur établi par les phases précédentes de rappel. Ces deux propriétés vont de pair avec deux contraintes : le rappel prend comme point de départ les types de concept présents dans des propositions et ces propositions, compte tenu de leur appartenance à un même texte, ne sont pas traitées indépendamment les unes des autres.

L'ensemble de ces propriétés est en pratique supportée par un mécanisme de propagation d'activité. Celui-ci opère au sein d'un réseau spécifique, construit sur la base des entités de la mémoire épisodique et des relations existant entre ces entités. Ce mécanisme est plus précisément articulé en deux phases. La première, appelée phase de focalisation du contexte, a pour objectif de délimiter la partie du réseau de propagation, donc de la mémoire, au sein de laquelle la sélection des entités formant le résultat du rappel sera réalisée. Cette délimitation permet de ne pas mobiliser l'ensemble de la mémoire épisodique et donc de rendre faisable l'application du processus de sélection même lorsque la mémoire est de très grande taille. Cette première phase est mise en œuvre par un mécanisme de propagation d'activité à flux constant.

La seconde phase du rappel consiste à valuer les éléments de la mémoire épisodique faisant partie de l'espace délimité précédemment en fonction de leur degré d'adéquation vis-à-vis à la fois de la configuration de types de concept soumise comme indice de rappel et du contexte établi par les phases de rappel antérieures. Cette valuation est assurée par une propagation d'activité au sein du réseau récurrent délimité par la première phase. La convergence de l'état du réseau est garantie par une fonction d'activation des nœuds provoquant la stabilisation progressive de leur activité par réduction de la prise en compte de leurs entrées.

La seconde opération de base de la mémoire épisodique que nous avons abordée dans ce chapitre est la mémorisation de nouvelles représentations de texte. Celle-ci s'effectue en trois étapes. La première consiste à rechercher au sein de la mémoire épisodique les épisodes et les UTs agrégés les plus proches de la représentation de texte à mémoriser. Cette première étape n'est ici considérée que du point de vue de son résultat. La façon dont elle est assurée dépend en effet de la façon dont l'analyse de texte produisant les représentations de texte est réalisée. Il est assez évident que la fonction de rappel est fortement sollicitée à cette occasion mais nous ne faisons pas d'hypothèse sur la forme de cette sollicitation.

La seconde étape de la mémorisation a pour rôle d'évaluer la similarité entre la représentation de texte considérée et les entités de la mémoire épisodique supposées en être proches. Le cœur de cette évaluation repose sur la similarité entre les UTs agrégées et les UTs textuelles. Les UTs agrégées sélectionnées sont donc passées en revue suivant l'ordre décroissant de leur pertinence supposée et comparées à celles de la nouvelle représentation de texte. Ce passage en revue est réalisé pour chaque UT textuelle tant qu'une similarité n'a pas été détectée avec l'une des UTs agrégées sélectionnées.

Cette recherche est suivie de la troisième étape, au cours de laquelle la mémorisation proprement dite de la représentation de texte est réalisée. En fonction des résultats de l'étape précédente, chacun de ses composants est agrégé à l'entité de la mémoire avec laquelle il a été trouvé similaire ou bien est mémorisé en tant que nouvelle entité agrégée lorsqu'aucun équivalent n'a pu être trouvé dans la mémoire épisodique. Cette démarche générale est appliquée pour tous les composants des représentations de texte, quel que soit leur niveau.

Il est possible également de trouver une présentation plus ou moins détaillée de ce modèle de mémoire dans [Ferret & Grau 1996], [Ferret & Grau 1997], [Ferret & Grau 1997] et [Ferret & Grau 1998].

Chapitre 7

L'abstraction de schémas

Nous présentons ici le stade ultime du processus d'apprentissage au sein de MLK, c'est-à-dire le passage des entités agrégées de la mémoire épisodique aux schémas de la mémoire pragmatique. Plus précisément, nous n'exposons dans ce chapitre qu'une première ébauche de ce processus : la construction d'un schéma à partir d'une Unité Thématique agrégée. Cette opération se caractérise par trois phases : la détection de la possibilité d'abstraire une UT agrégée en un schéma, la détermination des événements de l'UT agrégée que l'on doit retenir, et éventuellement généraliser, pour former le schéma et enfin, la construction du schéma proprement dit. Au travers de la présentation des méthodes mises en œuvre par ces trois phases, nous ferons apparaître la particularité principale de cette abstraction, en l'occurrence la nécessité, imposée par notre cadre général de travail, d'utiliser une théorie du domaine que l'on ne peut considérer ni comme complète, ni comme globalement cohérente[1].

1. Nature du problème

Le point que nous abordons ici constitue le dernier stade du processus d'apprentissage au sein de MLK tel que nous l'avons présenté au chapitre 3. Il recouvre la création de schémas par abstraction des entités de la mémoire épisodique. Cette phase prolonge et finalise le travail de généralisation déjà opéré par la mémoire épisodique. Nous nous limiterons dans le cas présent à examiner comment une UT agrégée peut être abstraite afin de créer un nouveau schéma. Il s'agit donc d'une étude préliminaire dans laquelle nous laisserons volontairement de côté des problèmes tels que : comment la structure de la mémoire épisodique en termes de relations de suivi thématique et d'organisation en épisodes peut être traduite au niveau de la mémoire pragmatique ou bien, en liaison avec ce problème, comment faire émerger différents niveaux de schémas et les organiser de façon hiérarchique?

L'intérêt de cette opération d'abstraction, opposée à la solution consistant à ne travailler qu'avec la mémoire épisodique, apparaît naturellement en reprenant les caractéristiques des schémas mises en évidence au §2 du chapitre 4. Ceux-ci représentent une forme de connaissance à la fois plus sûre et plus précise que les UTs agrégées de la mémoire épisodique. Elle est également plus générale du fait de son niveau d'abstraction

[1] Nous tenons à remercier tout particulièrement Gaël de Chalendar dont le travail de DEA [Chalendar 1997] sur le problème abordé dans ce chapitre a contribué très largement à nourrir notre réflexion.

plus élevé. Il est ainsi possible d'utiliser les schémas au delà de la simple analyse thématique pour les impliquer dans des tâches touchant d'autres dimensions de la compréhension de texte. Par ailleurs, leur nature et leur organisation s'accompagnent d'un mode d'évolution complémentaire de celui des UTs agrégées, fait de restructurations explicites conduisant à créer de nouveaux schémas, soit par spécialisation, soit par généralisation de schémas existants, ou encore, en utilisant des mécanismes de raisonnement par analogie. [Grau & Sabah 1985] présente un éventail de ces possibilités pour des schémas proches de ceux détaillés au chapitre 4.

L'opération d'abstraction qui met en œuvre ce passage vers les schémas définit sa spécificité par la présence de deux traits importants. Tout d'abord, elle donne lieu à un véritable changement de représentation et non à la simple construction d'un objet de même type, seulement différent du premier par le degré de spécialisation de ses constituants. Ensuite, cette transformation s'effectue sans faire intervenir de connaissances abstraites sur le domaine considéré. En cela, nous reprenons logiquement la même contrainte que celle imposée à la mémoire épisodique. Pour cette seconde phase de l'apprentissage dans MLK, nous nous démarquons donc à nouveau des travaux tels que [Mooney & DeJong 1985] qui s'inscrivent dans le courant de l'Explanation-Based Learning (EBL) [DeJong & Mooney 1986] ou de l'Explanation-Based Generalization (EBG) [Mitchell et alii 1986]. Nous nous situons au contraire dans une approche résolument inductive, de type Similarity-Based Learning (SBL).

Il faut rappeler que ce parti pris n'est pas définitif. Lorsque l'hypothèse d'une absence initiale de connaissances pragmatiques fournies a priori est retenue, comme c'est le cas ici, il faut pouvoir opérer l'abstraction des schémas sans avoir à s'appuyer sur d'autres schémas. À mesure du traitement d'un certain nombre de textes abordant les mêmes situations et du remplissage en conséquence de la mémoire épisodique, des schémas sont ensuite formés. Sans contrevenir à l'hypothèse initiale, ceux-ci peuvent alors être utilisés afin de contribuer à la formation de nouveaux schémas, lorsqu'ils en sont suffisamment proches. Le mécanisme premier d'abstraction des schémas doit donc être capable d'intégrer l'utilisation de ces connaissances nouvellement construites, lesquelles ne sont le plus souvent que partielles vis-à-vis de la situation considérée. C'est néanmoins un point que nous n'aborderons pas ici.

La création de nouveaux schémas repose donc essentiellement sur le contenu de la mémoire épisodique ainsi que sur les connaissances sémantiques, pour lesquelles nous supposons, au contraire des connaissances pragmatiques, l'existence d'une définition a priori. Les connaissances sémantiques forment en effet le socle sur lequel nous nous

appuyons pour apprendre les connaissances pragmatiques. Cette création se décompose en pratique en cinq étapes :

(1) détection de la possibilité d'abstraire une UT agrégée en un schéma;

(2) recherche des regroupements d'événements possibles au sein de l'UT agrégée;

(3) sélection des événements à retenir pour construire le schéma;

(4) généralisation des événements sélectionnés;

(5) construction du nouveau schéma.

La première de ces étapes se situe en amont du processus d'abstraction proprement dit puisqu'elle a pour objectif de déterminer si une UT agrégée est suffisamment stable et représentative pour être abstraite en un schéma. Si tel est le cas, on commence par examiner si certains de ses événements sont suffisamment proches les uns des autres pour donner lieu à une généralisation commune (étape 2). Cette étape vise à relâcher un peu les contraintes assez strictes imposées par la mémoire épisodique sur la similarité des événements lors la mémorisation des UTs des textes. Nous avons vu en effet au chapitre 6 que deux graphes ne peuvent être jugés similaires, et donc être agrégés, que s'il possèdent un prédicat de même type.

Cette contrainte présente à la fois l'avantage de la simplicité et celui du respect de l'information portée par les textes. Le premier point est particulièrement intéressant pour une mesure de similarité que l'on doit appliquer un grand nombre de fois puisqu'elle se traduit par un coût de mise en œuvre assez faible. Le second permet de mettre en évidence des différences inter-individuelles, comme le phénomène de la typicalité par exemple. Il n'est pas indifférent en effet de savoir qu'un événement est désigné préférentiellement par un type de concept plutôt que par son sur-type ou l'un de ses sous-types.

Néanmoins, la création d'un nouveau schéma passe par le gommage des éléments trop spécifiques des UTs agrégées et l'adoption d'un niveau de description suffisamment général. C'est dans ce cadre qu'une généralisation limitée des prédicats des graphes composant les UTs agrégées est réalisée, avec le but de fusionner certains de ces graphes et de les remplacer par un seul événement au niveau du schéma nouvellement formé.

La troisième étape de ce processus d'abstraction est chargée de sélectionner les événements intrinsèquement liés à la situation représentée en les distinguant de ceux dont la présence au sein de l'UT agrégée que l'on considère n'est que contingente. Cette sélection s'effectue bien entendu en tenant compte des regroupements opérés lors de l'étape précédente.

Le travail de la deuxième étape est accompli dans la perspective d'une généralisation des regroupements de graphes qu'elle réalise. Mais cette dernière opération est véritablement menée à bien par la quatrième étape du processus d'abstraction. La généralisation en question concerne plus largement tous les événements de l'UT agrégée ayant été sélectionnés à l'issue de la troisième étape, qu'ils soient formés d'un seul graphe agrégé ou du regroupement de plusieurs. La deuxième et la quatrième étape forment en réalité les deux parties d'une seule et même opération de généralisation des graphes de l'UT agrégée. Leur dissociation provient uniquement de la volonté de limiter son coût global en ne l'appliquant pas à des graphes incarnant des événements que l'on ne conservera pas dans le schéma final.

La cinquième et dernière étape est celle réalisant le changement de représentation à proprement parler puisqu'elle procède à la création d'un schéma à partir du résultat des étapes précédentes, lesquelles restent dans le cadre des UTs agrégées. Compte tenu des limites du travail présenté, cette opération est ici assez simple. Il s'agit essentiellement de transformer les graphes généralisés en références vers des schémas et de créer un en-tête pour le nouveau schéma en sélectionnant un graphe représentatif de l'UT agrégée. Dans un contexte plus large, chacun de ces sous-problèmes, auxquels d'autres pourraient venir s'ajouter, formerait en lui-même l'équivalent d'une étape.

2. Critères d'abstraction

2.1. Principes

Nous avons vu en préambule que le passage de la mémoire épisodique à la mémoire pragmatique est assimilé ici à une succession d'opérations consistant à abstraire une UT agrégée en un schéma. On fait ainsi de ce passage un processus purement local aux UTs agrégées. Les critères présidant à son déclenchement s'en trouvent simplifiés, de même que les conditions d'application de ces critères. On peut en effet déterminer si une UT agrégée est susceptible de donner naissance à un schéma uniquement en l'examinant lorsqu'elle, et elle seule, se modifie, c'est-à-dire lorsqu'elle s'agrège avec une UT provenant d'une nouvelle représentation de texte. Le coût d'application de ces critères s'en trouve également allégé : en n'ayant pas besoin de passer en revue la mémoire épisodique, on économise un coûteux parcours au sein d'un modèle de mémoire caractérisé par une structuration non hiérarchique.

Les critères d'abstraction sont eux-mêmes d'une application assez directe dans la mesure où par hypothèse, ils ne peuvent pas s'appuyer sur une théorie du domaine

existante et sur un cortège d'inférences plus ou moins complexes que celle-ci permettrait de mettre en œuvre. Ils doivent se contenter d'exploiter la forme des UTs agrégées et la façon dont celle-ci évolue au fur et à mesure des agrégations. Les indicateurs formels susceptibles d'intervenir en tant que critères d'abstraction se répartissent en deux grandes catégories suivant qu'ils permettent de répondre à l'une ou à l'autre des deux questions suivantes :

- l'UT agrégée considérée est-elle suffisamment développée pour que l'on puisse juger de sa capacité à être abstraite en schéma?
- le contenu de l'UT agrégée considérée montre-t-il une tendance suffisamment marquée et suffisamment persistante pour que cette UT soit abstraite en schéma ?

Dans la mesure du possible, on cherche d'abord à répondre à la première question et l'on ne s'attaque à la seconde qu'après avoir obtenu une réponse positive à la première. Ce séquencement va d'ailleurs de pair avec une complexification des critères. Dans la première catégorie, on trouve des indicateurs simples tels que le nombre d'agrégations ayant permis de construire l'UT agrégée considérée ou le nombre de graphes agrégés que cette même UT contient. Le second peut en pratique se ramener au premier. À moins de supposer l'agrégation de plusieurs exemplaires d'une même UT possédant peu de graphes, il est en effet fort improbable d'avoir une UT agrégée caractérisée par un nombre d'agrégations important et un petit nombre de graphes. En dépit de la difficulté à fixer ce type de valeur, la réponse à la première question s'obtient en confrontant le nombre d'agrégations de l'UT considérée à un seuil absolu.

La seconde catégorie d'indicateurs est plus complexe dans la mesure où il s'agit pour eux de synthétiser l'information attachée au contenu de l'UT agrégée candidate à l'abstraction. En généralisant les différentes variantes possibles, ces indicateurs se ramènent à une caractérisation de la distribution des poids des graphes de l'UT agrégée. Le but est plus précisément de détecter les graphes agrégés susceptibles d'être retenus pour intervenir dans la construction d'un éventuel nouveau schéma. On peut ensuite juger si ces graphes sont en nombre suffisant, notamment par rapport aux autres graphes de l'UT agrégée, et si leur présence est suffisamment stable.

2.2. Détail des critères d'abstraction

2.2.1. Vue d'ensemble des critères d'abstraction retenus

Nous avons choisi d'axer nos critères d'abstraction des UT agrégées autour de leur évolution au fur et à mesure des agrégations assurant leur formation. Plus précisément,

nous considérons qu'une UT agrégée peut être abstraite en un schéma seulement lorsque ceux de ses graphes rassemblant une part significative du total des poids de l'ensemble de ses graphes se retrouvent d'une agrégation à l'autre, et ce, pour un nombre minimum d'agrégations. Pour simplifier l'expression, nous parlerons dans ce qui suit de "tête" de l'UT à propos des graphes rassemblant une part significative du poids de l'ensemble de ses graphes. Pour paraphraser le principe exprimé ci-dessus, on peut donc dire qu'un nouveau schéma n'est créé à partir d'une UT agrégée que si le contenu de sa tête reste stable durant un nombre fixé d'agrégations.

En appliquant la grille d'analyse des critères d'abstraction présentée au §2.1, on voit que la détermination de la tête d'une UT agrégée et de son évolution renvoie à la seconde catégorie de critères tandis que la première catégorie est représentée par la contrainte d'un nombre minimum d'agrégations pour juger de la stabilité effective de la tête de l'UT.

2.2.2. Détermination de la tête d'une UT agrégée

La tête d'une UT agrégée est constituée des graphes de cette UT dont la somme des poids relatifs est supérieure strictement à 50% de la somme totale des poids relatifs des graphes constituant cette UT. En pratique, on classe les graphes par ordre décroissant de leur poids relatif dans l'UT et l'on retient pour former la tête de celle-ci les graphes jusqu'à ce que la somme de leurs poids dépasse la moitié de la somme de tous les poids relatifs des graphes de l'UT. Cette tête est définie par rapport à l'ensemble des graphes de l'UT, sans tenir compte de leur répartition entre les trois attributs de cette dernière.

La décision d'abstraction sera néanmoins négative si aucun graphe de l'attribut *Description* ne figure au sein de la tête de l'UT agrégée. De même qu'une UT, un schéma ne doit pas en effet avoir un attribut *Description* vide. Même si le contenu de la tête d'une UT agrégée n'est pas identique à ce que sera le contenu du schéma auquel elle donnera naissance, il y a suffisamment de corrélation entre les deux pour choisir de différer la décision d'abstraction lorsqu'un tel cas de figure se présente. Il suffit en effet de laisser une tendance plus nette se dégager au sein de cette UT agrégée en attendant quelques agrégations supplémentaires.

Considérée seule, la règle donnée ci-dessus ne contraint cependant pas suffisamment la sélection des graphes pour avoir une solution unique. Il est en effet fréquent que des graphes aient le même poids relatif, ce qui introduit un non-déterminisme lorsque l'on ne peut pas retenir tous les graphes ayant le même poids relatif. Dans un tel cas de figure, on impose de choisir les graphes de façon à équilibrer le nombre de graphes retenus entre les trois attributs de l'UT agrégée. Si l'on a sélectionné précédemment 3 graphes de l'attribut *Description*, 2 de *Circonstances* et 2 de *États Incidents* (configuration 3 - 2 - 2) et que l'on

doit encore choisir 2 graphes parmi 3 ayant le même poids et appartenant chacun à un attribut différent, on en choisira un de *Circonstances* et un de *États Incidents*.

En revanche, si l'on a la configuration 3 - 2 - 1 au lieu de 3 - 2 - 2, il reste encore une fois une incertitude entre les configurations finales 3 - 2 - 3 et 3 - 3 - 2. Pour lever ce non-déterminisme final, on fixe un ordre arbitraire sur les attributs. Dans l'ordre décroissant de priorité, un graphe est donc d'abord pris dans l'attribut *États Incidents*, ensuite dans *Description* et enfin dans *Circonstances*. Cet ordre tient compte de l'importance de chaque attribut du point de vue de la similarité des UTs ainsi que de l'importance statistique d'un graphe dans chacun des attributs. Ce dernier point justifie la préférence de *États Incidents* par rapport à *Description*, ce dernier comportant souvent plus de bruit que le premier.

2.2.3. Mesure de similarité des têtes d'UT agrégée et décision d'abstraction

À chaque agrégation concernant une UT agrégée, on applique le critère d'abstraction en mesurant la similarité entre la tête T´ de l'UT à la suite de cette agrégation, désignée dans ce qui suit par UTA´, et la tête T de cette même UT avant l'agrégation, référencée par UTA. La mesure de similarité entre T´ et T est donnée par :

$$SimTêteUT(T',T) = \frac{\sum_{i=1}^{n} wg'_i \; appartient(g'_i,T)}{\sum_{i=1}^{n} wg'_i} \quad \text{avec}$$

$$appartient(g'_i,T) = 1 \text{ si } g'_i \quad T$$
$$= 0 \text{ si } g'_i \quad T$$

g´¡ est un graphe de T´

wg´¡ est le poids relatif de g´¡ dans UTA´

i est l'indice énumérant tous les graphes composant T´

La similarité entre T´ et T correspond donc à la somme des poids relatifs dans UTA´ des graphes communs à T´ et T rapportée à la somme des poids relatifs dans UTA´ de tous les graphes composant T´. Cette mesure rend compte à la fois de l'étendue de la différence entre les deux têtes comparées, i.e. leur nombre de graphes différents, et de l'importance de cette différence vis-à-vis de l'UT agrégée considérée, i.e. le poids relatif des graphes différents dans cette UT.

Pour qu'une telle mesure soit représentative, il ne doit pas y avoir de trop grandes disparités quant au nombre de graphes rassemblés respectivement par T et T´. Cette condition est remplie lorsque l'UT agrégée regroupe déjà un certain nombre d'UTs

textuelles. En effet, même si l'agrégation d'une nouvelle UT entraîne des modifications dans les poids relatifs des graphes de l'UT agrégée, le rapport global du nombre de graphes sur le nombre d'agrégations ne change alors pas beaucoup. Dans le cas contraire, on aurait en fait peu de chances d'avoir une similarité entre les deux UTs. La situation est différente lorsque l'on se trouve dans les phases initiales de la formation d'une UT agrégée. Si les UTs sont assez différentes, à la fois quant au nombre de graphes et quant à la nature de ceux-ci, des variations assez marquées de la taille de la tête de l'UT agrégée sont possibles. Elles ne posent cependant pas problème dans la mesure où elles ne font que rendre compte de l'instabilité réelle de l'UT agrégée sur cette période.

Pour que l'abstraction d'une UT agrégée puisse être décidée, il est nécessaire que la mesure de similarité entre la tête de cette UT avant et après sa dernière agrégation soit supérieure à un seuil fixé a priori, le dépassement devant être enregistré pour un certain nombre d'agrégations. Nous avons expérimentalement choisi une valeur de 0,7 pour le seuil, la mesure de similarité entre têtes étant comprise entre 0 et 1, et une valeur de 3 agrégations pour la durée pendant laquelle cette propriété doit être observée.

Le choix de faire porter le critère d'abstraction des UTs agrégées sur leur évolution temporelle oblige par ailleurs à compléter leur structure. Chaque UT agrégée doit en effet conserver une référence vers les graphes formant sa tête, de même qu'une référence vers ceux qui formaient sa tête avant sa dernière agrégation. Enfin, elle doit également posséder un compteur indiquant combien d'agrégations à la suite ont eu lieu avec une mesure de similarité entre têtes supérieure au seuil fixé. Ce compteur est remis à zéro chaque fois que la mesure tombe en dessous de ce seuil.

Dans [Chalendar 1997], Gaël de Chalendar propose une autre méthode de détection de la capacité d'une UT agrégée à être abstraite. Cette méthode repose sur l'analyse de la distribution des graphes de l'UT agrégée tels qu'ils apparaissent à un moment donné au sein de celle-ci, sans tenir compte de son évolution. Elle se concrétise par ladétermination des graphes les plus représentatifs de l'UT. Ceux-ci sont dits émergents du fait qu'ils sont retenus ensuite pour former le nouveau schéma. Un graphe agrégé est considéré comme émergent si son nombre d'occurrences est supérieur à la somme de la moyenne des nombres d'occurrences des graphes de l'UT et de leur écart-type. L'abstraction de l'UT est décidée si la proportion de graphes émergents par rapport aux autres graphes est suffisante et si l'UT possède un nombre d'agrégations dépassant un seuil fixé a priori.

Théoriquement, les deux méthodes sont complémentaires. Une UT agrégée ne peut être abstraite en effet que si elle est stable et si ce vers quoi elle s'est stabilisée est intéressant. Dans la pratique, l'utilisation de l'une des deux est généralement suffisante. Nous accorderons néanmoins une préférence à celle fondée sur la stabilité des UTs

agrégées. Une UT agrégée ne peut en effet être stable du point de vue de l'agrégation si elle ne possède pas suffisamment de graphes émergents. Dans un tel cas, on aurait en effet des difficultés à trouver une similarité avec des UTs textuelles. Au contraire, la présence de graphes émergents ne garantit pas forcément la stabilité de l'UT agrégée, ceux-ci pouvant changer au cours du temps tout en restant en proportion acceptable.

3. Sélection et abstraction des événements

3.1. Principes

Cette partie couvre les étapes (2), (3) et (4) du processus d'abstraction et peut être vue globalement comme la détermination de ce que l'on retient de l'UT agrégée que l'on a décidée d'abstraire, en complémentarité avec l'étape suivante consistant à créer un nouveau schéma à partir du matériau sélectionné. Cette détermination passe par une généralisation des événements constitutifs de l'UT agrégée dans la mesure où l'on cherche à construire une connaissance suffisamment générale. Cette généralisation doit cependant respecter deux contraintes assez contradictoires : d'une part elle doit être réalisée en accord avec les principes généraux de l'abstraction de schémas, donc ne pas faire appel à des connaissances pragmatiques définies a priori, et d'autre part son coût doit rester acceptable.

L'impossibilité d'utiliser une théorie du domaine du fait de la première contrainte nous a conduit à adopter un biais de nature formelle pour la généralisation des graphes de l'UT agrégée. Nous avons repris pour ce faire l'idée développée dans [Chalendar 1997] d'un biais fondé sur l'attribution d'un coût à chaque opération de généralisation et de la fixation d'un coût maximum limitant de manière globale le nombre des opérations de généralisation possibles. Ce biais présente l'avantage de prendre également en considération, au moins pour partie, la seconde contrainte mentionnée ci-dessus.

En revanche, cette seconde contrainte nous a écarté de la solution également proposée dans [Chalendar 1997] et consistant à utiliser un algorithme générique de généralisation de graphes conceptuels, dans la lignée des travaux décrits dans [Mineau 1992] et [Ellis 1992]. Il nous est en effet apparu que l'exploitation des caractéristiques spécifiques des représentations manipulées, en particulier le fait que les UTs agrégées opèrent déjà une forme de généralisation, est susceptible de garantir une plus grande efficacité du processus ainsi qu'une plus grande pertinence des résultats obtenus. Cette exploitation se traduit en effet par la mise en œuvre d'un traitement plus différencié des différents sous-problèmes posés par une telle généralisation.

Plus globalement, nous nous différencions du travail décrit dans [Chalendar 1997] sur le plan du séquencement des différentes étapes du processus d'abstraction. Dans ce travail en effet, l'abstraction commence par une sélection des graphes à retenir pour construire le schéma et se poursuit par l'application d'un algorithme de généralisation des graphes sélectionnés. De par son caractère générique, cet algorithme réunit à la fois l'étape de détection des regroupements possibles d'événements et celle de généralisation des regroupements obtenus.

Cette différence de séquencement s'explique par les contraintes posées par l'algorithme de généralisation de graphes conceptuels. Celui-ci étant d'un coût assez lourd, il est préférable de l'appliquer sur le moins grand nombre de graphes possible, donc de le faire précéder par l'étape de sélection des graphes qui élimine une bonne partie des graphes formant le corps des UTs agrégées. Par ailleurs, son caractère mono-bloc ne permet pas de le découper en plusieurs phases applicables séparément.

3.2. Évaluation des regroupements possibles entre événements

3.2.1. Objectifs

Comme nous l'avons vu au §1, cette phase permet de relâcher les contraintes très strictes posées sur les prédicats des graphes lors de la constitution des UTs agrégées. Mais au delà, ce relâchement de contraintes s'inscrit dans la réalisation de deux objectifs à la fois complémentaires et concurrents :

- regrouper des graphes qui, individuellement, ne seraient pas retenus pour construire un nouveau schéma mais qui, rassemblés et généralisés, donnent un graphe qui sera sélectionné.
  On reprendra dans ce qui suit l'appellation de graphe émergent introduite au §2.2.3 pour désigner tout graphe d'une UT agrégée jugé suffisamment représentatif pour contribuer à la création d'un schéma à partir de cette UT;

- renforcer et généraliser des graphes répondant déjà aux critères d'émergence. Il s'agit le plus souvent de regrouper un graphe émergent avec un graphe qui ne l'est pas, donc de renforcer la présence d'un événement parfois exprimé d'une façon un peu différente de celle sous laquelle il apparaît de façon majoritaire.
  Le regroupement de deux graphes émergents est également possible mais sans doute plus rare. Il renvoie en effet à la situation d'une UT agrégée dans laquelle un même événement est exprimé tantôt d'une façon et tantôt d'une autre, sachant que

les deux modes d'expression sont à peu près équilibrés du point de vue de leur fréquence d'occurrence.

Compte tenu de l'absence de connaissances de référence pouvant servir de biais, il convient de contraindre autant que faire se peut le problème de la généralisation des graphes. C'est dans cet esprit que nous accorderons une priorité au premier objectif par rapport au second. Celle-ci permet de limiter le nombre de généralisations en fixant un ordre partiel sur les regroupements de graphes possibles.

La préférence accordée au premier objectif découle naturellement de l'impact respectif de chacun des deux choix sur le contenu du schéma final. Si un événement n'est pas émergent, il n'apparaîtra pas dans le nouveau schéma et aucun mécanisme de restructuration ultérieur ne pourra venir corriger cette décision initiale. En revanche, lorsqu'un événement est déjà émergent avant la phase de regroupement des graphes, il le sera toujours à l'issue de celle-ci. Il semble donc plus important de rendre émergents les événements qui ne sont pas détectés comme tels du fait de différences de forme d'expression que de renforcer des événements qui seront de toute manière retenus pour construire le nouveau schéma.

3.2.2. La détermination des graphes émergents

Compte tenu des choix réalisés ci-dessus concernant le regroupement des graphes, la première phase de l'opération d'évaluation des regroupements possibles entre événements consiste à déterminer quels sont les graphes émergents de l'UT agrégée considérée. Pour réaliser ce tri, on calcule pour chaque graphe de cette UT agrégée un cœfficient d'émergence proche de celui proposé dans [Chalendar 1997][1]. Ce cœfficient situe le poids absolu de chaque graphe par rapport à la référence constituée par la somme de la moyenne des poids absolus des graphes considérés et de leur écart-type moyen. Cette référence est appelée seuil d'émergence et se note S_{emg}. Plus formellement, le cœfficient d'émergence d'un graphe agrégé g est donné par :

$$c(g) = \frac{n_{agr}(g)}{\overline{n_{agr}} + \sigma(n_{agr})}$$

avec

$n_{agr}(g)$, le nombre d'agrégations, i.e. le poids absolu, du graphe agrégé g;

$\overline{n_{agr}}$, la moyenne des poids absolus des graphes agrégés à généraliser;

$\sigma(n_{agr})$, l'écart-type moyen des poids absolus des graphes agrégés à généraliser.

[1] La seule différence réside dans l'emploi d'un écart-type moyen (la moyenne de la valeur absolue des distances à la moyenne), plus étroit que l'écart-type quadratique habituellement utilisé.

Pour qu'un graphe agrégé soit jugé comme émergent, il faut que son cœfficient d'émergence soit strictement supérieur à 1 : c(g) > 1. La moyenne et l'écart-type des poids absolus sont calculés par rapport aux graphes composant l'attribut dans lequel se trouve le graphe g dont on évalue le cœfficient d'émergence.

Le fait de se limiter au niveau des attributs et de ne pas prendre comme référence tous les graphes de l'UT agrégée considérée est motivé par deux raisons. La plus évidente tient à ce que des graphes appartenant à des attributs différents ont des rôles différents du point de vue de la situation représentée. Pour évaluer une caractéristique d'un graphe, on ne peut donc pas prendre comme référence des graphes qui ne lui sont pas comparables sur le plan fonctionnel.

La seconde raison est plus directement liée au processus d'abstraction de schémas. On a vu précédemment que tous les attributs n'ont pas le même comportement du point de vue de l'agrégation. Il faut donc tenir compte de ces spécificités lors de l'abstraction si l'on veut éviter certains effets secondaires. En prenant comme référence tous les graphes d'une UT agrégée, on risque par exemple de n'avoir que très rarement des graphes émergents au sein de l'attribut *Circonstances*. La variété des circonstances communément rencontrées pour une situation fait que les graphes de cet attribut ont souvent un poids plus faible que leurs homologues des autres attributs, même lorsqu'ils sont liés à juste titre à la situation considérée.

À l'inverse, ne pas respecter la spécificité des attributs risque d'engendrer du bruit au niveau de l'attribut *Description* et dans une moindre mesure, au niveau également de l'attribut *États Incidents*. Encore une fois du fait du caractère peu affirmé de l'attribut *Circonstances*, la moyenne des poids de l'ensemble des graphes d'une UT est généralement plus faible que celle des poids des graphes de l'attribut *Description* ou de l'attribut *États Incidents*. Le cœfficient d'émergence des graphes sera donc plus élevé si la référence prise est l'ensemble des graphes de l'UT. Dans un attribut comme *Description*, comportant fréquemment un petit ensemble de graphes de poids fort et un ensemble plus large de graphes de poids faible, cet accroissement de c(g) risque de faire déborder la sélection des graphes du cadre délimité par le premier ensemble pour englober également une partie du second. La probabilité de retenir des graphes liés seulement de façon contingente à la situation considérée est alors plus grande.

3.2.3. L'algorithme de regroupement des graphes

Principes

Le but de cette phase est de relâcher partiellement la contrainte d'égalité entre les types des prédicats des graphes que l'on agrège. On se contentera donc à ce niveau de représenter chaque graphe agrégé par le type de son prédicat. On ramène de cette manière le problème traité à celui de la généralisation d'un ensemble de types de concepts. Comme nous l'avons vu en préambule, cette généralisation s'effectue sous une contrainte de coût. Chaque type de concept *t* se voit associer un capital de généralisation, appelé $C_{gen}(t)$. Lorsqu'un type de concept est généralisé, c'est-à-dire qu'il est remplacé par son sur-type immédiat[1], il transmet son capital de généralisation au type qui le remplace, moyennant une amputation correspondant au coût de l'opération de généralisation. Lorsque ce capital est épuisé, le type de concept ne peut plus être généralisé.

L'objet de cette étape étant de regrouper des graphes, la généralisation d'un type de concept ne s'effectue pas indépendamment de celle des autres types de concept considérés. De façon minimale, une généralisation consiste en effet à remplacer deux types de concept par leur sur-type commun minimal[2]. En dehors de la contrainte portant sur le capital de généralisation de chacun des deux types de concept impliqués, cette opération est soumise également à une contrainte sur son coût global : la somme des diminutions du capital de généralisation des deux types de concept doit rester inférieure à un seuil S_{cgen}. Bien entendu, pour deux types t1 et t2, on a initialement : $C_{gen}(t1) + C_{gen}(t2) > S_{cgen}$. Cette contrainte vise à limiter les généralisations trop importantes qui conduiraient à regrouper des événements trop disparates. Il faut rappeler à cet égard que la notion de typicalité d'un événement par rapport à une situation est importante. La généralisation cherche à s'affranchir des liens avec les épisodes particuliers mais elle ne doit pas conduire à perdre cette typicalité.

Le regroupement des graphes ne s'opère pas non plus sans tenir compte du rôle fonctionnel des graphes vis-à-vis de la situation. C'est pourquoi seuls les graphes appartenant à un même attribut sont regroupés entre eux.

Définition

Le mécanisme de généralisation exposé ci-dessus ouvre le plus souvent la voie à tout un ensemble de solutions. Il faut donc disposer d'un critère de jugement permettant de

---

[1] Ce sur-type est unique dans le cas présent puisque le treillis des types de concept est restreint à un arbre.

[2] La restriction du treillis des types de concept à un arbre implique également l'unicité du sur-type commun minimal de deux types de concept.

définir qu'une solution est plus satisfaisante qu'une autre. Puisque nous avons fixé au regroupement des graphes l'objectif premier de faire émerger de nouveaux graphes, nous avons choisi de privilégier les généralisations regroupant le plus grand nombre de types de concept tels que le poids absolu des graphes que ces types représentent dépasse S_{emg}, le seuil d'émergence. Lorsque les généralisations sont équivalentes du point de vue de ce critère, nous accordons notre préférence à celles possédant les types de concept les plus spécifiques. Ce choix est motivé, comme ci-dessus pour la contrainte sur le coût global d'une généralisation élémentaire, par le souci de respecter la spécificité de la situation représentée.

Le cadre exposé précédemment définit un espace de généralisation qu'il faut explorer afin de trouver la généralisation satisfaisant le critère recherché[1]. La stratégie la plus simple et la plus évidente est bien entendu celle consistant à engendrer toutes les généralisations possibles et à les tester les unes après les autres afin de déterminerlaquelle répond aux contraintes fixées. On a de cette manière la garantie de trouver une généralisation optimale. Mais il est évident qu'il s'agit d'une stratégie particulièrement coûteuse qui se heurte dans sa réalisation pratique à la complexité algorithmique intrinsèque au problème dès lors que le nombre de graphes est un peu important. Une mise en œuvre réaliste nécessite donc de sacrifier la garantie de trouver un optimum au profit d'heuristiques s'appuyant sur les spécificités du problème pour élaguer l'espace de généralisation.

La plus importante de ces heuristiques concerne la stratégie adoptée pour réaliser les regroupements. Afin de tenir compte des priorités définies en préambule et relatives au caractère émergent ou non émergent des graphes, nous avons organisé le processus de regroupement des graphes en trois phases successives :

- regroupement des prédicats des graphes non émergents[2];
- regroupement des prédicats résultant de la première passe avec ceux des graphes émergents;
- regroupement entre eux des prédicats résultant de la deuxième passe.

Ces trois phases sont de même nature dans la mesure où elles mettent en œuvre un algorithme identique d'exploration des généralisations possibles. Cet algorithme est néanmoins paramétré afin de prendre en considération les spécificités de chacune de ces

[1] Pour être exact, même lorsqu'on lui ajoute la contrainte de minimisation du niveau de généralité des types de concept, le critère de la maximisation du poids des graphes ne garantit pas la sélection d'une seule généralisation, même si la probabilité en est très forte.

[2] En utilisant le terme prédicat, on fait référence de façon plus précise au type du prédicat.

phases. Il repose sur trois niveaux de contrôle que l'on trouvera formalisés au travers de l'algorithme de la figure 7.1.

```
indicateurStabilité ← faux
Tantque non indicateurStabilité faire
    indicateurGénéralisation ← faux
    tri(P_gen)
    pred_gen ← premierÉlément(P_gen)
    construction de Q_gen
    Tantque non indicateurGénéralisation et non atteinteFin(P_gen) faire
        Pour tous les éléments (pred_i) de Q_gen faire
            stcm ← surTypeCommunMinimum(pred_gen,pred_i)
            Si (coûtGénéralisation(pred_gen,st) ≤ capitalGénéralisation(pred_gen)) et
            (coûtGénéralisation(pred_i,st) ≤ capitalGénéralisation(pred_i)) et
            (coûtGénéralisation(pred_gen,st) + coûtGénéralisation(pred_i,st) ≤ 3) alors
                mettre à jour Str_gen avec stcm
            FinSi
        FinPour
        Si nonVide(Str_gen) alors
            pred_sel ← choix dans Str_gen du type de concept répondant aux critères de la phase courante
            capitalGénéralisation(pred_sel) ← min(capitalGénéralisation(pred_gen) -
                coûtGénéralisation(pred_gen,pred_sel), capitalGénéralisation(pred_k) -
                coûtGénéralisation(pred_k,pred_sel)),
                avec pred_k, le prédicat avec lequel pred_gen a été généralisé
            indicateurGénéralisation ← vrai
            supprimer(pred_gen,P_gen)
            supprimer(pred_k,P_gen)
        Sinon
            pred_gen ← élémentSuivant(P_gen)
            construction de Q_gen
        FinSi
    FinTantque
    indicateurStabilité ← non indicateurGénéralisation et atteinteFin(P_gen)
FinTantque
ajouter les éléments de P_gen à configurationRésultat
renvoyer configurationRésultat
```

Fig. 7.1 - Algorithme de regroupement des prédicats d'un attribut d'UT agrégée utilisé par chacune des phases

Le plus élevé de ces niveaux est celui gérant la terminaison d'une phase et donc, le déclenchement de la phase suivante. Il lance la recherche de nouvelles généralisations tant qu'une stabilité de l'ensemble des prédicats considérés, en l'occurrence ceux d'un attribut d'UT agrégée, n'a pas été détectée. La stabilité d'un prédicat du point de vue de la généralisation est définie par l'impossibilité de trouver un autre prédicat avec lequel une généralisation respectant les contraintes de coût fixées est possible. Un cas particulier de cette définition de la stabilité est la situation dans laquelle le capital de généralisation d'un

prédicat est égal à zéro. Il faut préciser qu'à ce niveau de contrôle, les généralisations trouvées ne constituent pas des possibilités parmi lesquelles un choix ultérieur est à réaliser. Elles viennent s'ajouter les unes aux autres pour former une solution unique.

Le second niveau de contrôle est celui assurant le pilotage de la recherche d'une nouvelle généralisation. Il fait intervenir une deuxième heuristique essentielle afin de contraindre l'évolution au sein de l'espace des généralisations. Celle-ci consiste à établir un ordre de préférence a priori sur les prédicats que l'on choisit comme candidats à une généralisation.

À chaque itération du niveau supérieur de contrôle, on parcourt ainsi l'ensemble P_{gen} des prédicats disponibles à un moment donné de la phase selon le critère d'ordre fixé et l'on tente pour chacun d'entre eux de trouver une généralisation avec un ou plusieurs des autres prédicats de ce même ensemble P_{gen}. L'énumération des prédicats de P_{gen} s'arrête dès que l'on trouve un prédicat pour lequel une généralisation est possible. Dans ce cas, les prédicats impliqués dans la généralisation en question sont supprimés de P_{gen} et le type de concept qui en constitue le résultat y est ajouté. À l'itération suivante du niveau supérieur de contrôle, la recherche d'une généralisation sera effectuée à partir de ce nouvel ensemble P_{gen}, retrié afin de tenir compte du résultat de l'itération précédente. Une fois réalisée, une généralisation est donc acquise et ne sera pas remise en cause par une autre, éventuellement plus intéressante.

La recherche d'une généralisation prend fin également lorsque tous les prédicats de P_{gen} ont été parcourus sans qu'une généralisation n'ait pu être trouvée. On considère dans ce cas que la configuration de prédicats atteinte est stable, ce qui est le critère déclenchant le passage à la phase suivante.

Le contenu initial de P_{gen} dépend de la phase dans laquelle on se trouve. Au début de la première phase, P_{gen} rassemble l'ensemble des prédicats des graphes non émergents. Au début de la deuxième, cet ensemble ne regroupe au contraire que les prédicats des graphes émergents écartés lors de la première phase. Au début de la dernière enfin, il est formé du résultat complet de la deuxième phase.

L'ordre dans lequel on énumère les prédicats de P_{gen} est lui aussi dépendant de la phase considérée. Il repose à chaque fois sur un arrangement particulier de deux caractéristiques des prédicats : leur poids et leur degré de spécificité. Lorsqu'un prédicat fait partie directement d'un des graphes initiaux, son poids est égal au poids absolu du graphe dans lequel ce prédicat est présent. Lorsqu'il est le produit de la généralisation de deux prédicats ou plus, il est égal à la somme des poids des prédicats dont il est la généralisation. Le degré de spécificité d'un prédicat correspond quant à lui à sa profondeur dans le treillis des types de concept, sachant que plus un prédicat est profond

et plus il est spécifique. Dans les rares cas où l'association de ces deux critères ne détermine pas un ordre unique, on s'en remet à un choix arbitraire.

Pour la première phase, l'ordre d'énumération des prédicats de P_{gen} suit d'abord les valeurs décroissantes des poids des prédicats et, en cas de poids égaux, se conforme aux valeurs décroissantes de leur spécificité. Pour la deuxième phase, l'importance relative des deux critères est inversée : on privilégie d'abord les prédicats les plus spécifiques et, à spécificités égales, on traite en premier le prédicat de plus faible poids. Les principes d'ordonnancement pour la dernière phase sont les mêmes que pour la deuxième.

Dans le cas de la première phase, l'utilisation du poids des prédicats pour tracer un chemin dans l'espace des généralisations correspond à la transposition à un échelon local du critère global de jugement des généralisations. Celui-ci privilégie les configurations comportant le plus possible de prédicats ayant un poids les plaçant au-dessus du seuil d'émergence. On cherche donc au niveau local à former les généralisations ayant le poids le plus fort possible. Pour cela, on a choisi de privilégier les généralisations impliquant les prédicats de plus fort poids. Ces deux heuristiques ne garantissent pas une optimisation du critère global. Elles s'apparentent en effet à une stratégie de type Best first ou, dans le domaine du continu, à une descente de gradient : on s'oriente à chaque point de choix vers l'état maximisant de façon immédiate le critère à optimiser globalement en espérant qu'une succession de petites optimisations locales permettent de parvenir à la meilleure configuration[1].

En dépit de cette absence de garantie, on peut penser qu'il s'agit là d'un bon compromis entre coût et efficacité. Dans la mesure où le capital de généralisation associé à chaque prédicat reste faible, ce qui est le cas ici, la profondeur de l'arbre de recherche demeure faible elle aussi et il y a peu de chances de trouver une solution ne maximisant pas le critère de jugement retenu à chaque étape[2] tout en étant meilleure que celle donnée par la méthode exposée ici.

Le fait de privilégier au sein de chaque phase les prédicats les plus spécifiques obéit là encore à la logique consistant à ne pas perdre la typicalité des événements décrivant la situation décrite par l'UT à abstraire. Ce choix est mis en avant dans la seconde phase, et ensuite dans la dernière, étant donné que l'objectif premier du regroupement des graphes, c'est-à-dire l'émergence de nouveaux graphes, qui justifie de regrouper les prédicats de

---

[1] Pour être exact, nous nous situons dans une approximation d'une telle stratégie. D'une part, le fait de choisir le prédicat de plus fort poids n'implique pas, même à la suite d'une seule généralisation, d'obtenir le plus grand nombre de prédicats avec un poids au-dessus de S_{emg}. D'autre part, on n'examine pas après chaque généralisation quelles sont toutes les généralisations possibles à la suite pour choisir celle qui optimise le critère recherché.

[2] Une telle solution se traduit souvent par un chemin plus long en nombre d'étapes. Les contraintes de coût portant sur les généralisations en barrent donc l'accès.

plus fort poids, est déjà accompli à ce stade et ne pourra être poussé plus loin. Toutes les possibilités de regroupements entre graphes non émergents sont en effet épuisées lors de la première phase. Dans ces deux mêmes phases, l'utilisation du poids des prédicats comme critère de tri second favorisant les graphes émergents de faible poids se justifie par un souci d'équilibrage de la représentativité des différents événements du schéma qui sera formé. L'efficacité de l'équilibrage suppose bien entendu que les différences observées proviennent bien de la contrainte d'égalité des prédicats lors de l'agrégation.

Le troisième et dernier niveau de contrôle de l'algorithme de regroupement des graphes gère la recherche d'une généralisation pour un prédicat choisi par le deuxième niveau de contrôle et désigné dans ce qui suit par $pred_{gen}$. La stratégie retenue ici se distingue de celle des niveaux supérieurs puisqu'elle consiste à développer toutes les généralisations possibles avec les autres prédicats disponibles et à choisir celle répondant le mieux aux critères fixés. Elle s'appuie pour ce faire sur une structure de données, dénommée Str_{gen}, regroupant toutes les généralisations entre $pred_{gen}$ et un ou plusieurs prédicats de l'ensemble des prédicats disponibles, appelé Q_{gen}. Cette structure associe à un type de concept son poids ainsi que les prédicats dont il représente la généralisation.

La procédure consiste à énumérer tous les éléments de Q_{gen} , l'ordre étant dans ce cas sans importance, et à examiner si une généralisation est possible entre $pred_{gen}$ et le prédicat de Q_{gen} considéré. Cet examen tient compte à la fois de la contrainte portant sur le capital de généralisation des deux prédicats et de celle relative au coût total de la généralisation. Si le type de concept résultat de cette généralisation est déjà présent dans Str_{gen}, on se contente d'actualiser les données qui y sont rattachées en y ajoutant le prédicat courant venant de Q_{gen} et en augmentant son poids de celui de ce prédicat. S'il est absent de cette structure, on l'y ajoute avec ses données. En final, il suffit de parcourir Str_{gen} et d'appliquer les critères de sélection à chacun de ses types de concept pour obtenir la généralisation de $pred_{gen}$, à condition bien entendu que Str_{gen} ne soit pas vide.

L'ensemble Q_{gen} des prédicats disponibles ainsi que les critères de sélection finale de la généralisation retenue pour $pred_{gen}$ dépendent comme précédemment de la phase dans laquelle on se trouve. Le contenu de Q_{gen} est ainsi le complémentaire de celui de P_{gen} tel que toutes les généralisations caractéristiques de la phase considérée puissent être explorées. Dans la première phase, Q_{gen} est de ce fait composé de l'ensemble des prédicats correspondant aux graphes non émergents et dans les deux dernières phases, il rassemble les prédicats résultant de la phase précédente. En ce qui les concerne, les critères de sélection reprennent pour l'essentiel les critères de tri des prédicats de P_{gen}. Dans la première phase, on retient ainsi le type de concept de plus fort poids, avec une prime au plus spécifique en cas d'égalité des poids. Dans les deux dernières phases, le

type retenu est d'abord le plus spécifique et ensuite celui de poids le plus fort. La différence concernant ce dernier point par rapport aux critères de tri de P_{gen} s'explique par une nécessaire complémentarité : si l'on donne la priorité au renforcement des prédicats les plus faibles, il faut retenir leurs généralisations ayant le poids le plus fort.

Sur le plan de la généralisation des types de concept à proprement parler, nous avons retenu une valeur de 2 pour le capital de généralisation associé à chaque prédicat et la même valeur en ce qui concerne le coût total d'une généralisation élémentaire entre deux prédicats. En considérant que le passage d'un type à son sur-type se chiffre à un coût de 1, les valeurs choisies autorisent donc les configurations suivantes, avec p1 et p2, les deux prédicats à généraliser, et p, leur généralisation :

- p est le sur-type direct de p1 et de p2;
- p est le sur-type direct de p1, respectivement p2, et le sur-type à un pas de deux de p2, respectivement p1;
- p s'assimile à p1, respectivement à p2, et p2, respectivement p1, est un sous-type direct ou un sous-type à deux pas de p1, respectivement p2.

De façon générale, les généralisations trop poussées sont défavorisées par une contrainte supplémentaire portant sur le coût du passage d'un type à son sur-type. Afin de caractériser de façon assez primitive la notion de typicalité, nous avons en effet choisi d'alourdir ce coût dans la partie supérieure du treillis. On limite de cette façon les éventuelles sur-généralisations. Dans les deux tiers inférieurs du treillis, le passage d'un type à son sur-type possède un coût de 1 tandis que dans son tiers supérieur, ce coût est porté à 2. En tout état de cause, une généralisation ne peut ni dépasser, ni même égaler le niveau des primitives, c'est-à-dire le niveau des types immédiatement inférieurs au sommet du treillis (cf. chapitre 4).

La généralisation de deux prédicats p1 et p2 est encadrée par leur capital de généralisation mais en rétro-action, elle détermine le capital de généralisation du prédicat p qu'elle engendre. L'idée présidant à cette détermination est qu'un prédicat ne peut en aucun cas être généralisé au-delà de la limite fixée par son capital de généralisation initial, même si cette généralisation recouvre plusieurs généralisations élémentaires. Pour s'assurer du respect de cette contrainte, on soustrait au capital de généralisation de chacun des deux prédicats p1 et p2 le coût de sa généralisation jusqu'à p et l'on retient comme capital de généralisation de ce dernier le minimum des deux valeurs obtenues.

Compte tenu du capital de généralisation limité associé à chaque prédicat, il est par ailleurs aisé de se rendre compte que les généralisations sont de moins en moins fréquentes à mesure que les phases s'accumulent. Le capital de généralisation global de

l'ensemble des prédicats considérés, en l'occurrence ceux d'un attribut d'UT agrégée, diminue en effet à l'issue de chaque nouvelle phase et si le capital associé à chaque prédicat n'est pas significativement plus important que le nombre de phases – il lui est même inférieur dans le cas présent – il se trouve bien souvent épuisé quand on aborde les dernières phases ou tout du moins il est insuffisant pour que la jonction se fasse avec un autre prédicat. Cet effet est renforcé par l'augmentation du coût de remontée dans le treillis lorsque l'on passe dans la partie supérieure de celui-ci, ce qui intervient bien entendu plutôt lors des dernières phases.

Du point de vue de l'algorithme présenté ci-dessus, la marge de liberté accordée à la généralisation est une variable que l'on ajuste au travers de différents paramètres tels que le capital initial de généralisation des prédicats, le coût limite de généralisation de deux prédicats ou encore le coût de passage d'un type à son sur-type. Les valeurs que nous avons adoptées pour ces paramètres définissent ici une marge relativement faible. Dans un tel cas de figure, il est vraisemblable que les généralisations seront peu nombreuses. Il est alors intéressant, avant de tester les regroupements possibles, de détecter les prédicats que l'on ne pourra jamais généraliser par le simple fait que leur capital de généralisation ne leur permet pas de trouver un sur-type commun minimal avec autre prédicat.

Pour réaliser cette détection, il suffit de construire tous les couples de prédicats possibles et de déterminer pour chacun le coût d'atteinte d'un sur-type commun minimal. Si pour un prédicat, ce coût dépasse son capital initial de généralisation pour tous les couples dans lesquels il est impliqué, on sait qu'on peut le laisser de côté pour l'étape de réalisation des regroupements.

Exemple

Nous donnons ici un exemple du fonctionnement de cet algorithme de regroupement de graphes. Cet exemple est avant tout illustratif plus qu'il n'est véritablement réaliste du point de vue des données sur lesquelles il s'appuie. La configuration considérée est formée de 10 prédicats, désignés par de simples lettres et accompagnés de leur poids initial : A (8), B(7), C (6), D (4), E (3), F (3), G (2), H (1), I (1), J (1).

Leur situation dans le treillis des types de concept est représentée par la figure 7.2. La moyenne des poids de ces prédicats étant égale à 3,6 et leur écart-type moyen à 2,12, le seuil d'émergence S_{emg} est fixé à 5,72. Seuls les graphes contenant les prédicats A, B et C ont de ce fait le statut de graphe émergent. En appliquant la phase préalable de détection des prédicats non regroupables, on repère que A ne peut être regroupé avec aucun autre prédicat et sera donc laissé de côté. Enfin, le treillis est formé de 5 niveaux. Le tiers supérieur représente deux niveaux : le sommet et le niveau des primitives. Étant donné

que celui-ci ne peut être dépassé, le coût de passage d'un type à son sur-type sera ici toujours égal à 1.

1ère phase :

P_{gen} : D [2](4), E [2](3), F [2](3), G [2](2), H [2](1), I [2](1), J [2](1)[1]

1ère itération : $pred_{gen}$: D; Q_{gen} : E, F, G, H, I, J

→ pas de généralisation : les généralisations avec F ou E ne sont pas possibles du fait du coût limite fixé pour chaque opération de généralisation entre deux prédicats.

2ème itération : $pred_{gen}$: E (choix arbitraire entre E et F); Q_{gen} : D, F, G, H, I ,J

→ généralisation avec F; P_{gen} : T22 [1](6), D [2](4), G [2](2), H [2](1), I [2](1), J [2](1)

3ème itération : $pred_{gen}$: T22; Q_{gen} : D, G, H, I ,J

→ généralisation avec D; P_{gen} : G [2](2), H [2](1), I [2](1), J [2](1)

4ème itération : $pred_{gen}$: G; Q_{gen} : H, I ,J

→ généralisation avec H; P_{gen} : T13 [1](3), I [2](1), J [2](1)

5ème itération : $pred_{gen}$: T13; Q_{gen} : I ,J

→ pas de généralisation : la généralisation avec I n'est pas possible du fait de la contrainte de demeurer sous le niveau des primitives, c'est-à-dire ici sous P1, P2 et P3.

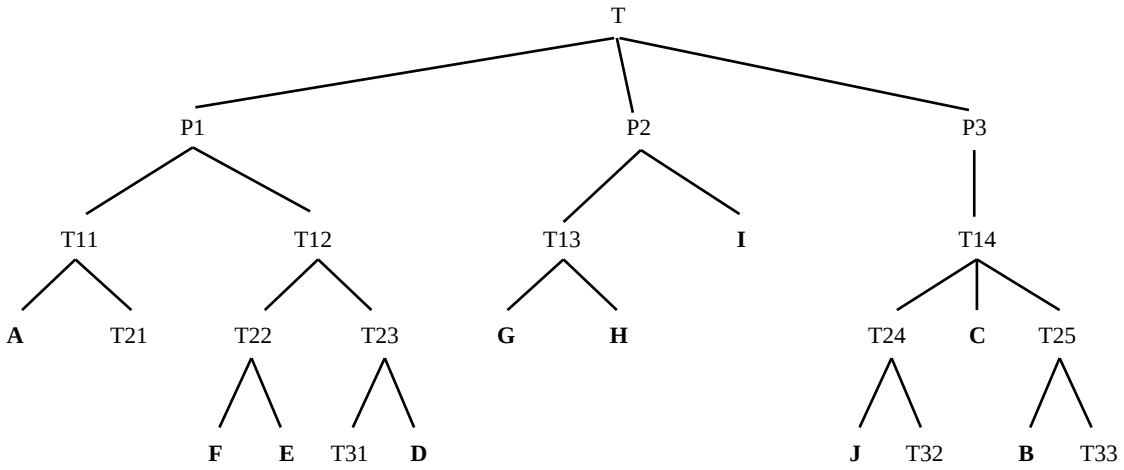


Fig. 7.2 - Treillis des types de concept utilisé pour l'exemple

On énumère ensuite les prédicats I et J sans plus de succès. Après la 7ème itération, on détecte donc la stabilité et l'on sort de cette première phase. En opérant le transfert des éléments restant dans P_{gen} vers configurationRésultat, on obtient pour cette dernière le contenu suivant : T12 [0](10), T13 [1](3), I[2](1), J[2](1).

[1] Les valeurs entre crochets correspondent au capital de généralisation des prédicats tandis que celles entre parenthèses sont leur poids.

2ème phase

P_{gen} : B[2](7), C[2](6)

1ère itération : $pred_{gen}$: B; Q_{gen} : T12, T13, I, J

→ pas de généralisation : la généralisation avec J n'est pas possible du fait du coût limite fixé pour chaque opération de généralisation entre deux prédicats.

2ème itération : $pred_{gen}$: C; Q_{gen} : T12, T13, I, J

→ généralisation avec J; P_{gen} : T14[0](7), B[2](7)

3ème itération : $pred_{gen}$: B; Q_{gen} : T12, T13, I

→ pas de généralisation

Une quatrième itération avec T14 ne donne pas non plus de nouvelle généralisation. Tous les éléments de P_{gen} ayant été parcouru une fois sans la survenue d'une généralisation, on peut clore la deuxième phase, avec configurationRésultat contenant alors :
T12 [0](10), B[2](7), T14[0](7), I[2](1).

3ème phase

P_{gen} : T12 [0](10), B[2](7), T14[0](7), I[2](1)

1ère itération : $pred_{gen}$: T12; Q_{gen} : B, T14, I

→ pas de généralisation : aucun autre prédicat ne peut être rejoint sans passer par le niveau des primitives

2ème itération : $pred_{gen}$: B; Q_{gen} : T12, T14, I

→ généralisation avec T14, Pgen : T14 [0](14), T12 [0](10), I[2](1)

3ème itération : $pred_{gen}$: T14; Q_{gen} : T12, I

→ pas de généralisation : aucun autre des deux prédicats possibles ne peut être rejoint sans passer par le niveau des primitives

Les deux itérations énumérant les derniers éléments de P_{gen} ne donnent pas non plus lieu à une généralisation, ce qui met donc un terme à cette dernière phase.
La configuration obtenue à l'issue de cette opération de regroupement des prédicats est donc : T14 (14), T12 (10), A (8), I(1).

Discussion

Ainsi que nous l'avons souligné précédemment, il semble raisonnable de penser qu'en dépit d'une stratégie ne garantissant pas une solution optimale du point de vue des critères retenus, l'algorithme de généralisation présenté s'en approche et l'atteint dans la plupart des cas. Il est cependant nécessaire de préciser que la contrainte portant sur le coût de généralisation de deux prédicats constitue une source de variabilité de ces résultats. Dans l'exemple de la figure 7.2, on a vu qu'il est possible de généraliser les prédicats E, F et D. En revanche, si l'un des deux membres du couple (E, F) disparaissait, la généralisation du membre restant avec D serait impossible car le coût total de

généralisation des deux prédicats, égal à 4, dépasserait alors le plafond, fixé à 3. Ce comportement peut paraître un peu indésirable au premier abord dans la mesure où l'écart entre E, F, D et T12 est toujours le même dans tous les cas.

En réalité, il ne fait que mettre au jour le mécanisme suivant : la contrainte considérée limite l'ampleur de la généralisation de tous les prédicats en général, et celle des prédicats isolés en particulier, c'est-à-dire des prédicats situés dans une partie du treillis dans laquelle leur densité est faible. Cette contrainte permet de s'adapter aux caractéristiques du treillis de types de concept servant de support à la généralisation. Si celui-ci est assez profond, on est obligé de fournir un capital de généralisation important aux prédicats. Les risques d'avoir des généralisations abusives augmentent en conséquence. Cette contrainte de coût plus globale limite alors les risques de telles sur-généralisations en garantissant que les généralisations de large ampleur se déroulent en plusieurs étapes, comme c'est le cas avec E, F et D dans l'exemple ci-dessus.

3.3. Sélection des événements représentatifs

Bien que le résultat final de la sélection des événements soit obtenu à l'issue de cette étape, une part importante du travail correspondant est en fait réalisée lors de la détermination des graphes émergents de chaque attribut de l'UT agrégée considérée, traitement réalisé préalablement à l'étape de regroupement des événements. C'est en particulier à cette occasion qu'est calculé S_{emg}, le seuil d'émergence.

Ce même seuil d'émergence est utilisé pour la présente étape. À la suite des regroupements entre événements, le nombre des prédicats a changé et leur poids absolu, au moins pour une part d'entre eux, a été modifié par rapport à la configuration initiale de l'UT agrégée. Nous conservons néanmoins la même référence pour sélectionner les prédicats qui contribueront à la construction d'un nouveau schéma. Le but de l'étape de regroupement n'est pas en effet de changer du tout au tout l'ensemble des événements à retenir mais de l'élargir un peu en relâchant les contraintes artificiellement strictes fixées par la similarité entre UTs. En particulier, les prédicats appartenant à des graphes émergents, ou éventuellement les prédicats qui en sont les généralisations, doivent faire partie d'office des événements retenus. Le cœur de cette étape de sélection est donc particulièrement simple puisqu'il consiste à ne conserver que les prédicats dont le poids absolu dépasse S_{emg}[1]. Dans l'exemple développé au §3.2, on retiendrait ainsi les prédicats T14, T12 et A, le poids de I, égal à 1, étant le seul à rester inférieur au seuil d'émergence, égal à 5,72.

[1] Par définition, c'est le cas des prédicats appartenant aux graphes émergents.

Le mécanisme décrit ci-dessus s'applique à chacun des attributs de l'UT agrégée à abstraire. Dans le chapitre 6, on a vu que ces attributs présentent des différences, notamment en ce qui concerne le nombre de graphes rassemblés ainsi que sur la répartition des poids de ces graphes. Il apparaît donc nécessaire, à la suite de la sélection précédente, d'opérer une sorte de lissage entre les différents attributs. Ceci est d'autant plus exact que le mode de sélection au sein des attributs conduit toujours à retenir au moins un événement, ce qui n'est pas forcément pertinent dans le cas d'un attribut tel que *Circonstances*. Ce lissage s'appuie sur le poids relatif des regroupements de graphes issus de l'étape précédente[1] (le terme 'regroupement' est utilisé de façon générique et un regroupement peut ne contenir qu'un seul graphe). Ce poids est en effet le caractère permettant de les ramener à la même référence, en l'occurrence leur importance vis-à-vis de leur UT agrégée.

Ce lissage s'opère en deux temps. Le premier consiste à éliminer les regroupements qui ont été sélectionnés mais que l'on juge insuffisamment représentatifs en comparaison de l'ensemble de ceux retenus. Pour ce filtrage, on s'appuie sur une référence définie comme suit : on détermine pour chaque attribut le minimum des poids relatifs de ses regroupements et l'on calcule la moyenne des minima obtenus pour l'ensemble des attributs. Lorsque le poids relatif d'un des regroupements retenus est inférieur à cette moyenne, il est éliminé des événements sélectionnés.

Dans un mouvement inverse du premier, le second temps du lissage consiste à récupérer un certain nombre de regroupements initialement laissés de côté lors du traitement individuel de chaque attribut. À l'échelle de l'UT agrégée, ces regroupements ont un poids au moins comparable à celui d'autres regroupements sélectionnés dans d'autres attributs que celui dans lequel ils figurent. Le fait de ne pas avoir été sélectionnés précédemment résulte simplement de la présence au sein de leur attribut d'un nombre conséquent de regroupements de plus fort poids, ce qui gomme leur représentativité plus globale. Pour compenser cet effet, on intègre donc parmi les regroupements sélectionnés tous ceux dont le poids relatif est supérieur ou égal au poids relatif du regroupement sélectionné ayant le plus faible poids relatif (en prenant bien entendu comme référence le résultat du premier temps du lissage).

Ce mouvement consistant à supprimer certains événements et au contraire, à en ajouter d'autres, doit néanmoins garantir en final le respect des contraintes minimales d'existence d'un schéma. En l'occurrence, ce sont les mêmes que pour les autres UTs : il faut au

[1] Le poids relatif des regroupements de graphes est calculé de la même façon que celui des graphes. La seule différence est qu'il peut être éventuellement supérieur à 1.

moins que le schéma possède un attribut *Description* non vide. On peut trouver étrange qu'un schéma puisse avoir un de ses attributs vide. Il s'agit là en fait d'un compromis acceptable. Il faut constater en effet que d'une part, le fait de retenir systématiquement un événement pour un attribut ne garantit pas sa pertinence et que d'autre part, le processus permettant de savoir si un schéma pourra être créé avec ses trois attributs rassemblant des événements pertinents est un peu lourd pour être appliqué comme critère systématique d'abstraction d'un schéma. Une solution un peu plus coûteuse que celle décrite ici consiste à suivre le cheminement décrit ci-dessus mais à repousser l'abstraction lorsque l'on constate que le contenu d'un des attributs ne s'est pas dessiné de façon suffisamment précise.

3.4. Généralisation des événements

Après que les graphes ont été regroupés en fonction de leur prédicat et qu'ils ont été sélectionnés, il reste à transformer chaque regroupement en un nouveau graphe, celui-ci étant issu de la généralisation des graphes formant le regroupement. Cette généralisation s'effectue en trois étapes. La première d'entre elles consiste à apparier tous les graphes du regroupement et à les joindre de façon à n'obtenir qu'un seul graphe agrégé. À ce stade, chaque rôle[1] du graphe rassemble l'ensemble des concepts occupant ce rôle dans les différents graphes regroupés. Dans le but de former un véritable graphe conceptuel à partir du regroupement initial, la deuxième étape est chargée de généraliser l'ensemble des concepts existant pour chaque rôle du graphe. Cette généralisation n'intervient que si le rôle en question possède une présence suffisamment affirmée. La dernière étape, enfin, est complémentaire de la deuxième puisqu'elle a pour but de supprimer les concepts dont le rôle est plus accessoire et/ou la présence plus épisodique.

En pratique, les deux premières étapes sont réalisées par une seule et même opération. Celle-ci est équivalente à l'agrégation des graphes intervenant lors de la constitution de la mémoire épisodique. Les deux opérations se différencient principalement sur deux points. L'un d'entre eux réside dans le fait que la nouvelle opération fait intervenir plusieurs graphes agrégés alors que l'agrégation définie précédemment est réalisée entre un graphe agrégé et un graphe conceptuel 'normal'. Sur le plan de la nature des graphes, la différence ne relève que du niveau de la mise en œuvre informatique puisqu'un graphe conceptuel 'normal' est parfaitement équivalent, du point de vue informationnel, à un graphe agrégé ne regroupant qu'un seul graphe. La disparition de la différence de nature entre les arguments supprime par ailleurs la dissymétrie de l'agrégation originelle et

[1] Le rôle d'un concept dans un graphe est défini par l'ensemble de ses relations avec les autres concepts du graphe.

permet de ramener la fusion de plusieurs graphes à une succession de fusions de paires de graphes.

Le second point de différenciation concerne le cadre de généralisation adopté. Celui-ci n'est plus le graphe canonique associé au type des prédicats des deux graphes mais le graphe canonique associé au type représentant la généralisation de tous les prédicats des graphes regroupés. Pour se replacer le plus possible dans le cadre d'une agrégation habituelle, il suffit de remplacer, préalablement à l'opération, le type du prédicat de tous les graphes par la généralisation retenue pour le regroupement. Comme lors d'une agrégation habituelle, les types des concepts occupant les mêmes rôles sont généralisés et les concepts occupant des rôles non présents dans les autres graphes sont ajoutés. La généralisation des types s'effectue également de la même manière. Elle consiste à prendre leur sur-type commun minimal lorsque celui-ci est inférieur au type du concept équivalent dans le graphe canonique associé au type du prédicat généralisé. S'il n'existe pas de généralisation inférieure à ce type, on retient le type de concept rassemblant le plus grand nombre d'occurrences. Cette opération d'agrégation modifiée réalise donc bien à la fois le travail de la première et de la deuxième étape de généralisation des regroupements.

La troisième étape de ce processus ne concerne que les concepts n'ayant pas d'équivalent au niveau du graphe canonique associé au type du prédicat généralisé. Ils représentent en pratique des caractéristiques associées aux objets (leur couleur, leur taille, leur forme, ...), aux personnes (leur rôle social, leur métier, ...) ou aux actions (caractéristiques aspectuelles, spatio-temporelles, ...). La présence de ces caractéristiques est le plus souvent très contingente mais certaines d'entre elles peuvent être néanmoins typiques d'une situation. C'est pourquoi la dernière étape de généralisation est chargée de statuer sur leur présence dans le schéma nouvellement créé.

Le critère adopté repose à la fois sur le poids du concept et son degré de spécificité : pour être retenu, le nombre d'occurrences d'un tel concept ramené au poids absolu du regroupement doit être supérieur à un seuil fixé a priori; d'autre part, le type généralisant l'ensemble des types des occurrences de ce concept ne doit pas dépasser une limite elle aussi fixée a priori par rapport au type le plus spécifique de ces mêmes occurrences. Pour le seuil concernant le poids des concepts, nous avons fixé une limite à 0,5. L'ampleur de la généralisation dans le treillis des types est quant à elle restreinte à 1 : le sur-type commun minimal de l'ensemble des occurrences doit être leur sur-type direct.

4. Construction des schémas

Toutes les étapes précédentes visent en définitive à la préparation de celle-ci puisqu'elles sont chargées de mettre en évidence les éléments d'une UT agrégée sur lesquels on peut s'appuyer afin de construire un schéma. Cette étape a pour mission de réaliser le changement de représentation proprement dit. Celui-ci se définit au travers de trois grands problèmes, s'identifiant chacun à la construction d'une des trois composantes d'un schéma : son entête, son corps et ses rôles. Nous n'aborderons ici, et encore que de façon relativement partielle, que la construction du corps et de l'entête d'un schéma. Nous reprendrons pour l'essentiel les propositions faites à ce sujet dans [Chalendar 1997].

4.1. Construction du corps du schéma

À l'issue de l'étape précédente, on dispose pour chaque attribut de l'UT agrégée à abstraire d'un ensemble de graphes conceptuels représentant des événements à la fois suffisamment typiques de la situation considérée et suffisamment généraux pour figurer dans une représentation abstraite de cette situation. Néanmoins, il s'agit là de la représentation sémantique de ces événements. Or, au niveau de la mémoire pragmatique, la représentation pragmatique d'une situation, c'est-à-dire un schéma, est décrite comme l'agencement des représentations pragmatiques de situations plus élémentaires.Autrement dit, le corps d'un schéma est composé d'un ensemble de références vers d'autres schémas. La construction de ce corps implique donc de transformer les graphes conceptuels obtenus précédemment en références vers des schémas représentant les mêmes événements.

Pour transformer un graphe en référence, deux cas de figure se présentent : soit le schéma représentant la situation évoquée par le graphe existe déjà au sein de la mémoire pragmatique et il suffit donc de le retrouver pour y faire référence; soit un tel schéma n'existe pas et il est nécessaire de le créer, au moins sous la forme minimale d'un schéma terminal (cf. chapitre 4). Par la suite, celui-ci pourra être éventuellement complété par l'abstraction d'une autre UT agrégée.

Dans les deux cas, la première étape de cette transformation consiste à rechercher un schéma dont l'entête s'identifie au graphe conceptuel considéré. Nous utilisons pour cela la structure d'indexation (cf. paragraphe implémentation du chapitre 4) qui accompagne la mémoire pragmatique. Cette structure de nature hiérarchique se fonde sur le type du prédicat de l'entête des schémas, sur le type d'un de ses objets et sur celui de la relation qui les unit. Lors de l'indexation d'un schéma, le choix de l'objet est réalisé d'abord par

rapport à sa relation avec le prédicat : cette relation doit être directe et faire partie des relations présentes dans le graphe canonique associé au type du prédicat. En cas de sous-détermination de ce premier critère, on retient l'objet dont le type de concept est le plus spécifique, donc supposément le plus discriminant.

En reprenant ces critères d'indexation, on est donc capable de déterminer l'objet du prédicat à prendre en considération lorsque l'on recherche un schéma à partir d'un graphe supposé représenter son entête. Si l'index permet d'accéder au moins à un schéma, on vérifie que celui-ci s'accorde avec le graphe à l'origine de sa recherche en testant l'égalité du graphe avec l'entête du schéma trouvé. Ce test est réalisé en procédant à une projection de l'un dans l'autre et vice et versa. Il y a égalité si les deux projections réussissent. Dans ce cas, on peut ajouter au corps du schéma en construction une référence vers le schéma trouvé dans la mémoire.

Si la recherche ne ramène aucun candidat intéressant, on crée donc un nouveau schéma composé uniquement d'un entête s'identifiant au graphe à transformer en référence et on l'ajoute à la mémoire pragmatique (ce qui passe par son indexation). Ce préalable ayant été accompli, tout se déroule au niveau du corps du schéma en construction comme dans le cas précédent.

La création d'une référence vers un autre schéma suppose non seulement la donnée d'un graphe et d'un schéma ayant ce graphe comme entête mais implique également la donnée d'une importance et d'une spécificité. Cette dernière est calculée selon la mesure exposée au chapitre 4. L'importance est dérivée quant à elle du poids du graphe au sein de l'UT agrégée source du nouveau schéma. Plus précisément, du fait des regroupements d'événements, il est possible que le poids relatif d'un regroupement, et donc du graphe qui en est la généralisation, dépasse 1,0. Pour ramener tous les poids à des valeurs comprises entre 0 et 1, on choisit de diviser le poids de chaque référence par celui ayant la valeur la plus grande. On aura donc toujours au moins une référence au sein du schéma possédant une importance égale à 1,0.

4.2. Définition de l'entête du schéma

Comme l'a montré le paragraphe précédent, la construction du corps d'un nouveau schéma est un problème auquel on peut apporter une solution satisfaisante dans une grande partie des cas. Il n'en est pas de même avec la définition de son entête. Cette tâche consiste en pratique à trouver une caractérisation sémantique de la situation représentée par le schéma. Il s'agit de faire le lien entre les connaissances pragmatiques et les connaissances sémantiques. Le seul moyen de mettre à jour ce lien est encore une fois de

s'appuyer sur le contenu de la mémoire épisodique. La relation de déviation thématique entre deux UTs agrégées traduit en particulier le fait qu'un événement de l'une des UTs est développé par l'autre UT. Cette définition s'identifie tout à fait au problème posé : la relation unit en effet la représentation sémantique d'un événement, i.e. un graphe dans l'UT source de la déviation, à une caractérisation pragmatique possible de ce même événement, i.e. l'UT agrégée cible de la déviation.

On déduit immédiatement de cette constatation une façon de déterminer l'entête d'un nouveau schéma. Il suffit en effet de trouver une relation de type déviation thématique ayant comme cible l'UT agrégée que l'on abstrait et d'adopter comme entête le graphe qui est à la source de cette déviation. Bien entendu, comme l'UT qui le contient n'a pas forcément fait l'objet d'une abstraction préalable, il faut transformer le graphe agrégé visé en un graphe conceptuel 'normal'. Il suffit pour cela de ne retenir que la partie des concepts agrégés résultant de la généralisation des instances et d'appliquer les principes de suppression des concepts accessoires mis en œuvre lors de la généralisation des regroupements d'événements.

On ne peut néanmoins garantir que toute UT agrégée que l'on abstrait est la cible d'une relation de déviation thématique permettant d'appliquer le principe exposé ci-dessus. Cela est d'autant plus vrai que ces relations sont plus difficiles à mettre en évidence par l'analyse thématique que les relations de changement, naturellement plus marquées. Par ailleurs, il n'est pas certain non plus que la détermination du point de départ de ces relations soit toujours très fiable. Une façon de lever cette incertitude est de s'appuyer sur les poids des relations et des graphes agrégés. Ceux-ci ne sont en revanche d'aucune aide pour trouver une autre solution lorsqu'ils invalident celle examinée initialement.

On en conclut que l'utilisation des relations de déviation thématique n'est possible que dans une minorité de cas, à moins de faire de leur présence un critère d'abstraction. Cette dernière possibilité imposerait toutefois de repousser l'abstraction à un terme très lointain pour la plupart des UTs agrégées.

Ne pas changer les critères d'abstraction oblige donc à trouver une façon de définir l'entête d'un nouveau schéma en l'absence de relation de déviation thématique. La solution retenue consiste à choisir comme entête le graphe possédant le poids le plus fort. Bien entendu, il ne s'agit là que d'une solution de secours dans la mesure elle contrarie la structure habituelle des schémas. En effet, on obtient de cette manière des schémas contenant des références circulaires. Même imparfait, cet entête peut être tout de même utile dans les tâches habituellement dévolues à cette partie des schémas. Ainsi, il lui est tout à fait possible de servir de déclencheur pour le schéma auquel il est associé lors de tâches de type compréhension de texte. Dans la mesure où il représente l'événement le

plus important du schéma, il a en effet toutes les chances d'apparaître dans les textes évoquant la situation concernée.

5. Exemple

Afin d'illustrer l'ensemble du processus d'abstraction, nous avons choisi de l'appliquer à l'UT agrégée de la figure 6.9, obtenue à l'issue du chapitre 6.

1ère étape : critères d'abstraction

Tête de l'UT à la suite de 3 agrégations (T3)

Somme des poids des graphes de l'UT (Sp) = 7,61, d'où un seuil fixé à 3,805

graphes retenus : Poignarder (1,0), Arrêter (1,0), Être_mort (0,66), Être_emprisonné (0,66), Soutenir (0,33), SeQuereller (0,33)

Tête de l'UT à la suite de 4 agrégations (T4)

Sp = 6,75, d'où un seuil fixé à 3,375

graphes retenus : Poignarder (1,0), Arrêter (0,75), Être_mort (0,75), Être_emprisonné (0,5), Menacer (0,25), Croire (0,25)

SimTête(T4,T3) = 0,86

Tête de l'UT à la suite de 5 agrégations (T5)

Sp = 6,8, d'où un seuil fixé à 3,4

graphes retenus : Poignarder (1,0), Arrêter (0,6), Être_mort (0,6), Être_emprisonné (0,4), Habiter (0,4), Être_blessé (0,4)

SimTête(T5,T4) = 0,76

Tête de l'UT à la suite de 6 agrégations (T6)

Sp = 6,37, d'où un seuil fixé à 3,185

graphes retenus : Poignarder (1,0), Arrêter (0,5), Être_mort (0,5), Être_blessé (0,5), Habiter (0,33), Être_emprisonné (0,33), Pénétrer (0,33)

SimTête(T6,T5) = 0,99

On constate sur ses trois dernières agrégations que cette UT agrégée peut être considérée comme suffisamment stable. La similarité entre sa tête à l'état Tn+1 et à l'état Tn reste en effet toujours au-dessus du seuil fixé (égal à 0,7). Le processus d'abstraction peut donc être lancé pour créer un nouveau schéma.

2ème étape : regroupement des événements

Détermination des graphes émergents

Circonstances : $S_{emg} = 1,11 + 0,20 \approx 1,31$

graphes émergents : Habiter

Description : $S_{emg} = 1,72 + 1,06 \approx 2,78$

graphes émergents : Poignarder, Arrêter

États Incidents : $S_{emg} = 2 + 0,8 = 2,8$

graphes émergents : Être_mort, Être_blessé

Regroupement des prédicats

1ère phase : regroupement des prédicats des graphes non émergents

aucun regroupement possible, quel que soit l'attribut considéré.

2ème phase : regroupement des prédicats des graphes non émergents et des prédicats des graphes émergents

Description : selon la hiérarchie des prédicats verbaux présentée au chapitre 4, *Poignarder* est un sous-type direct de *Frapper*. On peut donc regrouper les deux prédicats et les remplacer par le type de concept *Frapper*, doté d'un poids absolu de 7.

États Incidents : *Être_guillotiné* est un sous-type de *Être_exécuté*, lui-même sous-type de *Être_mort*. Les prédicats *Être_mort* et *Être_guillotiné* peuvent donc être regroupés pour donner le type de concept *Être_mort*, doté d'un poids absolu de 4.

3ème phase : regroupement de tous les prédicats issus de la 2ème phase

aucun autre regroupement possible.

3ème étape : sélection des événements

Seuil d'émergence

le regroupement des événements n'a pas fait apparaître de nouveaux événements par rapport à la situation initiale.

Suppression des regroupements globalement trop peu représentatifs

moyenne des minima = (0,33 + 0,5 + 0,5) / 3 = 0,44

On supprime donc de l'attribut *Circonstances* le regroupement ne contenant que le graphe *Habiter* .

Ajout de regroupements globalement représentatifs

poids relatif minimum des regroupements sélectionnés : 0,5

Pas d'ajout de regroupement supplémentaire.

4ème étape : généralisation des regroupements d'événements

Le regroupement ne contenant que le graphe *Arrêter* peut être généralisé directement : il n'y a ni jointure de graphes à effectuer, ni suppression de concepts. Il suffit de reprendre les types de concept généralisés du graphe agrégé.

Regroupement ayant comme prédicat *Frapper*

Les relations *objet* et *partie_de* ne font pas partie du graphe canonique direct de *Frapper*. Pour être conservées, elles doivent donc comporter au moins 4 occurrences (le poids du regroupement est de 7 et le seuil, fixé à 0,5). Leur poids

absolu n'étant que de 3, elles sont supprimées. Les types de l'*agent* et du *patient,* les relations communes aux deux graphes regroupés, sont par ailleurs généralisés tous les deux pour donner le type *Homme*.

Regroupement ayant comme prédicat *Être_mort*

La seule manifestation de la fusion des deux graphes est la généralisation du type de concept de la relation *source* en *Humain.*

Regroupement ayant comme prédicat *Être_blessé*

Ce regroupement ne comporte qu'un seul graphe mais celui-ci possède une relation, la relation *manière,* ne faisant pas partie du graphe canonique direct du prédicat. Cette relation n'étant présente que dans une seule occurrence du graphe sur le total des trois occurrences rassemblées, elle est par conséquent supprimée.

5ème étape : construction du schéma

À l'issue de cette étape, on obtient le schéma de la figure 7.3. Les quatre graphes construits précédemment ont été transformés en références vers d'autres schémas, eux-mêmes éventuellement construits à cette occasion. La spécificité associée à chaque référence n'apparaît pas dans la mesure où elle dépend du contenu global de la mémoire pragmatique, non défini dans le cas présent. Le schéma obtenu possède également un attribut *Circonstances* vide que nous n'avons pas fait figurer ici.

Schéma FrapperÊtre_mort

spécialisationDe: Schéma

Entête:
[Frapper; prédicat: vrai]
{→ (agent) → [Homme: *x1],
→ (patient) → [Homme: *x2]
→ (instrument) → [Arme_blanche: *x3]
};

Attribut Description

Schéma Frapper
importance: 1.0;
[Frapper; prédicat: vrai]
{→ (agent) → [Homme: *x1],
→ (patient) → [Homme: *x2]
→ (instrument) → [Arme_blanche: *x3]
}.

Schéma Arrêter
importance: 0.43;
[Arrêter; prédicat: vrai]
{→ (agent) → [Policier: *x4],
→ (patient) → [Homme: *x5]
}.

FinAttribut Description

Attribut ÉtatsIncidents

Schéma Être_mort
importance: 0.57;
[Être_mort; prédicat: vrai] → (source) →
[Humain: *x6].

Schéma Être_blessé
importance: 0.43;
[Être_blessé; prédicat: vrai] → (source) →
[Humain: *x7].

FinAttribut ÉtatsIncidents

FinSchéma FrapperÊtre_mort

Fig. 7.3 - Schéma abstrait à partir de l'UT agrégée de la figure 6.9

De même, nous n'avons pas fait figurer les rôles du schéma du fait de leur aspect purement formel dans le cadre de généralisation actuel. Un nouveau rôle est en effet construit automatiquement pour chaque concept d'un graphe transformé en référence

vers un schéma. C'est pourquoi chaque concept d'un entête de schéma associé à une référence possède une variable qui lui est spécifique. La fonction intrinsèque des rôles qui est de rendre compte des liens d'identité entre les acteurs de différents événements n'est donc pas assumée. Leurs propriétés sont définies dans le même esprit de façon minimaliste : l'importance d'un rôle est égale à l'importance de l'événement dont ils font partie et son type par défaut s'identifie à son type.

En ce qui concerne l'entête du schéma, le graphe de plus fort poids, en l'occurrence le graphe *Frapper*, a été retenu en l'absence de relation de déviation thématique ayant comme cible l'UT agrégée à abstraire. Le nom du schéma est quant à lui construit en prenant la concaténation des deux prédicats de plus grande importance. Enfin, aucun mécanisme de hiérarchisation des schémas n'ayant été spécifié, les schémas nouvellement créés sont rattachés sous le sommet de la hiérarchie des schémas.

6. Limites

La dimension de la compréhension de texte mise en avant dans le cadre d'ANTHAPSI est l'analyse thématique. C'est donc avec le souci de contribuer en priorité à cette analyse thématique que les mécanismes d'apprentissage ont été élaborés. De ce point de vue, la capacité la plus importante que doivent supporter les connaissances apprises est celle de reconnaître que deux événements appartiennent à la même situation. Selon cette optique, un schéma est avant tout considéré comme un regroupement d'événements intervenant dans une même situation. Sur ce plan, les résultats du mécanisme d'abstraction des schémas tels qu'ils apparaissent au travers d'exemples comme celui développé au §5 répondent dans leurs grandes lignes aux attentes fixées.

En revanche, si l'on souhaite représenter au travers des schémas des connaissances suffisamment générales et précises pour intervenir dans des mécanismes plus pointus de compréhension de texte, l'abstraction de schémas présentée ici comporte certaines insuffisances auxquelles il faudrait remédier. L'essentiel de ces insuffisances réside dans l'inexistence d'un mécanisme permettant de construire les rôles d'un schéma sur la base de leur véritable signification et pas seulement dans le but de garantir la bonne formation de ce schéma. Autrement dit, il faudrait être capable dans l'exemple ci-dessus de déterminer que l'*agent* de *Frapper* est le *patient* de *Arrêter* et occupe le cas *source* de *Être_mort* et *Être_blessé*.

La réalisation de ces distinctions ne doit pas seulement intervenir à l'issue de la création des schémas mais doit constituer une préoccupation tout au long du processus d'abstraction d'une UT agrégée et même au delà, lors de la formation de cette dernière.

Pour illustrer cette nécessité, il suffit d'examiner plus attentivement le cas de la référence vers le schéma *Être_mort*. Cette référence est le fruit du regroupement de deux graphes de l'UT agrégée de la figure 6.9 : les graphes *Être_mort* et *Être_guillotiné*. Sur le plan des connaissances sémantiques, il est pertinent de regrouper ces deux graphes puisqu'ils contiennent tous les deux une même information, à savoir la mort de l'individu auquel s'applique le prédicat. En revanche, sur le plan de la compréhension fine des tenants et aboutissants de la situation, il s'agit d'une aberration puisque le graphe *Être_mort* s'applique à la victime de l'attentat tandis que le graphe *Être_guillotiné* concerne l'auteur de l'attentat. En fusionnant ces deux graphes, on perd donc une distinction très importante, non pas sur le plan thématique, mais pour l'utilisation du schéma dans le cadre de la compréhension causale des textes.

Le même problème peut surgir en amont lorsque des UTs textuelles que l'on agrège comportent plusieurs graphes (en pratique, rarement plus de deux) dotés d'un même prédicat. Le critère retenu au chapitre 6 pour déterminer quel graphe s'agrège avec quel autre graphe dans un tel cas repose en effet sur la force de la similarité et non sur la similarité des rôles vis-à-vis de la situation.

Dans un cas comme dans l'autre, le fait de ne pas tenir compte de cette notion de rôle provient des difficultés qu'elle soulève. Au niveau de la mémoire épisodique, il est assez évident qu'intégrer les rôles dans la similarité entre UTs bloquerait probablement la plupart des agrégations en intégrant des contraintes formelles trop strictes compte tenu de l'absence de marge de manœuvre impliquée par l'absence de connaissances a priori sur le domaine. Le nombre important de rôles agrégés obtenus pour une seule UT agrégée en dépit d'un seuil de similarité entre rôles assez bas en est une illustration assez claire.

Le fait d'évacuer cette difficulté au niveau de la mémoire épisodique la reporte naturellement au niveau de la mémoire pragmatique. Il devient en effet difficile de dégager les relations entre les concepts de différents graphes dès lors qu'une UT agrégée amalgame des UTs textuelles qui ne sont pas nécessairement homogènes de ce point de vue. Il est bien entendu envisageable de revenir a posteriori au niveau des épisodes textuels pour retrouver ces relations, ainsi que le permet la mémoire épisodique, mais il n'est pas forcément évident de pouvoir retrouver des tendances suffisamment marquées dès que l'on met en jeu plus de deux graphes. Les rôles agrégés constituent une autre forme d'aide pour aller dans ce sens en mettant au moins en évidence quelques relations stables entre des concepts de différents graphes.

L'exploitation des relations, notamment causales, entre événements est une autre source d'informations pour construire les rôles. Ces relations peuvent être particulières et issues de l'UT agrégée abstraite ou bien être plus générales et provenir d'autres types d'apprentissage, comme celui mis en œuvre dans [Pazzani 1991]. Le premier cas est

également l'occasion de faire apparaître des relations causales ou temporelles entre les références formant le corps du schéma, problème que nous n'avons pas abordé dans le processus de généralisation présenté. L'existence de regroupements entre événements pose en particulier des questions sur le devenir des relations qui y sont attachées.

Tous les problèmes liés à l'abstraction des schémas ne se cantonnent pas qu'au strict domaine des rôles ainsi qu'on peut déjà le voir avec les relations intra-UTs. Le regroupement des prédicats, comme celui réalisé ci-dessus entre *Frapper* et *Poignarder*, renvoie ainsi aux interrogations portant sur l'opportunité et l'ampleur des généralisations. Connaissant la situation représentée, on sait dans ce cas précis qu'une telle généralisation n'est pas nécessaire. *Frapper* est en l'occurrence sans doute un peu trop général. Toute la difficulté consiste à trouver des critères pour appuyer cet avis en dehors des connaissances sur la situation que l'on cherche à apprendre. Dans le cas présent, on peut remarquer que le poids des deux prédicats est assez différent (6 pour le poids absolu de *Poignarder* et 1 pour celui de *Frapper*), facteur que l'on ne prend pas en compte pour le moment dans la généralisation des prédicats. Il semble en fait raisonnable de considérer qu'un prédicat est d'autant moins généralisable que son poids est important, ce dernier servant d'indicateur de pertinence et de typicalité par rapport à la situation. Rien ne garantit néanmoins que ce facteur aille systématiquement dans le sens souhaité.

Généraliser implique de regrouper ce qui est similaire mais également de dissocier ce qui est incompatible. Le regroupement des prédicats devrait donc posséder un pendant chargé de détecter les événements qui ne sont pas compatibles. Dans le cas d'une tentative d'assassinat par exemple, il n'est pas possible de regrouper dans le même schéma le cas dans lequel la victime meurt et celui où elle survit. Cette tâche renvoie plus généralement au problème de l'apprentissage de concepts disjonctifs. Les schémas ne représentant que des conjonctions d'événements, le seul moyen de rendre compte de disjonctions est de créer plusieurs schémas frères, chacun d'entre eux rendant compte d'une des possibilités.

La question la plus importante reste cependant celle de la détection de ces incompatibilités. Comme dans le cas des rôles, la principale possibilité consiste à revenir au niveau des différents épisodes ayant contribué à forger l'UT agrégée considérée. Il faut repérer les événements ou les configurations d'événements ne survenant jamais ensemble. La tâche est rendue d'autant plus ardue qu'elle ne peut être dissociée de celle de regroupement des prédicats. Il ne faut pas en effet regrouper des événements incompatibles et à l'inverse, il ne faut pas confondre deux prédicats très proches avec deux événements incompatibles. Deux prédicats proches dans une UT agrégée font souvent référence à un même événement exprimé de façon différente en fonction des épisodes. Il est alors normal qu'ils n'apparaissent jamais ensemble car le choix d'un des modes d'expression est généralement exclusif de l'autre. Plus globalement, les relations

intra-UTs constituent comme pour les rôles une aide potentielle à exploiter. Des événements liés par des relations causales ou temporelles ne sont pas en effet des événements incompatibles.

La dernière limitation dont nous ferons état ici concerne la structuration hiérarchique des schémas créés. Selon la procédure actuelle, un schéma nouvellement créé, que l'on appellera S_{nouv} dans ce qui suit, est invariablement raccroché au sommet de la hiérarchie. Une première façon d'aller plus loin peut être réalisée assez simplement. Il suffit en effet de trouver dans cette hiérarchie un schéma possédant un entête dont le prédicat est un sur-type du prédicat de l'entête de S_{nouv}. Ce sur-type doit bien sûr être le plus spécifique possible. On utilise pour cela l'index associé à la mémoire épisodique. Lorsqu'un candidat S_{st} a été trouvé, on projette son entête dans celui de S_{nouv}. Si la projection réussit, on peut rattacher S_{nouv} sous S_{st}. Cela suppose également de rattacher les éventuels fils de S_{st} sous S_{nouv}. La situation est plus complexe en cas d'échec de la projection. Il faut en particulier déterminer si le blocage ne provient pas de concepts accessoires ne faisant pas partie du graphe canonique associé au prédicat de l'entête et que l'on pourrait laisser de côté. Mais nous ne pousserons pas l'analyse plus loin dans le cadre de ce travail.

7. Implémentation

Les chapitres précédents nous ont montré que les pré-requis nécessaires à la mise en œuvre du mécanisme d'abstraction présenté, en l'occurrence la mémoire épisodique et la mémoire pragmatique, ont été implémentés sans restriction par rapport aux spécifications décrites. Le mécanisme d'abstraction en lui-même n'a pas fait l'objet d'une mise en œuvre similaire, tout du moins dans sa totalité. La détermination des graphes émergents et la construction finale du schéma sont en effet des éléments communs avec le travail de Gaël de Chalendar, travail que celui-ci a implémenté dans le cadre de son stage de DEA. L'abstraction des schémas n'a donc été testé que manuellement. Néanmoins, le degré de spécification de la procédure générale et des algorithmes plus spécifiques a été poussé suffisamment loin pour que ne subsiste plus d'ambiguïté. Par ailleurs, compte tenu du faible nombre d'UTs agrégées inhérent à l'ampleur du travail de modélisation manuelle requis pour leur constitution, le test manuel n'a pas été un obstacle à la validation de l'algorithme sur les données disponibles.

Récapitulatif

Dans ce chapitre, nous avons abordé le problème de l'abstraction des connaissances contenues dans la mémoire épisodique visant à la formation de nouveaux schémas au sein de la mémoire pragmatique. Plus précisément, nous nous sommes limité à l'abstraction d'une UT agrégée en un schéma. Cette abstraction se décompose en cinq étapes. La première en forme l'amorce puisqu'elle a pour objectif de déterminer quand l'abstraction d'une UT agrégée doit avoir lieu. Le critère retenu pour supporter cette décision repose sur l'analyse de l'évolution de la stabilité de l'ensemble formé par les graphes les plus significatifs de l'UT agrégée. Lorsque cet ensemble est suffisamment stable sur une période donnée, l'abstraction peut être déclenchée.

Les trois étapes suivantes sont étroitement liées et assurent la sélection et la généralisation des événements de l'UT agrégée à abstraire qui figureront dans le schéma nouvellement créé. La première de ces étapes est en pratique la plus importante et la plus complexe de l'ensemble du processus d'abstraction dans sa forme actuelle. Elle consiste à regrouper des événements de même nature mais présents comme autant de graphes distincts au niveau des UTs agrégées du fait de la non identité des types de leurs prédicats. Ce regroupement s'effectue donc par une généralisation contrôlée des types des prédicats des graphes de l'UT agrégée concernée. Le contrôle de la généralisation est exercé par l'affectation à chaque prédicat d'un capital de généralisation limité et l'association d'un coût à chaque opération de généralisation. L'application d'une stratégie spécifique d'exploration des différentes possibilités de généralisation permet de privilégier les regroupements entre événements maximisant le nombre d'événements susceptibles d'être sélectionnés pour former le nouveau schéma.

L'étape suivante a pour but de procéder à cette sélection des événements de l'UT agrégée à abstraire que l'on considère comme suffisamment représentatifs pour constituer le corps du schéma nouvellement créé. Cette sélection s'opère uniquement en s'appuyant sur les poids des événements. Schématiquement, elle consiste à ne retenir que le événements dont le poids dépasse un seuil déterminé par la somme de la moyenne de ces poids ajoutée à leur écart-type moyen.

Le dernier volet du triptyque annoncé est la fusion des graphes formant les différents regroupements de façon à ne représenter un événement que par un seul graphe. Cette fusion s'accompagne de leur généralisation. Celle-ci se traduit à la fois par une généralisation des concepts intervenant de façon récurrente dans les différents graphes en occupant le même rôle et une suppression de ceux dont la présence n'est qu'épisodique et le rôle, accessoire. L'étape s'appuie pour l'essentiel, plus précisément pour la fusion des

graphes et la généralisation des concepts, sur une adaptation de l'opération d'agrégation utilisée dans le cadre de la constitution de la mémoire épisodique.

La dernière étape réalise le changement de représentation entre UT agrégée et schéma en construisant un schéma à partir des événements sélectionnés et généralisés d'une UT agrégée. À l'occasion de cette construction, les graphes conceptuels représentant ces événements sont transformés en références vers les schémas développant ces événements sur le plan pragmatique. Au besoin, de nouveaux schémas, dans un premier temps terminaux, sont créés pour faire exister les situations correspondantes au niveau de la mémoire pragmatique. La construction d'un schéma se traduit également par la définition d'un entête, chargé de faire le lien entre l'évocation de la situation représentée par le schéma au niveau sémantique et ce dernier.

Année 1998

UNIVERSITE DE PARIS-SUD
U.F.R. SCIENTIFIQUE D'ORSAY

THÈSE

(Volume 2)

ANTHAPSI : un système d'analyse thématique et d'apprentissage de connaissances pragmatiques fondé sur l'amorçage

Olivier FERRET

MM.	Brigitte GRAU	Examinateur
	Daniel KAYSER	Examinateur
	Yves KODRATOFF	Examinateur
	Maria Teresa PAZIENZA	Rapporteur
	Gérard SABAH	Directeur
	Pierre ZWEIGENBAUM	Rapporteur

Chapitre 8

L'analyse thématique de MLK

Les représentations de texte étant au cœur de MLK, il est nécessaire d'accorder un intérêt tout particulier au processus assurant leur construction. C'est l'objet du présent chapitre. Compte tenu de la nature des représentations de texte, il apparaît que leur élaboration repose principalement sur une analyse thématique des textes. Celle-ci réutilise principalement les connaissances contenues dans la mémoire épisodique de MLK. Elle met en œuvre dans ce but une forme particulière de raisonnement à base de cas. Elle utilise en effet non pas les représentations de texte précédemment accumulées en tant que telles mais plutôt le résultat de leur agrégation. Cette utilisation passe par le processus de sélection de connaissances présenté au chapitre 6. Le mécanisme d'analyse considéré permet de réaliser une segmentation thématique des textes ne faisant pas d'hypothèse a priori sur la structuration du discours. Il lui est ainsi possible de traiter des textes au style dit très entrelacé. Les autres dimensions de la construction des représentations de texte sont également abordées dans ce chapitre mais dans une moindre mesure, certaines ayant déjà été évoquées au chapitre 5.

1. Le problème de l'analyse thématique

1.1. La notion de thème

1.1.1. Le point de vue de la linguistique et de l'analyse du discours

La notion de thème est caractérisée par une forme de paradoxe. Elle est à la fois très intuitive et assez difficile à cerner. Tout le monde est capable d'en donner une définition informelle : le thème d'un texte, d'une conversation ou d'une partie de l'un ou de l'autre représente ce dont 'parle' l'unité textuelle considérée; il en est le sujet. En revanche, les tentatives pour formaliser plus avant cette notion n'ont pas abouti pour le moment à la définition d'un cadre unifié de description. Cette difficulté n'est sans doute pas étrangère au caractère trop intuitif de la notion qui laisse penser qu'elle n'est sans doute pas le moyen de description adéquat du phénomène que l'on tente de capturer.

Dans [Brown & Yule 1983], Brown et Yule mettent ainsi en avant que la notion de thème est clairement une façon intuitive et commode de décrire dans le domaine de l'analyse du discours un principe unificateur plus profond rendant compte du fait qu'une unité de discours 'parle' de quelque chose et que l'unité qui la suit 'parle' d'une autre

chose. Ils préfèrent d'ailleurs définir la notion de thème de façon opératoire plutôt que d'essayer d'en cerner une caractérisation de nature descriptive : dans le cadre d'un discours, un thème est selon leur conception attaché à une unité de discours délimitée par deux changements de thème.

Dans [Todorov 1972], Todorov souligne dans le même esprit la pauvreté de l'appareil conceptuel existant pour aborder cette notion. Les études linguistiques qui ont été faites à son sujet se sont principalement attachées à deux unités de discours situées aux deux extrémités du spectre des unités possibles, la proposition et la phrase d'un côté et le texte de l'autre, sans parvenir à faire émerger de véritable niveau intermédiaire. En ce qui concerne les propositions et les phrases, on trouve la distinction originellement opérée par les linguistes du Cercle de Prague entre thème et rhème. Le thème représente dans ce cas ce dont parle le locuteur tandis que le rhème recouvre l'information apportée à propos de ce thème.

Nous nous situons là cependant à un niveau trop élémentaire par rapport à ce que recouvrent nos Unités Thématiques puisque la distinction thème/rhème se situe au niveau de la détermination du prédicat d'une proposition. [Martin 1992] présente les travaux de Fries sur la généralisation de cette notion de thème à une échelle plus large en introduisant des hyper-thèmes, conditionnant les thèmes et se plaçant eux-mêmes sous la dépendance de macro-thèmes. On retrouve à cet égard un peu la séparation entre super-structure, macro-structure et micro-structure opérée par Kintsch et van Dijk [Kintsch & Dijk 1978]. Néanmoins, si la distinction thème/rhème s'appuie sur un certain nombre de critères objectivables, cela semble moins évident de la mise en évidence des hyper-thèmes et des macro-thèmes.

À l'autre extrémité de l'échelle des unités textuelles, des travaux dans le prolongement de ceux de Greimas [Greimas 1966], comme ceux de Rastier [Rastier 1989], ont cherché à rendre compte de la notion de thème dans l'ensemble d'un texte, d'un ouvrage et même d'une œuvre entière en s'appuyant sur le concept d'isotopie, c'est-à-dire sur la récurrence de regroupements d'éléments sémantiques élémentaires, appelés sèmes. L'avantage de cette approche réside dans sa facilité d'application à des unités discursives de tailles différentes. Rastier la qualifie d'ailleurs de "Sémantique Descriptive Unifiée". Son inconvénient, qui est un corollaire inhérent à son intérêt, est de ne pas définir un éventail d'unités discursives de différents niveaux, en particulier sur le plan thématique.

Ce manque se fait d'autant plus sentir que les travaux réalisés en analyse du discours sur la définition d'unités discursives ne se sont généralement pas placés sur ce plan. Des formalistes russes tels que Propp [Propp 1970] jusqu'à la sémiotique narrative de Greimas [Greimas 1970, Greimas 1966] et aux grammaires de récit [Rumelhart 1977],

les critères présidant à la définition des unités discursives ont été plutôt de nature fonctionnelle que de nature thématique. Même en sortant de ce cadre, surtout appliqué aux textes narratifs, et en s'orientant vers des travaux plus proches des textes eux-mêmes, s'appuyant sur la détection de marques textuelles plutôt que sur des hypothèses de structuration a priori, il est difficile de trouver une prise en compte de la dimension thématique. Dans [Charolles 1993], Charolles précise ainsi que "la notion de thème est extrêmement difficile à préciser" et qu'elle "met en jeu toutes sortes d'habiletés linguistiques et non linguistiques qui sont très globales et *apriori* indécidables". Quant à s'intéresser à ce que les textes évoquent, les entités qu'ils introduisent et les chaînes référentielles dans lesquelles elles sont impliquées lui apparaissent à cet égard plus exploitables sur le plan de l'analyse du discours.

1.1.2. Le point de vue du traitement automatique des langues

Le peu de succès apparent de la notion de thème en linguistique n'est sans doute pas étranger au fait qu'en l'abordant, on entre dans le domaine de ce qui est exprimé au travers de l'usage de la langue mais que l'on n'est plus par là même dans le domaine de la langue elle-même. On sort donc de la sphère de la linguistique.

Le traitement automatique des langues n'est pas confronté à ce problème. Il peut se contenter de définir la notion de thème par rapport à des connaissances de référence. Dans ce cadre, un thème s'identifie à un regroupement de connaissances sur le monde se rapportant à un même objet. Ces connaissances recouvrent donc un ensemble d'actions, d'états, de personnages (au sens large puisqu'il s'agit aussi bien d'objets, de personnes que de lieux) et de relations de différents types (causales, temporelles, ...) liant ces actions et ces états. Cependant, on conserve à ce niveau une difficulté de définition puisque l'objet reste toujours à cerner. Dans le cas d'ANTHAPSI, il correspond à une situation prototypique du monde qui nous entoure, avec tout le flou que nous avons souligné au chapitre 1 inhérent à cette notion.

Cette notion de situation définit implicitement un certain degré de généralité. Si un thème rassemble tout ce qui a trait au domaine médical, il se situera clairement à une granularité supérieure à la situation. À l'inverse, s'il se limite au contenu d'une recette de cuisine, il sera vu comme un simple événement du point de vue d'une situation. Ces exemples révèlent que l'étendue possible de l'objet d'un thème est en pratique très large. Celui-ci dépend à la fois du texte traité et du but dans lequel on l'analyse. L'imprécision de la définition d'une situation montre par ailleurs que cerner cet objet ne peut être fait que de façon approximative. La meilleure garantie de ne pas tout mélanger tout en adoptant un

cadre de traitement homogène passe donc par une hiérarchisation des thèmes : à défaut de définir exactement les caractéristiques de leur contenu, on les organise relativement les uns aux autres en fonction du niveau de généralité de ce contenu.

Les thèmes manipulés dans le cadre du traitement automatique des langues peuvent également être très diversifiés quant à leur degré de structuration interne. Lorsque le domaine couvert est restreint, la représentation d'un thème peut être élaborée manuellement et donc, être très structurée, à la manière des schémas composant la mémoire pragmatique par exemple. En revanche, lorsqu'il s'agit de représenter un grand nombre de thèmes, comme en Recherche d'Informations, leur contenu est beaucoup plus frustre et moins précis[1]. Même en conservant une intervention essentiellement manuelle, il n'est plus possible d'avoir le même degré d'élaboration et l'on s'oriente alors vers des formes de connaissances telles que le thesaurus. On a aussi souvent recours à l'apprentissage automatique et en particulier aux techniques de regroupement conceptuel. Les thèmes formés se présentent alors comme des regroupements de mots, éventuellement pondérés en fonction de leur importance.

Quelle que soit la façon dont elles sont représentées, ces connaissances forment la référence permettant de reconnaître les thèmes des textes et de délimiter leur extension au sein de ceux-ci. L'analyse thématique, qui assure cette tâche, s'apparente à une forme de désambiguïsation. Chaque élément contribuant à la représentation d'un thème n'est pas en effet nécessairement spécifique de ce thème. Il peut également figurer dans les représentations d'autres thèmes : l'originalité d'un thème provient de la présence d'une configuration d'éléments et pas seulement de la présence de quelques uns, très spécifiques. Chaque unité textuelle considérée, il peut s'agir d'un mot plein ou d'un ensemble de mots comme une proposition, d'un concept ou de la représentation sémantique d'une proposition dans les cas les plus avancés, renvoie à donc à un ensemble de thèmes parmi lesquels un choix est réalisé en fonction du contexte courant, c'est-à-dire des unités qui ont déjà été appréhendées précédemment. La délimitation de l'extension d'un thème dans un texte s'effectue simplement en rassemblant toutes les unités du texte qui ont été rattachées à ce thème.

1.2. Définition de l'analyse thématique

Fondamentalement, l'analyse thématique répond aux objectifs définis ci-dessus à propos de la présentation de la notion de thème en traitement automatique des langues :

[1] En particulier, la frontière entre connaissances de nature pragmatique et connaissances de nature sémantique devient très floue et la représentation des thèmes rassemble dans la pratique les deux.

déterminer de quoi parle un texte, donc quels thèmes il aborde, et, dans la mesure où l'on souhaite pousser l'analyse un peu plus loin, reporter cette question à un niveau de granularité plus faible que l'unité discursive, ici le texte, que l'on considère.

L'étude des systèmes s'inscrivant dans le cadre de cet objectif général conduit à raffiner ce dernier en faisant apparaître trois sous-problèmes bien identifiés :

- la segmentation thématique;
- le suivi thématique;
- l'identification thématique.

La segmentation thématique consiste à découper les textes[1] en unités thématiquement homogènes. Bien que beaucoup de méthodes spécifiquement dédiées à ce sous-problème procèdent à un découpage des textes en blocs adjacents, il ne s'agit là que d'une façon assez élémentaire de traiter la question, généralement due à la faiblesse des moyens (tout spécialement les connaissances disponibles) mobilisés pour le résoudre.

Deux axes d'approfondissement se dessinent assez nettement. L'un d'entre eux réside dans la possibilité de définir des segments thématiquement homogènes dont le contenu ne respecte pas forcément la linéarité des textes. En pratique, un tel segment peut être formé d'un bloc de trois phrases sélectionnées à un endroit du texte considéré, d'un bloc de quatre phrases situé cinq paragraphes plus loin et enfin, d'un bloc de sept phrases éloigné de deux paragraphes du précédent. Cette discontinuité potentielle d'une même unité thématique permet en particulier de tenir compte de phénomènes tels que les interruptions ou les digressions.

Le second axe d'approfondissement envisageable pour ce sous-problème est la possibilité de définir des segments de différents niveaux et de les hiérarchiser. On obtient de cette manière un emboîtement de segments parallèle à une hiérarchie des thèmes abordés. Cette hiérarchisation permet de traiter de manière homogène des textes de longueurs différentes et d'adapter une même analyse à des besoins applicatifs divers.

En Recherche d'Informations, la tâche essentielle dévolue à une analyse thématique est de mettre en évidence les grands thèmes les plus représentatifs d'un document lors de son indexation. On peut donc se contenter des segments de plus haut niveau qu'elle peut mettre en évidence, correspondant à une segmentation assez grossière. En Extraction d'Informations, cette même analyse est utilisée pour localiser les parties du texte sur lesquelles il est important de se concentrer pour trouver les données précises que l'on recherche. On devra donc au contraire s'appuyer sur les unités thématiques les plus

[1] On parlera ici de texte mais il faut considérer cette appellation comme générique. La même analyse est opérante pour des dialogues par exemple.

élémentaires que la segmentation peut déterminer afin de ne pas passer à côté de certaines informations.

En ce qui le concerne, le suivi thématique consiste à déterminer quelles sont les relations unissant les différents segments délimités par la segmentation. Cette détermination s'effectue entre segments de même granularité. Cette tâche est étroitement dépendante du degré d'élaboration de la segmentation réalisée. Lorsque celle-ci ne produit qu'une série de blocs de texte adjacents, la seule possibilité se résume à faire apparaître une relation de changement de thème entre chaque segment et celui qui le suit. Celle-ci reste d'ailleurs implicite dans de tels cas.

En revanche, si le degré de segmentation est plus fin, il devient envisageable, et même nécessaire, de faire la différence entre ce qui constitue un changement de thème net et ce qui ne relève que d'une déviation de thème, c'est-à-dire d'un approfondissement d'un point particulier d'un thème. Avec une segmentation trop grossière, les déviations de thème ne peuvent être perçues dans la mesure où elles se trouvent fusionnées au thème qu'elles contribuent à détailler à cause de leur proximité avec lui. Ce phénomène est renforcé par le fait que les déviations de thème sont souvent prises en sandwich entre deux parties du développement du thème qu'elles précisent. Le passage d'un segment à un autre est alors effectivement synonyme d'un changement de thème. Avec une segmentation plus fine au contraire, les déviations se différencient par rapport aux thèmes plus importants et un changement de segment n'est plus synonyme d'un changement de thème, d'où la nécessité de définir les relations précises entre segments.

Le dernier sous-problème de l'analyse thématique dégagé ici, celui de l'identification thématique, s'intéresse à la caractérisation des unités thématiques définies par la segmentation. Il est donc lui aussi assez dépendant du résultat de celle-ci. Cette dépendance est néanmoins moins fondamentale que dans le cas précédent. Bien que le résultat de l'identification thématique soit différent pour un même texte en fonction du résultat de la segmentation, cette dernière ne remet pas en cause la possibilité de réaliser la caractérisation des segments d'un texte dans la mesure où celle-ci ne change pas fondamentalement suivant qu'un segment corresponde à un paragraphe ou à un texte entier.

Cette caractérisation peut revêtir des formes assez diverses. Le cas le plus évident est celle de la mise en relation d'un segment avec une représentation de référence du thème qui est sensé être son objet. La tâche d'identification thématique s'identifie alors à une tâche d'étiquetage telle qu'elle est pratiquée au niveau sémantique ou au niveau morpho-syntaxique. En l'absence d'une telle représentation de référence des thèmes, la caractérisation d'un segment est réalisée en se fondant sur ses caractéristiques

intrinsèques ou par rapport aux caractéristiques d'autres segments. Sans prétendre à l'exhaustivité, on peut ainsi mentionner la possibilité de construire de façon automatique un nom à partir de son contenu, ce qui est souvent utile dès qu'il y a interaction avec un utilisateur, ou simplement le fait de détecter si le segment a supposément le même thème qu'un autre segment traité précédemment ou si au contraire, son objet n'a encore jamais été rencontré.

À titre de comparaison, il est intéressant de confronter le découpage de l'analyse thématique exposé ci-dessus aux différentes tâches distinguées dans le cadre du programme d'évaluation Topic Detection and Tracking (TDT) [Wayne 1997] organisé par le Département Américain de la Défense (DARPA) concernant le même sujet. La notion de thème est ici restreinte à celle d'événement, correspondant à un sujet au sens journalistique du terme (l'attentat du RER de la station S^{t} Michel, l'explosion en vol du Boeing de la TWA, ...). Trois tâches sont mises en évidence dans ce cadre :

- Segmentation. Il s'agit de découper un flot continu de texte en segments évoquant chacun un événement unique. Il s'agit de la tâche de base constituant le préalable aux deux suivantes;

- Détection ou Identification. La tâche de détection consiste à repérer un nouvel événement par rapport à ceux déjà rencontrés dans un flot continu de texte. Elle doit être remplacée dans les étapes ultérieures du programme par une tâche d'identification. Celle-ci est chargée de reconnaître que plusieurs segments font référence au même événement au sein d'un flot continu de texte. Cette tâche est réalisée de façon non supervisée : aucune représentation de l'événement concerné n'est en effet donnée a priori;

- Tracking. Contrairement à la tâche d'identification, on fournit au préalable la représentation d'un événement et l'on cherche à extraire, à partir d'un flot continu de texte segmenté, tous les segments relatifs à cet événement. Plus précisément, la tâche inclut également la construction de la représentation de l'événement puisque ne sont donnés que les matériaux nécessaires à la construction de cette représentation, en l'occurrence un certain nombre d'exemples de textes relatifs à l'événement pisté (accompagnés également de contre-exemples).

Dans le contexte de ce programme d'évaluation, on retrouve le problème de la segmentation posé en tant qu'élément de base de l'analyse thématique puisque la tâche chargée de le résoudre est un préalable aux deux autres. La tâche Tracking est quant à elle une application du problème de l'identification thématique. Il s'agit en effet de caractériser chaque segment, non par rapport à un ensemble de thèmes, mais par rapport à un seul. La caractérisation est donc binaire : soit le segment considéré est en relation avec le thème

pisté, soit il ne l'est pas. La partie de la tâche en relation avec la construction de la représentation du thème à pister n'est en revanche pas du ressort de l'analyse thématique de notre point de vue.

Des considérations du même ordre interviennent lorsque l'on s'intéresse aux tâches Détection et Identification. En première analyse, ces deux tâches apparaissent elles aussi comme des applications du problème de l'identification thématique. La tâche Identification est effet identique à la tâche Tracking si l'on ne tient pas compte de la présence préalable ou non d'une représentation du thème pisté et la tâche Détection consiste à étiqueter chaque segment par le caractère de nouveauté de son thème.

En fait, les deux ont un fond commun important fondé sur le mélange étroit entre analyse thématique proprement dite et apprentissage de connaissances sur les thèmes. Dans le cas de la tâche Identification, l'apprentissage porte sur un seul thème, en l'occurrence celui que l'on veut pister. Il est en effet intéressant de bénéficier de l'apport d'information venant de chacun des segments repérés comme relevant de ce thème afin d'améliorer progressivement les performances du processus de pistage sur ce thème. Dans le cas de la tâche Détection, le champ est plus large puisqu'il faut construire de façon incrémentale une représentation de tous les thèmes rencontrés, de façon à déterminer si celui d'un nouveau segment est déjà connu. La problématique se rapproche alors fortement de celle que nous développons dans le cadre d'ANTHAPSI concernant la liaison entre compréhension et apprentissage. Compte tenu des contraintes concernant la nécessité d'aboutir à des systèmes véritablement opérationnels, même au détriment d'une certaine précision, les études menées jusqu'à présent dans le cadre du programme TDT se rapprochent du contenu de SEGCOHLEX et de SEGAPSITH.

1.3. Les travaux concernant l'analyse thématique

1.3.1. Vue d'ensemble

Les travaux portant sur l'analyse thématique se répartissent en première approximation en deux grands courants. On distingue ceux s'appuyant sur un suivi du focus ainsi que sur des connaissances élaborées, résultant d'une modélisation manuelle effectuée pour un domaine très restreint, et ceux, opérant sur des ensembles de textes beaucoup plus larges et ne faisant appel qu'à des connaissances peu structurées et peu précises, voire pas de connaissances du tout. Cette distinction en recoupe une autre, complémentaire: les premiers travaux prennent place au niveau conceptuel tandis que les seconds n'interviennent qu'au niveau lexical. Les différences de moyens se traduisentévidemment

par des différences de résultat. Les systèmes relevant du premier courant accordent une importance toute particulière au suivi des thèmes et à la façon dont ils s'enchaînent. Cette capacité leur confère également celle de segmenter les textes mais une telle segmentation n'apparaît pas nécessairement de façon explicite. Ce niveau de précision est en revanche inaccessible aux systèmes relevant du second courant pour lesquels la tâche exclusive est celle de segmentation thématique des textes.

Les travaux détaillés respectivement dans [Grosz & Sidner 1986] et [Grau 1983] sont représentatifs du premier courant. Seul [Grau 1983] est véritablement spécifique de l'analyse thématique. [Grosz & Sidner 1986] s'inscrit plus généralement dans le cadre d'une structuration du discours fortement déterminée par les intentions. Le recoupement avec la dimension thématique est néanmoins très fort, notamment pour ce qui est des textes, dans lesquels intention et thème sont souvent étroitement mêlés, tendance qui est moins marquée pour les dialogues.

Le second courant rassemble une plus grande diversité de travaux, reflétant la diversité des moyens utilisés pour segmenter les textes : [Hearst 1997] pour l'utilisation de la récurrence des mots, [Morris & Hirst 1991] pour celle d'un thesaurus, [Kozima 1993] pour l'exploitation d'un dictionnaire sous forme électronique ou encore [Litman & Passonneau 1995], fondé sur le repérage de marques linguistiques.

Dans ce qui suit nous présentons plus en détail les travaux relevant du premier courant dans la mesure où il s'agit du cadre le plus proche de la segmentation thématique développée pour MLK. Les travaux spécifiques du second courant seront approfondis au chapitre 9, à l'occasion de la présentation de SEGCOHLEX.

1.3.2. L'analyse thématique au niveau conceptuel

Grau

Le but du travail présenté dans [Grau 1983] et [Grau 1984] est explicitement de rendre compte de la cohérence des textes. Cette cohérence est supposée reposer sur les connaissances pragmatiques de référence partagées par le rédacteur et le lecteur d'un texte. Celles-ci sont représentées par des schémas proches de ceux décrits au chapitre 4. Chacun d'entre eux incarne un thème. L'objet de l'analyse développée est de déterminer à quel cas se rattache toute nouvelle proposition d'un texte significative sur le plan thématique :

- la proposition participe au développement du thème en cours;

- la proposition marque le début du développement d'un point particulier du thème en cours;

- la proposition marque le début d'un changement de thème.

Le choix entre ces trois possibilités dépend de la relation trouvée au niveau des connaissances de référence entre la proposition et le contexte courant. Pour être précis, on ne considère pas la proposition en tant que telle mais on s'appuie plutôt sur le schéma qu'elle évoque au travers de son entête. Le contexte est lui-même formé d'un ensemble de schémas, représentant les thèmes "accessibles" :

- le thème actif. Il s'agit du thème en cours de développement;

- le thème qui était actif avant le thème actif courant dans le cas où le second est une déviation du premier. On revient au développement d'un thème à la suite d'une parenthèse à propos d'un de ses points particuliers;

- le thème qui était actif lors d'un changement de thème ayant mené au thème actif courant. Ce dernier est alors considéré comme une interruption que l'on clôt et sur laquelle on ne reviendra pas;

- le thème principal. La détermination de celui-ci constitue la première phase de l'analyse. On suppose donc qu'il est explicité au début du texte. Le thème principal est en fait le premier thème rencontré qui se trouve confirmé au moins une fois. Sa spécificité est qu'il peut être réintroduit à tout moment du texte.

Ce contexte est bien entendu mis à jour tout au long de l'analyse. Pour trouver un lien entre une nouvelle proposition et le contexte courant, on essaie de trouver un lien entre le schéma S_{prop} évoqué par la proposition et ceux composant le contexte S_{contxt}. On parcourt pour cela les liens de référence entre schémas ainsi que les liens de hiérarchisation. Si S_{prop} est identique au thème actif, on reste dans le développement du thème actif. Si un chemin est trouvé dans le graphe des schémas entre S_{prop} et les schémas de S_{contxt}, on conclut que la nouvelle proposition marque le début d'une déviation thématique. Si en revanche, aucun lien n'est trouvé, on opte en faveur d'un changement de thème, à condition toutefois que le nouveau thème partage au moins un élément avec le thème actif. Dans le cas contraire, il y incohérence.

Pour matérialiser le suivi des thèmes, une structure spécifique est construite au fur et à mesure de l'analyse. Cette structure est un arbre rassemblant l'ensemble des thèmes explicités ainsi que ceux, dits inférés, ayant permis de faire le lien entre la proposition courante et le contexte lorsque ce lien n'était pas direct. Cet arbre a comme racine le thème principal du texte et chacun de ses nœuds possède comme fils les thèmes qui ont été

développés à partir de ce nœud, soit dans le cadre d'une déviation thématique, soit dans celui d'un changement de thème.

Grosz et Sidner

La proposition faite par Grosz et Sidner dans [Grosz & Sidner 1986] est une théorie générale sur la structuration du discours visant à élargir des théories plus spécifiquement développées pour rendre compte de la structuration des dialogues finalisés. L'originalité de cette théorie réside dans le fait qu'elle réunit de façon cohérente trois composantes essentielles manipulées en analyse du discours : la structure linguistique, la structure intentionnelle et la focalisation de l'attention.

La structure linguistique incarne la manifestation la plus immédiate de l'analyse du discours puisqu'elle spécifie le découpage des textes[1] en segments de discours. Chaque segment rassemble un ensemble de propositions assumant collectivement une fonction particulière vis-à-vis de l'ensemble du discours. Les segments sont organisés de façon hiérarchique. Un segment peut rassembler un ensemble de segments plus petits, et ainsi de suite jusqu'à parvenir aux propositions. La structure obtenue n'est pas un emboîtement strict puisqu'un segment peut être composé de propositions et de segments à un même niveau. Cette structure des textes se manifeste partiellement au travers d'indices linguistiques de diverses sortes : connecteurs, changements d'intonation lorsqu'il s'agit d'énoncés oraux, énoncés performatifs caractérisant explicitement un changement de sujet, changement de temps et d'aspect, ...

La structure intentionnelle explicite quant à elle les buts des différents locuteurs impliqués dans la production d'un discours. Ces buts sont spécifiques de chaque discours. Ils forment un ensemble ouvert que la structure intentionnelle n'a pas pour objectif de contraindre. En revanche, elle rend compte de la façon dont ces buts sont organisés en faisant l'hypothèse que leur mode de structuration est générique. Les buts composant la structure intentionnelle entretiennent ainsi deux types de relations structurelles : les relations de dominance et les relations de précédence. On dit qu'un but B1 domine un but B2 si la réalisation de B2 contribue à la réalisation de B1. Les relations de dominance confèrent à la structure intentionnelle son caractère hiérarchique. Chaque nœud de l'arbre obtenu est occupé par un but et les fils de ce nœud représentent les sous-buts intervenant dans la réalisation de ce but. Les relations de précédence prennent place en ce qui les concerne entre les buts d'un même niveau. Elles rendent compte de l'ordre, qui peut n'être que partiel, dans lequel ces différents buts doivent être satisfaits afin d'aboutir à la réalisation du but qui les domine.

---

[1] Là encore, la notion de texte est à prendre de façon générique et recouvre tout type de discours.

La structure intentionnelle est développée de manière parallèle à la structure linguistique. Chacun de ses buts est associé à un segment de discours et fixe ainsi son rôle vis-à-vis du segment qui le contient. C'est pourquoi on les dénomme plus spécifiquement *Buts de Segment de Discours* (BSD). Un but global est également défini pour l'ensemble du discours.

La troisième et dernière composante du modèle est de nature différente des deux précédentes. Il ne s'agit pas en effet de construire une structure rendant compte d'un aspect particulier d'un discours mais de définir en tout point de ce dernier sur quels éléments[1] l'attention est centrée. Cet ensemble d'éléments est appelé espace de focalisation. La focalisation de l'attention est le processus permettant d'assurer de façon coordonnée la construction de la structure linguistique et celle de la structure intentionnelle. Un espace de focalisation est associé à chaque segment de discours. Il contient également le BSD lié à ce segment. Les espaces de focalisation permettent de différencier le degré de saillance des objets du discours et des buts des locuteurs sur la base des relations existant entre les segments auxquels ils sont adjoints. Au-delà, ils contrôlent la visibilité de ces objets et des BSDs entre les segments.

Les espaces de focalisation sont utilisés à la fois dans la détermination des BSDs de la structure intentionnelle et dans le traitement de phénomènes liés à la structure linguistique, comme celui des anaphores par exemple. Ils restreignent le champ des possibilités à explorer et introduisent un ordre de préférence parmi celles-ci. Dans le cas des anaphores, ils permettent ainsi d'éliminer un certain nombre de référents potentiels et d'ordonner au moins partiellement les candidats restants.

La visibilité entre espaces de focalisation est régie par l'intermédiaire du mécanisme de focalisation de l'attention. Celui-ci est fondé sur l'utilisation d'une pile, appelée pile de focalisation, afin de prendre en compte le caractère arborescent des structures linguistique et intentionnelle. Le principe général est le suivant : chaque fois que le début d'un nouveau segment est détecté, l'espace de focalisation qui lui est associé est empilé; chaque fois que la fin d'un segment est décelée, son espace de focalisation est dépilé. Au sein d'un segment, il est possible d'accéder aux constituants de l'espace de focalisation lié à ce segment, appelé Ef, mais également aux éléments des espaces de focalisation situés au-dessous de Ef dans la pile de focalisation. Le degré de saillance décroît néanmoins à mesure que l'on s'enfonce dans cette pile, sachant que les éléments les plus saillants sont ceux du propre espace de focalisation.

[1] Ces éléments représentent les objets du discours, au sens large, leurs propriétés ainsi que les relations qu'ils entretiennent.

À la base, les mouvements d'empilement et de dépilement opérés dans la pile de focalisation sont sous la dépendance de la structure intentionnelle. En effet, l'espace de focalisation d'un segment S2 est empilé au-dessus de l'espace de focalisation d'un segment S1 parce que le BSD de S1 domine celui de S2. La relation de dominance immédiate entre deux BSDs détermine ce mouvement de base dans la pile de focalisation mais la structure intentionnelle constitue également le guide incontournable dans des cas où la gestion de la pile est plus complexe, comme lors des interruptions.

Une interruption correspond intuitivement à une rupture soudaine du discours. Plus précisément, on fait la distinction entre les interruptions dans lesquelles le nouveau segment, appelé ici S_{int}, n'a de rapport sur le plan intentionnel (de dominance ou de précédence) avec aucun des segments précédemment rencontrés et les interruptions où cette absence de rapport est limitée au dernier segment. Dans le premier cas, l'empilement du nouvel espace de focalisation est accompagné d'une absence totale de visibilité en dehors de cet espace. Le second cas est plus complexe puisqu'il se traduit par un dépilement temporaire (jusqu'à la terminaison de S_{int}) des espaces de focalisation situés entre celui de S_{int} et l'espace de focalisation du segment, appelé S_{ret}, dont le BSD est lié à celui du nouveau segment S_{int}. On se retrouve alors, du point de vue de la pile de focalisation, dans le même état que si l'empilement de l'espace de focalisation de S_{int} avait suivi de façon immédiate celui de S_{ret}.

Le mécanisme de focalisation de l'attention suppose un traitement des textes réalisé dans le respect de leur linéarité. Pour chaque proposition, on doit déterminer si elle débute un nouveau segment, si elle clôt le segment courant ou si plus simplement, elle ne fait que développer le segment courant. En pratique, on teste les deux premiers cas et en cas de réponse négative pour les deux, on conclut au troisième. La détermination du commencement ou de la terminaison d'un segment s'appuie sur deux types d'indications : la présence d'indices linguistiques explicites et/ou la possibilité de lier l'intention de la proposition à l'un des BSDs de la structure intentionnelle accessibles à ce moment du processus de traitement, c'est-à-dire un BSD associé à l'un des segments dont l'espace de focalisation est encore présent dans la pile de focalisation. Ce lien peut être établi à partir de principes généraux sur les interactions entre intentions et croyances en liaison avec le contexte courant représenté par la pile de focalisation ou en ayant recours à des connaissances plus ou moins générales sur l'objet du discours. Dans ce dernier cas, on se rapproche du travail de Grau exposé précédemment.

Bien que leur modèle ne s'attache pas spécifiquement à un point de vue thématique, Grosz et Sidner avance l'hypothèse selon laquelle la notion de thème du discours peut être identifiée à celle de BSD, sans néanmoins avancer davantage d'arguments dans ce sens. Il

nous semble que cette affirmation n'est que partiellement juste. Lorsque les relations de dominance entre les BSDs sont établies sur la base de connaissances générales sur le domaine, on peut effectivement penser que les BSDs correspondent à des thèmes. Le modèle est d'ailleurs assez proche dans son esprit de celui proposé par Grau, ainsi que nous l'avons noté ci-dessus.

En revanche, cette parenté ne nous paraît plus justifiée lorsque les BSDs sont établies sur la base des intentions et croyances des rédacteurs/locuteurs et des lecteurs/auditeurs. Le fait de mélanger intention et thème entretient une ambiguïté qui persiste dans des travaux ultérieurs portant sur la segmentation du discours et faisant explicitement référence au modèle de Grosz et Sidner (cf. [Litman & Passonneau 1995] par exemple). Même si l'on peut penser que la notion d'intention est plus déterminante au niveau des dialogues et celle de thème plus importante au niveau des textes, le fait de conserver la distinction entre les deux permet d'être plus précis (cf. [Grau & Vilnat 1997] dans le cas des dialogues). Une intention a bien entendu toujours un objet mais il n'est pas évident que le découpage en intentions suive exactement le découpage selon les objets, même si les deux sont en relation étroite. Plusieurs intentions peuvent en effet avoir le même objet et à l'inverse, une même intention peut porter sur plusieurs objets.

1.4. Le problème de l'analyse thématique dans MLK

L'analyse thématique de MLK se situe un peu à cheval entre les deux grands courants dégagés ci-dessus. Elle relève de l'approche la plus élaborée dans la mesure où elle prend place au niveau conceptuel, manipule des connaissances structurées s'appuyant sur des concepts et enfin, met en œuvre un découpage des textes s'affranchissant de leur linéarité. En revanche, certaines de ses caractéristiques la différencie des travaux présentés. Ses connaissances ne possèdent en effet ni le même degré de généralité, ni le même degré de sûreté que celles utilisées dans le cadre de ces travaux. De plus, le contenu de l'analyse se limite à la tâche de segmentation et n'inclut pas le suivi des thèmes.

La spécificité de l'analyse thématique de MLK se définit plus précisément au travers d'un objectif et d'une contrainte sur les moyens. L'objectif est de pouvoir traiter des textes au style dit très entrelacé comme c'est le cas de beaucoup d'articles journalistiques. La contrainte de moyen est la nécessité d'utiliser la mémoire épisodique comme source principale de connaissances pragmatiques.

La contrainte posée sur les moyens est particulièrement forte puisque les connaissances pragmatiques sont à la base de la discrimination entre les thèmes dans les travaux que

nous avons présentés précédemment. Or les connaissances de la mémoire épisodique présentent des caractéristiques assez différentes de celles propres aux connaissances utilisées dans le cadre de ces travaux. Elles ne sont en effet ni cohérentes, ni générales, ni complètes. Il suffit pour s'en rendre compte d'examiner le contenu d'une UT agrégée.

Tout d'abord, celui-ci n'est pas cohérent. Les graphes agrégés qui composent l'UT sont le résultat d'un cumul indifférencié des événements de toutes les situations particulières rassemblées par l'UT. En aucun cas tous ces événements ne pourraient intervenir dans une seule et même situation. Une tentative de suicide par exemple peut très bien réussir dans un épisode et échouer dans un autre. Le premier cas se traduit par la mort de la personne impliquée alors que dans le second cas, celle-ci reste en vie. Si les deux UTs correspondantes sont agrégées, et aucune contrainte de cohérence ne l'interdit, on peut donc avoir simultanément une chose et son contraire, ce qui pose certains problèmes du point de vue de la généralisation, ainsi que nous l'avons vu au chapitre 7. De façon plus générale, une UT agrégée peut contenir des possibilités exclusives, sans d'ailleurs que celles-ci soient strictement contradictoires. Si l'on achète une voiture avec une somme d'argent donnée, il n'est pas possible d'utiliser cette même somme pour partir en voyage. Or rien n'interdit la présence des deux possibilités dans la même UT agrégée.

Le contenu d'une UT agrégée n'est pas non plus général dans la mesure où la contrainte d'égalité portant sur les prédicats des graphes interdit de fusionner des événements spécifiques pour former des événements plus généraux, donc valides pour un plus grand nombre de situations particulières, ce qui irait en outre dans le sens avec d'une cohérence accrue. Enfin, compte tenu du caractère progressif de la formation des UTs agrégées, rien n'assure que tous les événements propres à une situation aient été exprimés dans les textes rencontrés à un moment donné de la formation d'une UT agrégée. On ne peut donc pas non plus garantir la complétude de cette dernière.

Au delà, on ne peut pas non plus garantir que tous les thèmes abordés dans un texte soient représentés dans la mémoire épisodique. Tout dépend des textes ayant été précédemment traités. En tout état de cause, un thème est toujours nouveau vis-à-vis de la mémoire épisodique la première fois qu'il est rencontré dans un texte. Du point de vue de l'analyse thématique, il faut donc être capable de discriminer les différents thèmes d'un texte tout en étant en mesure dans le même temps de faire la part entre les thèmes déjà connus et ceux qui ne le sont pas encore.

Les caractéristiques de ces connaissances font évidemment que leur usage dans le cadre de l'analyse thématique ne peut être le même que celui qui est fait des connaissances pragmatiques dans des travaux tels que ceux de Grau ou de Grosz et Sidner. Alors que dans ces deux derniers cas, les connaissances pragmatiques se définissent essentiellement par les relations qu'elles entretiennent entre elles, les UTs agrégées se définissent par

rapport à des connaissances d'une autre nature, en l'occurrence la représentation sémantique des événements qu'elles rassemblent. Les principales relations entre les représentations des thèmes sont donc implicites puisque reposant sur le partage d'événements communs. On ne peut de ce fait s'appuyer sur des relations sûres et précises entre ces représentations pour rendre compte du suivi thématique. Par ailleurs, les problèmes d'incomplétude et de manque de cohérence et de généralité mentionnés précédemment obligent à raisonner plutôt de manière statistique que de façon très précise en se fondant sur un seul élément très bien structuré.

L'objectif mis en avant, c'est-à-dire le traitement de textes au style entrelacé, constitue lui aussi une contrainte importante dans la mesure où il limite la possibilité d'exploiter un modèle a priori de la structuration du discours. Le style entrelacé, que l'on retrouve souvent dans les textes journalistiques, est en effet caractérisé par le fait que les différents thèmes abordés par le texte considéré ne sont pas développés en séquence, avec éventuellement des interruptions marquées, mais de manière parallèle. La nécessaire linéarité du discours transforme ce parallélisme en une succession de passages d'un thème à un autre. Ces changements successifs ne s'accompagnent généralement pas de marques de cohésion et seules les connaissances du lecteur sur les thèmes concernés permettent de rendre compte de la cohérence du texte.

Le texte de la figure 8.1, sans être un représentant caricatural de ce phénomène, en offre une certaine illustration. On y trouve deux thèmes principaux : la grève proprement dite des ouvriers des plates-formes de mer du Nord et les conséquences de cette grève, en particulier sur le marché pétrolier. La structure globale du texte est à peu près la suivante. Sur les trois paragraphes constituant le texte, les deux premiers sont dédiés à la présentation de chacun des deux thèmes abordés, un paragraphe étant alors spécifique d'un thème : le premier introduit le thème principal, qui est la grève des ouvriers, tandis que le second présente le thème des conséquences de cette grève sur le marché pétrolier. Le dernier paragraphe, quant à lui, mélange des faits relatifs aux deux thèmes. Il est caractéristique du style entrelacé mentionné précédemment. Il commence par revenir sur la grève proprement dite ("Selon les représentants ... centaines d'ouvriers"), parle ensuite des conséquences sur la production ("Shell et BP, ... pas affectée") et enfin termine sur une information à nouveau relative à la grève ("Cinq arrêts ... mois dernier").

Dans le cas présent, le texte est assez court et le phénomène est donc limité. Avec des textes plus longs, abordant un nombre plus important de sujets, ce phénomène d'entrelacement peut s'avérer plus présent, en particulier lorsque le texte ne se contente pas seulement d'apporter des informations comme c'est le cas ici mais qu'il est le vecteur d'une discussion à propos de ces informations ou d'idées plus générales.

Grève de vingt-quatre heures sur les plates-formes pétrolières de mer du Nord

Les ouvriers des plates-formes de mer du Nord, qui demandent depuis plusieurs semaines la reconnaissance officielle de leur syndicat, l'amélioration des mesures de sécurité et la réintégration d'employés licenciés pour faits de grève, ont entrepris mercredi 12 septembre une nouvelle grève de vingt-quatre heures.

Cette nouvelle a inquiété le marché pétrolier, sensible aux moindres menaces pesant sur la production. Le brut brent britannique a encore progressé mercredi à 30,95 dollars le baril, contre 30,65 dollars la veille.

Selon un représentant du syndicat, la grève a été suivie par plusieurs centaines d'ouvriers. Shell et BP, les deux compagnies visées, indiquaient toutefois en milieu de journée que la production n'étaient pas affectée. Cinq arrêts de travail similaires avaient eu lieu le mois dernier. (AFP)

Fig. 8.1 - Exemple de texte de style journalistique (journal *Le Monde* - 14 septembre 1990)

Les travaux de Grau et de Grosz et Sidner, bien que reposant assez massivement sur des connaissances décrivant le domaine considéré, font certaines hypothèses sur les thèmes accessibles à un moment donné du traitement d'un discours afin de réduire l'espace de recherche au sein de ces connaissances. La pile de focalisation joue ce rôle dans le cas de Grosz et Sidner et le contexte courant dans celui de Grau. Ces hypothèses se traduisent bien évidemment par un certain nombre d'a priori concernant la structuration des textes sur le plan thématique. Même si les textes pris comme exemple pour illustrer ces modèles ont généralement une structure moins hachée que les textes journalistiques, ces a priori ne sont pas fondamentalement en contradiction avec la structure de ce type de textes.

Dans le modèle de Grosz et Sidner, la gestion d'un certain type d'interruptions, en l'occurrence celles dont le BSD est en rapport avec l'un des BSDs de la pile de focalisation, permet de rendre compte de la réintroduction à tout moment de thèmes déjà rencontrés. Dans le modèle de Grau, ce type de phénomène est appréhendé notamment par la présence constante du thème principal au sein du contexte courant, laquelle permet de réintroduire ce thème à tout moment ainsi que les thèmes qui lui sont directement liés.

Néanmoins, dans un cas comme dans l'autre, on peut douter de la possibilité d'appliquer les mécanismes de focalisation présentés sur une large échelle. Comme le soulignent Grosz et Sidner dans [Grosz & Sidner 1986], la détection d'un nouveau segment et la détermination du lien qui l'unit aux autres segments ne peuvent pas toujours être réalisées dès le début de ce segment. Pour contourner cette difficulté, Grosz et Sidner évoquent la possibilité d'anticiper la présence d'éléments explicites permettant de mettre

en évidence les BSDs en utilisant des indications partielles. Elles n'apportent cependant pas d'arguments véritablement généraux à l'appui de leur hypothèse. L'utilisation de marques linguistiques spécifiques (les "cue phrases" et les "clue words") est évoquée mais celles-ci ne peuvent au mieux être utilisées que pour détecter le début ou la fin d'un segment, pas pour déterminer le BSD qu'il porte.

Le même problème se pose à l'échelle du texte entier, qui est le segment de discours de plus haut niveau. Grau fait l'hypothèse que le thème principal d'un texte peut être identifié systématiquement au début de celui-ci. Même si le texte de la figure 8.1 va dans ce sens, il est également très facile de trouver des contre-exemples invalidant cette hypothèse (cf. le premier exemple de texte donné dans [Grosz & Sidner 1986]). Dans bon nombre de cas, on risque donc de se tromper de thème principal et compte tenu de l'importance de cette notion dans le modèle, en particulier pour des textes au style entrelacé, il est fort probable d'aboutir rapidement à la détection d'incohérences fantômes bloquant l'analyse.

Compte tenu des restrictions d'application accompagnant la définition d'une structuration a priori du discours, il nous semble préférable de réduire au minimum les hypothèses portant sur ce point et de nous reposer essentiellement sur les connaissances pragmatiques disponibles pour délimiter des segments thématiquement homogènes, même si ces connaissances n'ont pas nécessairement des caractéristiques idéales (cf. ci-dessus). En outre, il devrait être plus facile de structurer a posteriori les différents segments obtenus plutôt que de le faire en même temps qu'ils sont définis. Cette structuration devrait être dans le même temps plus fiable.

2. Une méthode de segmentation thématique fondée sur la mémoire épisodique

Avant de commencer l'exposé du mécanisme d'analyse de MLK, il est nécessaire de préciser que nous présentons dans ce chapitre davantage l'ossature d'une analyse thématique qu'une méthode directement opérationnelle et évaluée. La raison en est qu'une véritable évaluation demanderait la construction d'une mémoire épisodique de grande taille, ce qui représente un travail de modélisation manuelle de très grosse ampleur dans notre contexte de travail actuel. Plutôt que de simplement illustrer les principes de fonctionnement de cette analyse sur un exemple ad hoc, nous avons préféré mettre l'accent sur le développement d'un mécanisme similaire au sein de ROSA (cf. chapitre 10). À ce niveau en effet, ce mécanisme ne peut faire l'objet d'une véritable validation qui, si elle ne présume pas de la valeur de l'analyse de MLK, fournit néanmoins quelques indications quant à sa validité potentielle. D'autre part, par le biais de l'amorçage de ROSA vers MLK, le mécanisme d'analyse de ROSA représente sur un

plus long terme le moyen de faciliter la constitution d'une mémoire épisodique suffisamment grande et diversifiée.

La notion d'ossature, utilisée plus haut, renvoie plus précisément au fait que la concrétisation de la méthode d'analyse de MLK repose sur la définition d'un certain nombre de fonctions, telles que des mesures de similarité par exemple, dont le contenu exact ne peut être déterminé en pratique que par des expérimentations s'appuyant un jeu de test réaliste. Nous avons fait des propositions pour la définition de ces fonctions, à la fois en nous appuyant sur des principes a priori et sur les expériences menées dans le cadre de ROSA, tout en sachant que ces propositions n'ont pas vocation à être intangibles. Elles pourront en effet être remises en question lorsque l'analyse thématique de MLK pourra véritablement être testée.

2.1. Principes

L'objectif poursuivi par la méthode que nous présentons ici est de délimiter les différentes UTs composant la représentation d'un texte. Cette délimitation consiste plus précisément à déterminer combien d'Unités Thématiques forment cette représentation, c'est-à-dire combien de situations sont évoquées par le texte, et surtout à en définir le contenu, autrement dit à rattacher les propositions du texte[1] à ces différentes UTs. Même si elle ne recouvre qu'une partie de la construction des UTs, cette tâche en constitue la base incontournable et permet au moins de produire une forme minimaliste des UTs. Du point de vue de l'analyse thématique, elle se situe un peu au delà de la segmentation thématique "traditionnelle" dans la mesure où cette dernière se contente généralement de découper les textes en blocs contigus traitant chacun d'un thème spécifique ou assumant un rôle fonctionnel déterminé. Il est néanmoins évident que notre volonté d'aborder des textes à la structure complexe nous oblige à dépasser cette conception un peu restrictive de la segmentation des textes.

L'algorithme élaboré dans ce but s'inscrit dans la lignée de ceux, tels celui de Grau ou celui de Grosz et Sidner (cf. §1.3.2), gérant une notion de centre d'attention (focus) : on traite les propositions les unes à la suite des autres en respectant la linéarité de la forme de surface du texte. Pour chacune d'elles, on détermine l'ensemble des faits et des connaissances jugés comme les plus saillants compte tenu à la fois du résultat du traitement des propositions précédentes et de ce que la nouvelle proposition considérée apporte. Cet ensemble, que l'on dénommera contexte de la proposition, délimite ce qui est

[1] Nous utilisons ici le terme assez générique de proposition mais nous faisons référence plus précisément à sa représentation sémantique, c'est-à-dire à un graphe conceptuel.

accessible au processus d'analyse pour établir comment cette proposition s'insère dans la représentation de texte courante.

Comme nous l'avons indiqué au paragraphe 1.4, la façon dont une proposition est mise en relation avec les unités de discours plus larges (ici les UTs) ne repose pas dans notre cas sur un modèle très contraint de la structuration du discours et du suivi thématique. Nous ne faisons donc pas d'hypothèse quant à l'existence de liens de dépendance entre les propositions d'une même UT. Nous considérons que celles-ci peuvent tout à fait ne pas être contiguës et que cette absence de contiguïté globale ne conduit pas davantage à des regroupements par petits blocs. En cela, nous ne reprenons même pas l'hypothèse minimale qui est implicitement faite par l'essentiel des algorithmes de segmentation thématique et qui stipule qu'en l'absence de détection d'un changement de thème, l'unité considérée, ici une proposition, se rattache au segment de texte courant[1].

Le processus d'analyse est donc amené à gérer une liste d'UTs en cours de construction qui sont autant de candidats possibles pour l'intégration de chaque nouvelle proposition traitée. Plus précisément, ce processus doit apporter une solution aux deux problèmes suivants :

- il doit tout d'abord déterminer si la proposition courante fait ou ne fait pas référence à une situation nouvelle vis-à-vis de celles déjà évoquées par le texte et donc, décider quand créer de nouvelles UTs;

- s'il a établi que la proposition courante ne fait pas référence à une situation nouvelle, il doit encore reconnaître à quelle situation, donc à quelle UT en construction, cette proposition se rattache.

En l'occurrence, ces deux problèmes sont joints puisque le jugement de détection d'une nouvelle situation se définit en négatif par rapport à la possibilité de rattacher la proposition courante à l'une des situations déjà abordées. Le processus d'analyse thématique repose donc en grande partie sur l'opération d'évaluation de la cohérence d'une proposition vis-à-vis d'une UT en construction, évaluation qui conditionne la possibilité de rattachement d'une proposition à une UT.

Le processus global de construction des UTs se définit alors de la façon suivante. Soit P, une proposition d'un texte; si la cohérence de P vis-à-vis de l'une au moins des UTs en construction gérées au moment du traitement de P est jugée suffisante, la proposition P

[1] Il serait plus exact de dire que cette hypothèse n'est pas faite de façon explicite. En effet, dès lors que l'on prend en compte le contexte installé par les propositions précédentes dans la caractérisation de la proposition courante (cf. par exemple l'influence du contexte dans la sélection des UTs agrégées de la mémoire épisodique), on a tendance à rapprocher celle-ci des propositions qui la précède.

est intégrée à l'UT en construction pour laquelle cette cohérence est maximale. Au contraire, si l'on estime qu'elle ne se raccroche à aucune des situations déjà identifiées, on crée une nouvelle UT en construction à laquelle la proposition P est rattachée.

2.2. Segmentation en présence de connaissances apprises sur le domaine

L'évaluation de la cohérence d'une proposition par rapport à une UT en construction ne peut se faire par comparaison directe de ces deux entités dans la mesure où le nombre de concepts d'une proposition, voire celui d'une UT lorsqu'elle se trouve au début de sa formation, est insuffisant pour qu'une telle comparaison soit significative. Cette évaluation nécessite donc de faire appel à une source de connaissances sur les situations évoquées qui pourra seule déterminer si l'événement auquel réfère la proposition s'inscrit dans la situation représentée par l'UT en construction. Ce rôle de référence sur les situations est naturellement dévolu à la mémoire épisodique dans le cas présent. Cependant, compte tenu du caractère incertain des connaissances abritées par cette mémoire, il ne nous a pas paru possible d'exploiter les représentations de texte qu'elle contient à la manière d'un système de raisonnement à base de cas "traditionnel", c'est-à-dire en essayant de déterminer, pour juger de la cohérence d'une proposition par rapport à une UT en construction, si cette proposition, ou tout du moins une proposition proche ou liée, apparaît dans une UT de la mémoire similaire à l'UT en construction.

En revanche, la mémoire épisodique, conjuguée à son mécanisme de rappel, offre le moyen de réaliser ce jugement de cohérence par procuration. Plutôt que de comparer directement une proposition et une UT, on compare ici les représentations construites à partir des connaissances que cette proposition et cette UT ont permis de sélectionner au sein de la mémoire épisodique. Du point de vue du raisonnement à base de cas, cette façon de procéder met l'accent sur la phase de recherche en mémoire des cas en relation avec le problème à traiter et réduit à la portion congrue la phase d'adaptation. Cela va dans le sens du raisonnement à base d'expériences, présenté au chapitre 1 comme une spécialisation du raisonnement à base de cas adaptée au cadre défini par ANTHAPSI.

Les représentations associées à la fois aux UTs en construction et aux propositions sont plus précisément dénommées *contextes.* Le contexte d'une proposition s'identifie au contexte défini en préambule comme l'ensemble des connaissances entrant dans le champ du centre d'attention. Sur le plan pratique, il est constitué des UTs agrégées les plus activées de la mémoire épisodique obtenues à la suite d'une phase de rappel initiée à partir des concepts de la proposition, conformément au processus décrit à la section 3 du

chapitre 6. Chacune de ces UTs possède un poids, égal à son niveau d'activité à l'issue du rappel. Les UTs agrégées retenues sont les N UTs agrégées les plus activées parmi celles possédant un niveau d'activité dépassant le seuil $S_{sélUTA}$. Celui-ci est donné par la somme de la moyenne des niveaux d'activité de toutes les UTs agrégées activées et de leur écart-type. N spécifie la taille du contexte de la proposition. Au sein de celui-ci, les UTs agrégées sont ordonnées selon l'ordre décroissant de leur poids.

L'ensemble de concepts utilisé pour amorcer le rappel est en réalité formé non seulement des concepts de la proposition traitée mais également des concepts de la proposition qui la précède et ceux de la proposition qui la suit. Cet ensemble s'identifie à une fenêtre glissante d'une taille de 3 propositions que l'on déplace sur tout le texte de façon à ce que chacune de ses propositions en occupe le centre à un moment donné. Les concepts de la proposition occupant le centre de cette fenêtre se voient attribuer une activité initiale plus forte que les autres afin de marquer leur caractère prépondérant à ce stade de l'analyse. La taille exacte de la fenêtre est un paramètre qu'il convient d'ajuster sur la base de tests réalisés sur des textes disposant d'une segmentation de référence. La valeur de 3 est une sorte de valeur par défaut : elle est supérieure à la taille minimale, égale à 1, tout en restant suffisamment petite pour que la fenêtre n'inclut pas de propositions trop hétérogènes. La prise en compte ainsi réalisée de l'environnement d'une proposition répond au besoin de lisser dans le temps les valeurs de similarité entre le contexte des UTs en construction et le contexte des propositions. Une assise plus large des indices de rappel permet en effet de s'abstraire des micro-variations brusques pouvant résulter du passage d'une proposition à une autre. Ce phénomène se manifeste en particulier lorsqu'une des propositions est assez marquée thématiquement alors que celle qui la suit ou qui la précède est très générale.

Afin que la comparaison entre UT en construction et proposition soit possible, le contexte d'une UT en construction est de même nature que celui d'une proposition. Il est donc formé d'un ensemble d'UTs agrégées pondérées (cf. figure 8.2). Ce sont en l'occurrence les UTs de la mémoire revenant le plus souvent et avec le plus de force lorsque cette UT en construction est évoquée. Mais contrairement à ce qui se passe pour les propositions, ces UTs agrégées ne sont pas sélectionnées en faisant appel directement à l'associativité de la mémoire épisodique à partir du contenu de l'UT en construction. Ce contexte est concrètement le résultat de la fusion des contextes associés aux différentes propositions intégrées dans cette UT en construction. Cette fusion est bien entendu incrémentale : le contexte est mis à jour à chaque intégration d'une nouvelle proposition au sein de l'UT en construction.

La fusion d'un contexte d'UT en construction et d'un contexte de proposition se déroule en trois étapes. La première d'entre elles consiste à fusionner les deux listes d'UTs agrégées en recalculant le poids des UTs communes. Les UTs présentes dans seulement un des deux contextes conservent leur poids d'origine. La deuxième étape réalise le tri du résultat de la fusion précédente suivant l'ordre décroissant des poids des UTs agrégées. La dernière, enfin, produit le contexte actualisé en supprimant les UTs agrégées de plus faible poids au delà des *N* plus importantes. Les UTs agrégées du contexte de la proposition qui ne sont pas présentes dans le contexte de l'UT en construction peuvent ainsi remplacer des UTs de celui-ci si leur poids est supérieur au poids de ces dernières.

Le recalcul lors de la première étape du poids des UTs agrégées communes est donné par la fonction *F*, telle que :

$$poids(t+1, UTC_j, UTA_i) = F(poids(t, UTC_j, UTA_i), poids(Prop_t, UTA_i), t) \qquad [1]$$

avec $poids(t,UTC_j,UTA_i)$: poids au sein du contexte de l'UT en construction UTC_j de l'UT agrégée UTA_i après l'intégration de t proposition dans UTC_j;

$poids(Prop_t,UTA_i)$: poids de l'UT agrégée UTA_i au sein du contexte associé à la proposition $Prop_k$.

La fonction *F* fait partie des paramètres que nous avions évoqués en préambule de la présentation de la méthode d'analyse de MLK et dont la définition exacte devra être confirmée lors de futures expérimentations. Suivant l'importance qu'elle accorde respectivement au poids des UTA_i dans le contexte de l'UT en construction et au poids de ces mêmes UTs agrégées dans le contexte de la proposition, on choisit de favoriser une évolution lente ou bien rapide du contenu du contexte de l'UT en construction. De cette façon, on ajuste également son degré de sensibilité vis-à-vis des variations introduites par les propositions de textes tout au long du texte : par exemple, si une prime est donnée par *F* au poids des UTA_i du contexte de l'UT en construction, une évolution du contexte des propositions ne sera répercutée au niveau du contexte de l'UT en construction que si elle se manifeste suffisamment longtemps.

On définit ainsi le rôle exact que l'on fait jouer au contexte d'une UT en construction. Celui-ci peut être le réceptacle des tendances des différentes propositions de cette UT et produire finalement une représentation synthétique de celle-ci mettant en évidence les UTs agrégées les plus fréquemment sélectionnées lors de son évocation. Il est ainsi le moyen de définir les UTs agrégées avec lesquelles cette nouvelle UT est susceptible de s'agréger pour être mémorisée. Ce fonctionnement par accumulation est comparable au mode de construction des structures agrégées de la mémoire épisodique.

Mais le contexte d'une UT en construction est surtout un outil au service de l'analyse thématique. En cela, il doit favoriser le plus possible le rattachement des propositions aux UTs en cours de construction. La façon dont un même thème est évoqué tout au long d'un texte peut évoluer et il est à cet égard intéressant que la représentation de ce thème puisse suivre cette évolution en parallèle afin d'être le plus proche possible de la manière courante dont ce thème est évoqué. Pour obtenir ce résultat, la fonction *F* doit alors favoriser le poids des UTs agrégées du contexte des propositions par rapport au poids des UTs agrégées du contexte des UTs en construction.

Remarquons, pour finir cette discussion à propos de *F*, qu'une UT en construction ne présente pas les mêmes caractéristiques aux différents stades de son développement et que par conséquent, il pourrait être opportun d'adapter la définition de *F* à l'état de l'UT considérée (c'est l'objet de la variable *t* apparaissant dans la spécification de F donnée par [1]). Lorsqu'une UT en construction est au début de sa formation, elle ne regroupe par essence que peu d'éléments et ceux-ci ne sont pas nécessairement très représentatifs de la situation évoquée. Il est donc nécessaire de favoriser les apports des propositions venant progressivement compléter cette UT et affirmer ainsi sa définition. Dans cette phase, *F* doit donc favoriser l'influence du contenu du contexte des propositions par rapport au contenu du contexte de l'UT en construction. En revanche, lorsque l'UT a atteint un certain degré de stabilité (cf. §2.3 pour l'évaluation de cette stabilité), seules les modifications significatives apportées par de nouvelles propositions sont à prendre en compte par le contexte de cette UT. *F* doit alors conférer une plus grande inertie au poids des UTs agrégées composant ce contexte.

En l'absence de retour possible sur les choix réalisés, nous avons choisi de donner à *F* la forme la plus simple possible permettant de tester si besoin est les différentes options mentionnées ci-dessus. *F* est ainsi définie comme une fonction linéaire vis-à-vis des poids des UTs agrégées :

$$poids(t+1, UTC_j, UTA_i) = \alpha(t)\ \ poids(t, UTC_j, UTA_i) + \beta(t)\ \ poids(Prop_t, UTA_i)) \quad [2]$$

Nous avons retenu plus précisément la fonction *F* obtenue pour $\alpha(t)=1$ et $\beta(t)=1$. Elle correspond en l'occurrence au premier mode de fonctionnement du contexte de l'UT en construction exposé précédemment, autrement dit le fonctionnement par accumulation des contextes des propositions. À notre sens en effet, celui-ci représente le meilleur compromis entre les différentes possibilités offertes par la forme de *F*.

Le choix de *F* permet de fixer précisément la procédure de construction du contexte d'une UT en construction. Une proposition en cours de traitement et une UT en construction se retrouvent ainsi caractérisées de manière uniforme. Il est dès lors possible

d'évaluer la compatibilité des deux en ayant recours à une mesure de similarité opérant sur leurs contextes. Cette mesure ne porte que sur les UTs agrégées communes aux deux contextes. La présence d'une UT dans l'un des deux et pas dans l'autre ne véhicule pas en effet un sens particulier. Afin de capturer le plus finement possible les variations touchant les UTs agrégées communes d'un contexte à l'autre, la mesure d'évaluation de la similarité entre contextes repose sur les quatre facteurs suivants :

- l'importance en termes de poids des UTs agrégées communes aux deux contextes par rapport à l'ensemble des UTs du contexte de la proposition. Cette importance est évaluée par le rapport entre la somme des poids des unes et la somme des poids des autres;

- l'importance en termes de poids des UTs agrégées communes aux deux contextes par rapport à l'ensemble des UTs du contexte de l'UT en construction. Comme pour la proposition, cette importance est évaluée par le rapport des poids;

- l'importance en termes de nombre d'entités des UTs agrégées communes aux deux contextes par rapport aux UTs composant les contextes. Cette importance est donnée par le rapport entre le nombre des premières et la taille des contextes. Cette dernière, égale à *N*, est la même pour les deux contextes, ce qui explique que ce facteur n'est pas dédoublé comme dans le cas des poids;

- la différence d'ordonnancement des UTs agrégées communes aux deux contextes. Étant donné que plus cette différence est grande et moins les contextes sont similaires, on prend plus exactement le complémentaire par rapport à 1 de cette valeur. La différence d'ordonnancement est évaluée de la façon suivante : pour chaque UT agrégée commune aux deux contextes, on calcule la valeur absolue de la différence entre le numéro d'ordre, appelé également rang, de l'UT dans un contexte et son rang dans l'autre contexte. La différence d'ordonnancement est la moyenne arithmétique de ces valeurs, normalisée par la différence maximale de rang possible, soit *N-1*. Cette normalisation permet d'obtenir une valeur entre 0 et 1. Plus formellement, cette différence s'écrit :

$$diffRang\,(Ctxt_{UTC}, Ctxt_{Prop}) = 1 - \left(\frac{\sum_{c=1}^{p} \left| rang(UTA_c, Ctxt_{UTC}) - rang(UTA_c, Ctxt_{Prop}) \right|}{(N-1)\ p} \right)$$

avec $Ctxt_{UTC}$: contexte de l'UT en construction;
$Ctxt_{Prop}$: contexte de la proposition;
UTA_c : UT agrégée commune aux deux contextes;
p : nombre d'UTs agrégées communes aux deux contextes.

Ces quatre facteurs sont combinés selon une moyenne géométrique. On obtient ainsi la mesure de similarité entre contextes :

$$sim(Ctxt_{UTC}, Ctxt_{Prop}) = \left(\frac{\sum_{c=1}^{p} poids(UTA_c, Ctxt_{UTC})}{\sum_{i=1}^{N} poids(UTA_i, Ctxt_{UTC})} \cdot \frac{\sum_{c=1}^{p} poids(UTA_c, Ctxt_{Prop})}{\sum_{i=1}^{N} poids(UTA_i, Ctxt_{Prop})} \cdot \frac{p}{N} \; diffRang(Ctxt_{UTC}, Ctxt_{Prop}) \right)^{1/4} \quad [3]$$

Les valeurs des quatre facteurs se situent toutes entre 0 et 1. En conséquence, cette mesure de similarité est encadrée par les mêmes valeurs.

Globalement, on peut dire que ses trois premiers termes caractérisent l'importance des UTs agrégées communes aux deux contextes considérés vis-à-vis de chacun d'entre eux. Son dernier terme rend compte quant à lui de la similarité des caractéristiques que possèdent ces UTs communes au sein de chacun de ces contextes. En l'occurrence, on a retenu comme caractéristique le rang de l'UT agrégée dans le contexte.

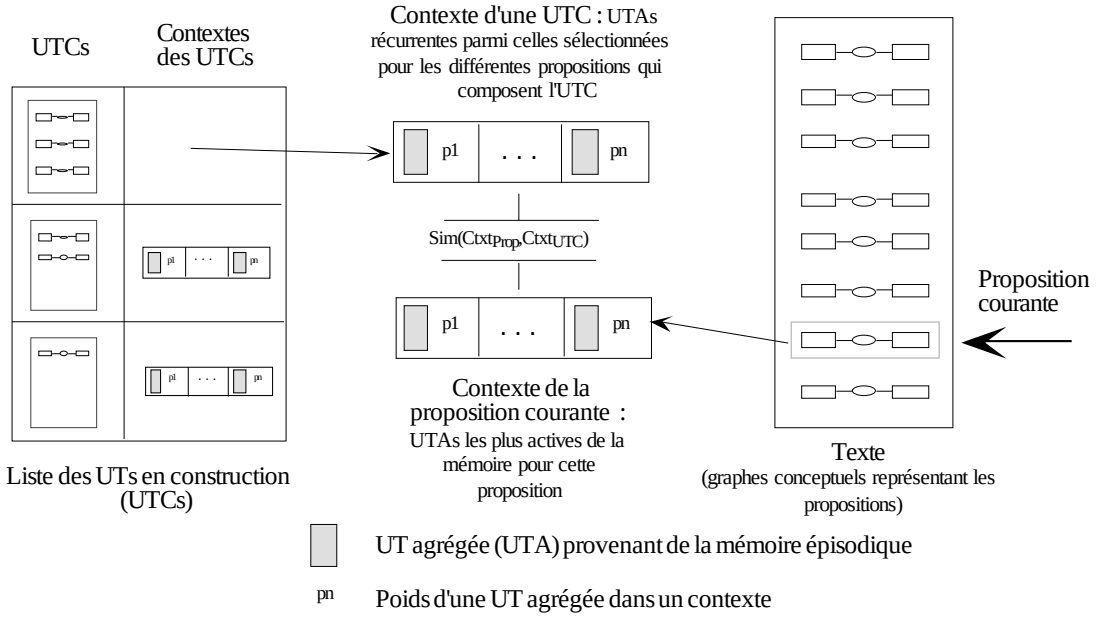


Fig. 8.2 - Comparaison des contextes des propositions et des UTs en construction

Comme la fonction *F*, la mesure de similarité entre contextes fait partie des paramètres de l'algorithme qui sont susceptibles d'être adaptés en fonction des résultats d'expérimentations futures. Il existe d'ailleurs un certain lien de dépendance entre cette mesure et la fonction *F*. Si la mesure de similarité est assez stricte, le problème de l'évolution du contenu du contexte des UTs en construction se pose en effet de façon très

affaiblie puisque les propositions qui sont rattachées à une UT ont alors un contexte très proche de celui de cette UT.

Une fois que les moyens d'évaluer la cohérence entre une proposition d'un texte et une UT en cours de construction ont été précisés, on peut mettre en œuvre les principes décrits au §2.1. Pour déterminer si la proposition en cours d'analyse se rattache à l'une des UTs en construction, on applique la mesure de similarité définie ci-dessus entre le contexte de chacune des UTs en construction et le contexte de la proposition considérée (cf. figure 8.2). Si l'une au moins de ces valeurs est supérieure à $S_{nouvUTC}$, le seuil en dessous duquel une nouvelle UT en construction doit être créée, la proposition est intégrée à l'UT en construction de plus forte valeur de similarité. Si en revanche cette condition n'est pas remplie, une nouvelle UT doit être introduite, UT dont le premier constituant sera la proposition considérée.

Dans le cas par exemple du texte de la figure 5.2 évoquant une tentative d'assassinat à l'encontre de Martin Luther King (cf. chapitre 5), l'analyse se déroulerait comme suit, en supposant que l'on dispose d'UTs agrégées relatives aux différents thèmes abordés[1] : le traitement de la première proposition ("je me trouvais dans un grand magasin de Harlem") entraîne la création, en quelque sorte par défaut, d'une première UT. À ce stade, la fenêtre de délimitation des indices de rappel, que l'on appellera *fenêtre texte* dans ce qui suit, ne regroupe que les deux premières propositions. La deuxième proposition ("entouré de quelques centaines de personnes") est véritablement très générale et n'évoque pas de situation spécifique. Les UTs agrégées sélectionnées pour former le contexte de cette première UT ne sont donc liées qu'au contenu de la première proposition. Elles font plus précisément référence aux situations mettant en scène prototypiquement un grand magasin, comme celle consistant à faire ses courses par exemple.

La deuxième proposition est pour sa part rattachée à l'UT nouvellement créée : son manque de spécificité n'introduit pas de nouvelle situation et la fenêtre texte centrée sur elle (propositions (1), (2) et (3)) inclut toujours la première proposition. Ajouté à l'effet de mémoire du mécanisme de rappel, ce dernier point conduit donc à sélectionner à nouveau des UTs agrégées proches du contexte de la seule UT existante. Dans le même temps, la troisième proposition ("j'étais en train de dédicacer des exemplaires de mon livre 'Stride toward Freedom'"), plus marquée thématiquement que les deux premières (types de concept *Dédicacer* et *Livre*), entraîne déjà la sélection d'UTs agrégées liées au

[1] L'analyse présentée ici n'est qu'un scénario mais celui-ci s'inspire de l'application d'un mécanisme d'analyse proche, en l'occurrence celui de SEGAPSITH, sur le même texte. Il est par ailleurs évident que le déroulement de l'analyse est étroitement dépendant des connaissances disponibles.

domaine de l'édition et du livre. Le contexte de la deuxième proposition se trouve ainsi partagé au moins entre deux thèmes et le rattachement de cette dernière ne s'appuie pas sur une similarité très forte. La mise à jour qui s'en suit du contexte de l'UT en construction concernée y introduit certaines des UTs agrégées en relation avec le thème de l'édition et du livre sélectionnées par la troisième proposition.

Le rattachement de celle-ci à la même UT que les deux premières est également le fait d'une similarité assez faible : la fenêtre texte contient alors une proposition très générale, la deuxième, une proposition liée à un sujet non abordé jusqu'à présent, la quatrième ("qui relate le boycottage des autobus de Montgomery en 1955-56"), et une proposition qui n'est en rapport qu'avec une partie du contexte de l'UT en construction considérée. Ce dernier lien, renforcé par la prime accordée aux concepts de la proposition centrale et conjugué au fait que la fenêtre texte ne fait pas référence majoritairement à un autre thème, est néanmoins suffisant pour qu'une nouvelle UT en construction ne soit pas créée.

On constate à l'occasion de ce premier morceau d'analyse que la décision de compatibilitéentre une proposition et une UT en construction est parfois assez fragile pour les toutes premières propositions faisant référence à une nouvelle situation. Elle devient naturellement beaucoup plus franche à mesure que l'UT représentant cette situation prend davantage corps : seules les UTs agrégées apparues de façon récurrente subsistent en effet au sein de son contexte. L'apparition d'une nouvelle situation est donc plus facilement détectable si la situation précédente a été suffisamment développée.

Ce début d'analyse fait également apparaître l'intérêt d'utiliser des contextes formés de plusieurs UTs agrégées pour représenter les propositions et les UTs en construction. La proposition (3) fait référence à une situation de séance de dédicace. Il est possible que cette situation soit représentée en tant que telle par une UT agrégée de la mémoire épisodique, auquel cas celle-ci pourra être utilisée pour contextualiser la proposition en question. Dans un cadre fortement marqué par la progressivité de l'apprentissage, il est encore plus probable qu'une telle UT n'existe pas, mais que d'autres UTs agrégées liées tout de même au domaine du livre et de l'édition soient disponibles. L'essentiel, en l'occurrence, est que ces UTs puissent témoigner de l'attachement d'un ensemble de types de concept, ici ceux des propositions (3) et (5), à un même thème. Les utiliser pour former les contextes des propositions et des UTs en construction permet de détecter la compatibilité d'une proposition et d'une UT sans avoir nécessairement une représentation de référence de la situation évoquée par cette UT. Cet aspect est particulièrement important pour l'apprentissage de nouvelles situations et la démarche adoptée va à cet égard au delà de ce qu'autorise un raisonnement à base de cas "traditionnel".

L'analyse du texte se poursuit quant à elle par le traitement de la quatrième proposition. Celle-ci fait référence à une nouvelle situation mais avec les propositions (3) et (5) ("j'apposais ma signature sur une page"), le contenu de la fenêtre texte se trouve fortement orienté vers le thème du livre. Compte tenu de la présence de ce thème dans le contexte de la seule UT existante, ceci conduit à rattacher également la quatrième proposition à cette UT. Ce cas illustre l'importance que revêt la taille de la fenêtre texte. Elle détermine en effet la taille minimale, en nombre de propositions, que doit avoir l'évocation d'une situation pour donner lieu à la création d'une UT spécifique. Si le sujet de la quatrième proposition avait été développé par une cinquième et même une sixième proposition, un changement de thème aurait sans doute été détecté avec le regard en avant et en arrière limité à une seule proposition que nous avons ici. Avec une seule proposition parmi les trois de la fenêtre texte, le thème du boycottage des bus se retrouve en revanche toujours minoritaire et n'a donc que peu de chances d'émerger en tant que tel, sachant d'autre part qu'il est encadré par deux propositions thématiquement marquées.

La cinquième proposition constitue un point de transition, dont le traitement est susceptible de poser problème. Lorsqu'elle est centrée sur cette proposition, la fenêtre texte contient en effet trois propositions ((4), (5) et (6-7), composée des propositions (6) ("je sentis quelque chose de pointu s'enfoncer brutalement dans ma poitrine") et (7) ("je venais d'être poignardé à l'aide d'un coupe-papier, par une femme"), réunies par l'analyse sémantique) faisant référence à trois sujets distinctement marqués (une action de boycottage, une séance de dédicace et une tentative de meurtre). Seuls la prédominance de l'activité des concepts de la proposition centrale et l'effet d'inertie caractérisant le rappel permettent de rattacher cette cinquième proposition à l'UT en construction active depuis le début du texte.

Le traitement de la proposition (6-7) entraîne en ce qui la concerne un changement suffisant de la configuration des UTs agrégées les plus actives de la mémoire épisodique pour que la similarité avec le contexte de la seule UT en construction tombe en dessous de $S_{nouvUTC}$ et provoque la création d'une nouvelle UT. Les types de concept de (6-7) deviennent en effet prédominants au sein de la fenêtre texte tandis que le thème du livre n'est plus représenté que par une proposition périphérique et ne bénéficie plus véritablement de l'inertie du rappel. Par ailleurs, la proposition (8) ("qui devait être reconnue folle par la suite"), même si elle ne renforce pas obligatoirement le thème de (6-7) – la folie n'intervenant pas dans toutes les tentatives de meurtre – va plutôt dans son sens que dans celui du thème du livre.

Ce changement dans la configuration des UTs agrégées activées se reproduit au niveau de la proposition (9) ("on me transporta d'urgence à l'Hôpital de Harlem"). Cette

transition introduit néanmoins une différence par rapport à la répartition des propositions entre les UTs de la représentation proposée au chapitre 5 pour le même texte. Dans cette représentation (cf. figure 5.5), les propositions (8) et (9) sont en effet affectées à une UT TentativeAssassinat, en même temps que la proposition (6-7). Or, s'il est peut être possible d'affecter les propositions (6-7) et (8) à la même UT, il est beaucoup moins probable d'y ajouter la proposition (9). Dans le cas de (8), tout dépend de ce que le type de concept *Fou* permet de sélectionner dans la mémoire épisodique. Si les UTs agrégées concernées font davantage référence aux actes criminels qu'au monde médical, il est possible de raccrocher (8) à la même UT que (6-7) en bénéficiant à la fois de la présence de (6-7) dans la fenêtre texte centrée sur (8) et de l'inertie du rappel.

Pour (9) en revanche, un tel rattachement paraît plus difficile car les propositions constituant la fenêtre texte centrée sur (9) renvoient plus au monde médical qu'au monde criminel, sauf bien entendu si des UTs allant dans le sens contraire sont présentes dans la mémoire épisodique. Bien que l'attachement au thème de l'hospitalisation est faible pour (10) ("je restai de longues heures sur un lit"), la conjonction avec (9) dans la fenêtre texte contribue à renforcer globalement la sélection de ce thème. Voici donc un cas dans lequel l'algorithme proposé donnera probablement un résultat différent de la segmentation de référence proposée.

L'explication de cette différence réside notamment dans le fait que nous avons affaire ici à une déviation, c'est-à-dire à une sorte de glissement thématique plutôt qu'à un net changement de thème. La capacité à détecter cette évolution thématique ainsi que le moment où on la détecte dépendent alors pour beaucoup de la valeur donnée au seuil $S_{nouvUTC}$. En le plaçant plus ou moins haut, on rend également plus ou moins fréquente la création de nouvelles UTs. On contrôle ainsi la forme des représentations de texte produites et donc plus largement, la forme de la mémoire épisodique dans laquelle elles seront intégrées.

Avec une valeur haute de ce seuil, on aura tendance à créer davantage d'UTs pour un même texte, celles-ci étant aussi plus spécifiques. En particulier, on détectera mieux les déviations de thème, mais bien sûr avec un certain risque de bruit. Celui-ci se manifeste par la création de plus d'UTs que nécessaire lorsque les propositions ne sont pas très marquées thématiquement, comme c'est le cas par exemple des propositions (10), (11) ("on faisait mille préparatifs") et (12) ("pour extraire l'arme de mon corps"). Si le seuil au contraire est bas, on intégrera au sein d'une même UT des propositions plus hétérogènes sur le plan thématique et en final, on obtiendra des UTs thématiquement plus étendues.

Faisons remarquer, pour clore cette analyse, que le rattachement de (12) à la même UT que (9), (10) et (11) n'apparaît pas nécessairement évident. Les types de concept *Arme* et

Corps ont en effet plus de chances de renvoyer à des situations liées à la criminalité, au terrorisme et à la guerre qu'à des situations liées au monde médical. En outre, la seule autre proposition présente dans la fenêtre texte, la proposition (11), est très neutre du point de vue thématique. Il est donc assez probable de rattacher (12) à la même UT que (6-7). Ce n'est d'ailleurs pas sans justification puisque (12) fait effectivement référence à la situation représentée par cette UT. Seule la présence de propositions plus marquées par le thème de l'hospitalisation venant à la suite de (12) permettrait d'aller à l'encontre de cette tendance.

2.3. La prise en compte de l'incomplétude de la mémoire épisodique

Le principe exposé ci-dessus repose implicitement sur l'hypothèse que la mémoire épisodique contient des UTs agrégées proches des situations évoquées par les textes analysés. Dans un cadre comme MLK où apprentissage et compréhension sont étroitement mêlés, il est cependant nécessaire de ne pas se cantonner à faire évoluer les représentations des situations déjà présentes en mémoire mais de permettre également d'ajouter à cette dernière de nouvelles situations. Nous les appellerons dans ce qui suit des *situations inédites* et les UTs qui les représentent seront appelées *UTs inédites*. Au delà de sa capacité à discerner la manifestation dans les textes de différentes situations possédant une représentation dans la mémoire épisodique, l'analyse thématique de MLK doit donc aussi être capable de faire la distinction entre l'évocation de situations connues et celle de situations inédites.

De ce point de vue, le critère qu'elle utilise pour créer de nouvelles UTs est ambigu. La similarité entre le contexte de la proposition courante et le contexte de chacune des UTs en construction peut en effet ne pas dépasser le seuil $S_{nouvUTC}$ pour deux raisons différentes :

- les situations formant le contexte de la proposition sont différentes de celles rencontrées précédemment au cours de l'analyse du texte. C'est le cas d'une évolution thématique vers une situation représentée en mémoire mais différente des situations évoquées par les UTs du texte déjà construites;

- le contexte de la proposition ainsi que celui de l'UT en construction à laquelle cette proposition devrait être rattachée ne sont pas significatifs compte tenu de l'absence au sein de la mémoire d'UTs agrégées entrant en résonance avec le contenu thématique de la proposition et de l'UT. C'est le cas d'une proposition devant être rattachée à une UT représentant une situation inédite du point de vue de la mémoire épisodique.

Pour lever cette ambiguïté, il faut pouvoir reconnaître la manifestation d'une situation inédite. Cette reconnaissance intervient à deux niveaux :

- au niveau de la proposition en cours d'analyse d'abord; si le seuil $S_{nouvUTC}$ n'est franchi pour aucune des UTs en construction, on contrôle que le contexte de la proposition en question peut être considéré comme *significatif*;

- au niveau de la liste des UTs en construction ensuite; dans le cas où le contexte de la proposition courante laisse à penser que celle-ci doit se rattacher à une situation inédite, on regarde dans la liste des UTs en construction si l'une d'entre elles ne représente pas une telle situation.
  On notera que l'évaluation du caractère inédit d'une UT en construction est réalisée dynamiquement, ce qui permet de ne pas cataloguer définitivement comme inédite une UT sur laquelle, à un moment donné de l'analyse, on n'aurait pas un recoupement important entre les informations apportées par le texte et les informations présentes en mémoire.

Pour détecter que le contexte d'une proposition est non *significatif,* on s'appuie encore une fois sur l'associativité de la mémoire épisodique ou plus exactement sur le négatif de cette propriété : une situation inédite ne pouvant être mise en relation avec une ou plusieurs UTs agrégées proches de cette situation, l'activité de la mémoire reste en effet à un niveau faible lorsque les propositions qui y font référence sont présentées en tant qu'"entrées" de la mémoire. Compte tenu des propriétés du mécanisme de rappel (activité des unités du réseau de propagation non bornée, ...), cette faiblesse de l'activité est cependant difficile à détecter en toute généralité.

On s'appuie donc sur une autre caractéristique de l'activité de la mémoire en pareille situation : non seulement celle-ci est faible mais elle est également très uniforme. On définit ainsi que le contexte d'une proposition est non significatif lorsqu'il est établi à partir d'une configuration d'activité de la mémoire épisodique caractérisée par un taux de relief inférieur à un seuil pré-déterminé. Plus précisément, ce taux est évalué en déterminant la proportion d'UTs agrégées, par rapport à toutes celles ayant été activées, possédant une activité dépassant le seuil fixé par la somme de la moyenne de ces activités et de leur écart-type.

L'évaluation du caractère inédit d'une UT en construction repose quant à elle sur les deux critères suivants :

- la valeur moyenne des poids associés aux UTs agrégées formant le contexte doit être la plus faible de toutes les moyennes similaires calculées pour les contextes de l'ensemble des UTs en construction;

- le second critère exploite l'incohérence caractérisant une UT inédite du point de vue de la mémoire épisodique. Une telle UT représente en effet un assemblage de propositions sans justification puisque réalisé en dehors de toute référence à des situations déjà rencontrées. Pour matérialiser le degré de cohérence d'une UT en construction, on calcule son niveau de stabilité. Celui-ci se définit comme inversement proportionnel au degré d'évolution du contenu de l'UT. Il est évalué à chaque fois qu'une nouvelle proposition est intégrée à l'UT en construction. Il est le résultat de l'application de la mesure de similarité entre contextes exposée précédemment entre le contexte de l'UT avant intégration de la nouvelle proposition et le contexte après cette intégration. Le niveau de stabilité d'une UT renvoyant à une situation représentée en mémoire augmente à mesure que de nouvelles propositions lui sont ajoutées du fait de la constance des UTs agrégées amenées par les contextes de celles-ci. Au contraire, celui d'une UT inédite ne montre aucune tendance à la progression et reste bas du fait de l'hétérogénéité des UTs agrégées amenées par chaque proposition intégrée par rapport à celles déjà présentes dans le contexte de l'UT. Pour déclarer qu'une UT en construction est inédite, on vérifie donc également que son niveau de stabilité se situe en dessous d'un seuil donné.

Lors de l'analyse d'un texte, dans le cas où toutes les valeurs de similarité entre le contexte de la proposition courante et le contexte des UTs en construction sont inférieures à $S_{nouvUTC}$, on détermine si le contexte de la proposition est ou n'est pas significatif. Dans l'affirmatif, on sait que l'on se trouve dans le cas de l'apparition dans le texte d'une nouvelle UT pour laquelle la mémoire épisodique possède des UT agrégées proches. On peut donc créer une nouvelle UT en construction avec cette proposition. Dans le cas contraire, on passe en revue toutes les UTs en construction afin de leur appliquer les critères de détection d'une UT inédite. Si une telle UT est reconnue, la proposition courante y est intégrée. Autrement, elle donne lieu à la création d'une nouvelle UT en construction.

Il convient de remarquer qu'il n'est pas possible de faire ici la différence entre plusieurs éventuelles situations inédites puisque par définition, les connaissances nécessaires à une telle différenciation ne sont pas disponibles. Si l'on détecte une proposition susceptible d'appartenir à une situation inédite, on ne peut en effet créer de nouvelle UT inédite que si une telle UT n'a pas déjà été détectée comme telle au sein de la liste des UTs en construction.

Il faut préciser toutefois que l'amorçage de l'analyse thématique de MLK par celle de SEGAPSITH est justement destiné à pallier cette insuffisance. On se reportera au paragraphe 3.2 du chapitre 10 pour avoir une description de la façon dont cet amorçage peut être réalisé.

Finalement, le processus d'analyse thématique de MLK se résume par l'algorithme de la figure 8.3.

```
Pour toutes les propositions du texte (PC) faire
    phase de rappel : définition du contexte de PC
    simMax ← 0
    Pour toutes les UTs en construction (UTC) faire
        valSim ← sim(contexte(PC),contexte(UTC))
        Si (valSim > simMax) alors
            simMax ← valSim
            UTrattachement ← UTC
        Fin_si
    Fin_pour
    Si (simMax < S_nouvUTC) alors
        Si estInédite(PC) alors
            UTC ← élémentSuivant(ListeUTsEnConstruction)
            Tantque ((UTinédite = nil) et (UTC <> nil)) faire
                Si estInédite(UTC) alors
                    UTinédite ← UTC
                Fin_si
                UTC ← élémentSuivant(ListeUTsEnConstruction)
            Fin_tantque
            Si (UTinédite <> nil) alors
                intégration(PC, UTinédite)
            Sinon
                nouvelleUTC ← créationNouvelleUTEnConstruction
                intégration(PC, nouvelleUTC)
            Fin_si
        Sinon
            nouvelleUTC ← créationNouvelleUTEnConstruction
            intégration(PC, nouvelleUTC)
        Fin_si
    Sinon
        intégration(PC, UTrattachement)
    Fin_si
Fin_pour
```

Fig. 8.3 - Algorithme de la segmentation thématique de MLK

À l'issue de la segmentation thématique, l'ensemble des délimitations d'UTs définies sont transformées en une représentation de texte (cf. §3). Cette étape est suivie de la mémorisation de la représentation de texte obtenue, ce qui passe en particulier par la mémorisation de ses UTs. Cette dernière est réalisée en évaluant la similarité de chacune de ces UTs avec les UTs agrégées qui constituent leur contexte. Ces UTs agrégées sont testées suivant l'ordre décroissant de leur poids. L'UT du texte est agrégée avec la première de ces UTs agrégées pour laquelle on trouve une valeur de similarité suffisante.

Si la condition n'est remplie pour aucune d'elles, ce qui peut arriver notamment pour les UTs inédites, l'UT est mémorisée en tant que nouvelle UT agrégée, suivant le processus décrit au chapitre 6.

3. La construction des représentations de texte

En faisant apparaître les différentes UTs d'un texte et en déterminant leur contenu, la segmentation thématique définit à la fois l'ossature et les éléments de base des représentations de texte. La construction de ces dernières ne s'arrête cependant pas à ce stade. Quatre autres tâches doivent effet être réalisées pour la mener à bien de façon complète :

- la détermination des relations thématiques existant entre les UTs;
- la définition du statut de chaque UT en tant qu'UT principale ou secondaire;
- la répartition des propositions de chaque UT entre ses différents attributs;
- la mise en évidence des relations de nature temporelle et causale existant entre les propositions des UTs.

Les deux premières tâches forment la partie de l'analyse thématique de MLK dédiée au suivi de l'évolution thématique. Elles permettent de structurer les représentations de texte à un niveau où les UTs sont considérées comme des briques de base. Les deux dernières tâches s'attachent au contraire à la structuration interne des UTs. À ce titre, elles sortent un peu du champ de notre étude puisqu'elles sont plutôt du ressort de processus d'analyse temporelle et causale. Nous évoquerons néanmoins quelques pistes envisageables les concernant.

3.1. Le suivi thématique

Des algorithmes comme ceux de Grosz et Sidner ou de Grau ne dissocient pas la segmentation du suivi thématique. Cette conception est applicable lorsque les textes traités se conforment à un modèle du discours pré-établi, par exemple une structure hiérarchique emboîtant les segments les uns dans les autres comme des poupées russes, et/ou lorsqu'il est possible de se reposer sur un ensemble de connaissances à la fois stables et précises. Compte tenu de nos hypothèses de travail, il apparaît préférable dans notre cas de réaliser les tâches de suivi thématique à l'issue de la segmentation. Ceci nous permet en effet d'appuyer nos jugements en la matière sur une vue d'ensemble du découpage thématique

réalisé. Précisons qu'à ce stade de notre travail, nos propositions restent embryonnaires : elles se limitent à quelques heuristiques simples et à l'ébauche de voies à explorer.

3.1.1. Les relations entre les Unités Thématiques

Pour déterminer quelles sont les relations thématiques existant entre les UTs produites par la segmentation, nous nous appuyons principalement sur la façon dont les différentes UTs se manifestent dans la séquence des propositions formant le texte. Pour cela, nous appliquons les deux principes suivants :

- si une ou plusieurs propositions, notée(s) Prop A, appartenant à une UT A apparaissent dans la séquence des propositions du texte entre deux propositions appartenant à une UT B, notées Prop B, on en déduit qu'il existe une relation de déviation thématique entre l'UT A et l'UT B. L'UT source de la déviation est dans ce cas l'UT s'étant manifestée la première dans le texte.

(1)		(2)	
	Prop A-1		Prop B-1
	Prop A-2		Prop B-2
	Prop B-1		Prop A-1
	Prop B-2		Prop A-2
	Prop A-3		Prop B-3
	Prop A-4		Prop A-3

  Le cas (1) est le plus classique : un segment correspondant à la totalité de l'UT B est inclus dans un autre segment, correspondant lui à l'UT A. On considère alors qu'il existe une relation de déviation thématique allant de l'UT A vers l'UT B.
  Le cas (2) est un peu plus complexe : le thème de l'UT A et celui de l'UT B sont développés en même temps, par allers et retours successifs. Il s'agit d'une manifestation du style enchevêtré que nous avons évoqué précédemment et dont nous avons eu un aperçu au travers du texte de la figure 8.1. On considère ici qu'il existe une relation de déviation thématique allant de B vers A.
  Quel que soit le cas de figure, la proposition de l'UT source retenue pour être le point de départ précis de la déviation est la dernière proposition de cette UT apparaissant avant la première proposition de l'UT déviation. Il s'agit ainsi de A-2 dans le cas (1) et de B-2 dans le cas (2);

- si la première proposition d'une UT A vient immédiatement à la suite de la dernière proposition d'une UT B, on infère que les UTs A et B sont liées par une relation de changement de thème. C'est le cas illustré ci-dessous, A-n étant la dernière proposition de l'UT A :

(3) Prop A-1
...
Prop A-n
Prop B-1
...

Même s'ils permettent d'établir des relations thématiques pertinentes dans un certain nombre de cas, ces deux principes n'ont pas la prétention de fournir une bonne solution de façon systématique. C'est particulièrement vrai pour le second, qui se présente plutôt comme une sorte de règle par défaut. Si une UT B se rapporte à la dernière proposition d'une UT A, elle est développée dans le texte à la suite des propositions de A et conduit à une configuration semblable à (3). Cette dernière est alors interprétée comme un changement de thème au lieu d'être reconnue comme une déviation. Une telle erreur pourrait en l'occurrence intervenir entre l'UT *TentativeAssassinat* et l'UT *Hôpital* de la représentation de texte de la figure 5.5 (cf. chapitre 5).

Le premier principe se trouve lui aussi contredit dans le cas des phénomènes d'interruption du discours : le développement d'un sujet est interrompu brutalement par une intervention portant sur un thème qui n'est pas lié au précédent et qui ne sera pas réabordé ultérieurement. Nous sommes alors en présence d'un changement de thème et non d'une déviation comme tendrait à l'indiquer le premier principe. Ces interruptions se rencontrent surtout dans les dialogues et sont peu fréquentes dans les textes écrits.

Ces deux contre-exemples montrent qu'il est nécessaire de compléter les deux principes présentés par d'autres moyens. Les résultats de la segmentation thématique nous en offre au moins un. À l'issue de cette opération, chaque UT construite se retrouve dotée d'un contexte caractérisant cette UT du point de vue de la mémoire épisodique. Par définition, on peut considérer que les deux UTs impliquées dans une relation de déviation sont plus proches d'un point de vue thématique que deux UTs impliquées dans une relation de changement de thème. L'UT cible de la déviation détaille en effet une partie de l'UT qui en est la source. Cette différence devrait a priori se retrouver au niveau des contextes associés aux UTs. Dans une configuration telle que (3) ou (1), on mesurerait ainsi la similarité entre le contexte de l'UT A et celui de l'UT B. Si la similarité trouvée est trop faible, on conclurait en faveur d'un changement de thème et dans le cas contraire, on opterait alors pour une déviation.

Un procédé similaire pourrait d'ailleurs être utilisé afin de déterminer précisément la proposition faisant l'objet d'une déviation. Nous avons proposé ci-dessus que celle-ci soit assimilée à la dernière proposition de l'UT source de la déviation précédant la première proposition de l'UT qui en est la cible. Il ne s'agit cependant que d'une heuristique, qui pourrait être complétée par une mesure de la similarité entre le contexte de la proposition en question, en l'occurrence celui établi lors du traitement de cette proposition par le processus de segmentation, et le contexte final associé à l'UT incarnant la déviation. Les valeurs de cette mesure pour plusieurs propositions candidates

pourraient ainsi être comparées. Le principal inconvénient de ce procédé est l'obligation qu'il impose de conserver une trace du contexte établi lors de la segmentation pour chaque proposition d'un texte.

Le dernier moyen envisageable de lever une éventuelle ambiguïté entre une relation de déviation thématique et une relation de changement de thème consiste à utiliser la mémoire épisodique en tant que base de cas. Si l'on trouve en effet une relation de déviation assez récurrente entre deux UTs agrégées de la mémoire similaires aux deux UTs du texte considérées, il paraît raisonnable d'opter en faveur d'une relation de déviation. Le raisonnement à base de cas ne peut néanmoins être utilisé qu'à la condition de disposer déjà d'un certain nombre de cas de référence. Il doit donc obligatoirement être amorcé. Les deux procédés présentés auparavant constituent un bon moyen de fournir un ensemble de cas de départ, la mémoire épisodique se chargeant sur un certain terme de faire le tri entre les relations significatives et celles qui ne le sont pas.

3.1.2. Le statut des Unités Thématiques

Ainsi que nous l'avons souligné au §3.2 du chapitre 5, la détermination du statut d'une UT dépend pour l'essentiel de sa situation dans le réseau de relations thématiques présent dans une représentation de texte. On retient ainsi comme UT principale toute UT qui n'est pas la cible d'une relation de déviation thématique et qui n'entretient de relation de changement thématique, si elle en entretient, qu'avec d'autres UTs principales. Toutes les autres UTs sont marquées comme UTs secondaires.

3.2. La structuration des Unités Thématiques

La répartition des propositions d'un texte entre les différents attributs d'une UT et la mise en évidence des relations existant entre ces propositions ont besoin de s'appuyer sur des moyens d'analyse temporelle et causale prenant en compte l'absence de connaissances de référence. Les moyens utilisables dans notre cadre de travail nous apparaissent être les suivants :

- les travaux reposant uniquement sur la forme des textes, c'est-à-dire sur un ensemble de marques linguistiques de surface (connecteurs, type des verbes, temps des verbes, etc.). Nous avons déjà évoqué au §3.3.2 du chapitre 5 l'utilisation possible des travaux de ce type – plus spécifiquement ceux concernant la détermination de la valeur aspectuo-temporelle des énoncés – pour la répartition des propositions entre les différents attributs des UTs. Leur utilisation pour la mise à jour des relations temporelles entre les propositions est donc tout à fait

envisageable. Par ailleurs, des travaux de même nature portant sur l'extraction de relations causales par ce type de moyens peuvent également être exploités [Jackiewicz 1996]. Globalement, il faut souligner que ces méthodes donnent des indications sérieuses mais plus rarement des certitudes. Elles contribuent donc à introduire une certaine incertitude dans les représentations de texte construites;

- le raisonnement à base de cas à partir du contenu de la mémoire épisodique. Comme pour les relations thématiques, l'idée est d'exploiter un raisonnement à base de cas plus "classique" que celui mis en œuvre lors de la segmentation. En l'absence de connaissances de référence sur le domaine, ce raisonnement ne peut relever pour l'essentiel que d'une approche syntaxique, c'est-à-dire fondée sur la similitude de structure des cas. On décide ainsi d'affecter une proposition à un attribut plutôt qu'à un autre en se fondant son affectation dans d'autres UTs similaires ou l'on établit une relation entre deux propositions sur la base de sa présence répétée également dans d'autres UTs similaires. Bien entendu, l'utilisation de ce type de raisonnement se heurte toujours au problème de l'amorçage : il faut qu'un autre mécanisme puisse produire les premiers cas. Une intervention humaine dans ce sens pourrait être envisagée mais en pratique, elle se limitera toujours à un domaine restreint;

- l'utilisation de connaissances temporelles et causales apprises dans un cadre différent. Nous avons vu au chapitre 2, avec l'apprentissage de type TDL (Theory-Driven Learning) proposé par Pazzani, que des travaux existent sur l'apprentissage automatique de relations causales élémentaires à partir de textes. L'utilisation de techniques de détection des relations temporelles et causales telles que celles évoquées au premier point conjuguée à celle de méthodes d'apprentissage est également une voie intéressante pour acquérir ce type de connaissances. Celles-ci peuvent ensuite être utilisées pour la structuration des UTs au même titre que des connaissances fournies a priori. La seule restriction à cet usage réside dans la nécessité de déterminer quand ces connaissances, dont la formation se déroule en parallèle de celle de la mémoire épisodique, sont suffisamment sûres pour être exploitées dans notre cadre.

Parmi ces trois types de moyens, le premier est sans doute celui qui doit être privilégié à court terme. Pour chaque nouveau domaine abordé, il sert en effet de support à la mise en place des deux autres. Ceux-ci peuvent ensuite prendre son relais en produisant des résultats plus assurés, fruit de l'accumulation des textes traités. Rappelons pour finir, à la suite du §3.3.2 du chapitre 5, que les connaissances sémantiques sur les actions, lorsqu'elles sont suffisamment élaborées, recèlent des connaissances de nature causale permettant de réaliser des inférences quasi-automatiques à propos des conséquences immédiates des actions. En l'occurrence, ces connaissances sont également utilisables

afin de mettre en évidence les relations de cette nature intervenant entre les propositions des UTs.

4. Discussion et extensions

Le premier point à souligner concernant l'analyse thématique de MLK est bien entendu la nécessité d'implémenter celles, parmi ses composantes, dont le degré de spécification est suffisamment élevé pour franchir cette étape. C'est le cas plus particulièrement de la segmentation thématique, laquelle a d'ailleurs déjà été implémentée dans le cadre de ROSA sous une forme un peu simplifiée, ainsi que nous le verrons au chapitre 10. Au niveau de MLK, les seuls points restant à préciser sont les valeurs de certains paramètres numériques comme la taille de la fenêtre texte ou bien encore celle des contextes. On pourra s'inspirer dans un premier temps de certaines leçons tirées de ROSA mais seuls des tests réalisés avec l'algorithme précis de MLK permettront de leur donner une valeur adéquate. D'autre part, le comportement de cet algorithme dans certaines conditions limites doit être testé afin de procéder à d'éventuels ajustements. C'est le cas des débuts et des fins de texte ou encore des UTs au tout début de leur formation, lorsqu'elles ne regroupent qu'une seule proposition.

L'état d'avancement des autres dimensions de la construction des représentations des textes est plus limité. Les principes relatifs à la détermination des relations thématiques et au statut des UTs sont fixés mais il reste encore à en déduire des algorithmes précis, même si cette tâche semble assez directe. Le seul point flou à propos de ces principes réside dans l'utilisation possible du raisonnement à base de cas mais ceci doit être envisagé comme une extension. En revanche, tout ce qui a trait à la structuration interne des UTs reste encore à définir précisément. Nous ne donnons ici que des pistes dont l'exploration demandera encore beaucoup de travail. Du point de vue de l'analyse thématique, cette partie n'est cependant pas primordiale et n'empêcherait pas de mener des tests poussés sur la segmentation et le suivi thématiques.

Ces tests se heurtent comme précédemment à la nécessité d'une intervention humaine pour réaliser l'analyse sémantique des textes. Toutefois, même avec un jeu de test restreint constitué manuellement, il devrait être possible de vérifier quelques propriétés de base de l'analyse thématique. En particulier, on pourra essayer de juger de sa capacité d'auto-cohérence, c'est-à-dire examiner si les textes dont les représentations ont permis de construire la mémoire épisodique se retrouvent segmentés de la même façon que dans la représentation qui en a été mémorisée. Précisons que pour mener de tels tests, il est

nécessaire de compléter la mise en œuvre de la propagation d'activité en même temps qu'il faut implémenter l'analyse thématique définie dans ce chapitre.

À côté de la mise en œuvre et du test de l'analyse thématique sous sa définition actuelle, son application à un texte tel que celui relatant la tentative d'assassinat de Martin Luther King laisse entrevoir quelques améliorations possibles. Celles-ci visent en particulier à éviter que les propositions les moins spécifiques sur le plan thématique, donc les plus difficiles à rattacher à l'UT dont elles devraient faire partie, ne donnent lieu à la création de nouvelles UTs ou ne se rattachent toutes à une UT inédite. Le but n'est pas pour autant d'introduire un modèle de structuration a priori des textes car nous ne souhaitons pas perdre la flexibilité du mécanisme actuel.

La solution que nous préconisons consiste à offrir la possibilité de différer temporairement la décision de rattachement d'une ou de plusieurs propositions. Comme nous le verrons au chapitre 10, cette solution va dans le sens également retenu pour la segmentation de SEGAPSITH : un changement de thème n'y est en effet confirmé qu'après une certaine période d'observation.

Dans le cas présent, la procédure pourrait être la suivante : lorsqu'une proposition P_{deb} se rattachant sans ambiguïté à une UT en construction déjà existante est suivie d'une proposition identifiée comme ayant un contexte non significatif, on traite les propositions qui la suivent jusqu'à rencontrer une proposition P_{fin} au contexte significatif. Si celle-ci se rattache à la même UT que P_{deb} et que L_{interv}, le nombre de propositions séparant P_{deb} de P_{fin}, n'est pas trop important (seuil S_{inedit}, à définir), on fait l'hypothèse que ces propositions se rattachent à la même UT que P_{deb} et P_{fin}. En revanche, si L_{interv} dépasse S_{inedit}, on suppose que les propositions correspondantes font référence à une situation inédite dont on recherche une représentation parmi les UTs en construction. Si une telle UT n'existe pas, une nouvelle UT en construction est créée pour contenir ces propositions. Une incertitude persiste dans le cas où L_{interv} reste inférieur à S_{inedit} mais que l'UT de rattachement de P_{deb} est différente de celle de P_{fin}. On n'a pas en effet de moyen de déterminer à laquelle des deux UTs en question les propositions doivent être rattachées. Seuls des tests sur un ensemble diversifié de textes peuvent valider les choix réalisés en offrant dans le même temps des bases suffisantes pour prolonger ces mécanismes et lever éventuellement ce genre d'indécisions.

Dans ce dernier cas, mais aussi plus généralement, on peut essayer d'exploiter la nature des propositions afin d'évaluer leur potentialité à marquer un changement de thème. La nature d'une proposition fait référence dans ce cas à son rôle dans la phrase, notamment du point de vue syntaxique. Le fait qu'une proposition soit principale ou subordonnée, son type, si c'est une subordonnée, ou le type des subordonnées qui

l'environnent, si c'est une principale, ne sont sans doute pas des informations indifférentes du point de vue de la segmentation. Si une proposition n'est pas thématiquement marquée mais qu'elle est par exemple une subordonnée relative associée à une proposition que l'on sait rattacher à une UT en construction, il est probable qu'en rattachant cette proposition à la même UT, on tombe juste dans une grande majorité des cas. Ceci reste néanmoins un sujet d'étude à explorer de façon plus systématique.

La dernière extension que nous aborderons a trait aux rapports entre l'analyse thématique de MLK et celle que l'on peut mener en utilisant les schémas de la mémoire pragmatique, à la façon de [Grau 1983]. Suivant la logique de l'emploi privilégié des méthodes fondées sur les connaissances les plus sûres et les plus générales, la préférence sera donnée à l'analyse thématique exploitant la mémoire pragmatique lorsque des schémas correspondant aux situations évoquées existent en son sein. Dans le cas contraire seulement, on fait appel à l'analyse thématique de MLK. Néanmoins, la situation est rarement strictement binaire. La mémoire pragmatique peut ne contenir qu'une partie des schémas nécessaires à l'analyse complète du texte considéré. Selon l'étendue du manque, on peut penser faire appel à l'un ou l'autre mécanisme d'analyse en lui adjoignant au besoin les services de celui n'ayant pas été retenu.

La collaboration directe des processus d'analyse est cependant difficile dans la mesure où ils n'obéissent pas à la même logique : l'analyse de MLK dissocie la segmentation thématique du suivi alors que celle proposée par Grau unit ces deux dimensions dans un même algorithme. En revanche, une collaboration plus indirecte est envisageable par le biais des connaissances.

Du point de vue de l'analyse de MLK, les schémas peuvent ainsi être utilisés de la même façon que les UTs agrégées dans les contextes des propositions et les contextes des UTs en construction. Il suffit pour cela que le mécanisme de rappel par propagation d'activité soit étendu à la mémoire pragmatique. Ceci ne pose pas de problème particulier puisque des poids sont associés aux relations entre les schémas dans cette optique. La seule opération à réaliser consiste donc à élargir le réseau de propagation d'activité en reprenant dans ses grandes lignes la structure du réseau associé à la mémoire épisodique puisque les schémas sont structurellement très proches des UTs agrégées.

L'intérêt de cette utilisation des schémas est néanmoins assez limité, surtout si l'on choisit de conserver les UTs agrégées à partir desquelles les schémas ont été construits. En fait, les schémas pourraient s'avérer surtout utiles lors de la structuration des représentations de texte. Par les liens qu'ils entretiennent entre eux, ils permettent en effet de reconnaître facilement les relations de déviation thématique sans avoir à les reconstituer à partir de la mémoire épisodique suivant un processus d'abstraction complexe et coûteux.

L’utilisation des UTs agrégées par le processus d’analyse thématique reposant sur la mémoire pragmatique apparaît plus prometteur. Elle permettrait en effet de suppléer à l’absence de certains schémas, sans lesquels l’analyse deviendrait impossible à ce niveau. Rappelons que pour l’essentiel, cette analyse consiste à sélectionner un schéma pour chaque proposition du texte considéré et à déterminer quelle est la nature du lien unissant ce schéma aux schémas présents dans le contexte courant. Celui-ci rassemble les thèmes pouvant être développés à un moment donné de l’analyse en considérant qu’il s’agit soit d’une déviation thématique, soit de la simple confirmation du thème actif. Si aucun lien n’est trouvé, on opte pour un changement de thème.

Dans ce cadre, les UTs agrégées permettraient de représenter un thème lorsqu’il n’est pas possible pour ce faire de sélectionner un schéma représentatif dans la mémoire pragmatique. Le thème courant associé à la proposition en cours de traitement serait alors matérialisé par un vecteur d’UTs agrégées de la mémoire épisodique établi par le rappel initié à partir de la proposition, de la même façon que dans l’analyse de MLK. Ce vecteur, mémorisé dans le contexte courant, serait ensuite mis à jour lors des évocations ultérieures de ce thème en fusionnant son vecteur dans le contexte courant avec le vecteur associé à la proposition réintroduisant ce thème, de la même façon que le contexte d’une UT en construction est actualisé par le contexte d’une nouvelle proposition.

Si elle est retenue, cette pluralité des formes de représentation des thèmes obligera à mettre en œuvre deux techniques différentes pour juger de la nature du lien entre la proposition traitée et le contexte courant, ou plus précisément l’ensemble des thèmes qui le composent. Lorsque l’un de ces thèmes sera représenté par un schéma, on suivra la procédure habituelle de recherche d’un lien au travers du réseau de schémas. Au contraire, lorsqu’il sera représenté par un vecteur d’UTs agrégées, on adoptera une procédure similaire à celle de l’analyse de MLK : on établit le contexte de la proposition grâce à une phase de rappel des UTs agrégées de la mémoire épisodique et l’on compare ce vecteur à celui représentant le thème considéré par l’intermédiaire d’une mesure de similarité. Si la valeur de similarité est suffisamment forte, on estime qu’il existe effectivement un lien entre la proposition et le thème et l’on opte en faveur d’une déviation. Au contraire, si cette valeur est trop faible, on se prononce dans le sens d’un changement de thème.

Récapitulatif

Dans les chapitres précédents, nous avons exposé la partie de MLK liée à l'apprentissage de connaissances. Ce chapitre nous a permis de présenter son complémentaire, c'est-à-dire la partie de MLK liée à l'analyse des textes. Cette analyse, essentiellement de nature thématique, vise à construire les représentations de texte décrites au chapitre 5, représentations servant de briques élémentaires pour la construction de la mémoire épisodique décrite au chapitre 6.

Compte tenu de la tâche à réaliser, nous avons commencé par faire l'examen de la notion de thème en linguistique. Ce tour d'horizon nous a conduit à constater que l'analyse thématique, dans un sens proche de celui où nous l'entendons ici, ne possède pas de statut véritable dans cette discipline. À défaut d'une définition bien établie, nous avons repris à notre compte la définition intuitive stipulant que l'analyse thématique consiste à mettre en relation des parties de discours avec des connaissances de référence représentant les thèmes que ces unités textuelles sont supposées évoquer. Autour de cette problématique, on met plus spécifiquement en évidence trois tâches : la segmentation en unités textuelles thématiquement homogènes, la caractérisation de l'enchaînement des thèmes, appelée suivi des thèmes, et l'identification du thème d'une unité textuelle. Dans le cas de MLK, seules les deux premières nous intéressent explicitement. La troisième est implicitement réalisée lors de la mémorisation des représentations de texte.

Les travaux concernant les deux premières tâches sont assez nombreux et divers. Dans ce chapitre, nous avons laissé de côté tous ceux relevant d'une approche quantitative opérant au niveau lexical pour ne retenir que ceux intervenant au niveau conceptuel. Nous avons plus précisément détaillé le système d'analyse thématique élaboré par Grau [Grau 1983] ainsi que le modèle plus général de Grosz et Sidner [Grosz & Sidner 1986] sur la structuration du discours. Dans les deux cas, l'analyse d'un texte est fondée sur l'utilisation conjointe de connaissances sur le domaine et d'attentes portant sur la façon dont un texte doit être structuré. Cette utilisation est contrôlée par l'intermédiaire d'un mécanisme de focalisation de l'attention. Les tâches de segmentation et de suivi thématique se trouvent par ailleurs étroitement intriquées l'une dans l'autre.

Le cadre de travail dans lequel nous nous plaçons se distingue de ces travaux sur deux points importants. D'une part, les connaissances sur le domaine y sont incertaines et incomplètes. D'autre part, les textes que nous considérons, notamment des textes journalistiques, se caractérisent par un style très entremêlé qu'il est assez difficile de faire entrer dans des structures pré-définies, en particulier celles qui sont habituellement manipulées. Ces caractéristiques nous ont conduit à définir une analyse thématique dans

laquelle la segmentation, qui délimite les UTs, et le suivi thématique, qui détermine les relations thématiques et le rôle des UTs, sont séparés. La première est fondée entièrement sur les connaissances existant au sein de la mémoire épisodique à propos des situations. Elle prend en compte par des mécanismes spécifiques l'incomplétude et l'incertitude de ces connaissances. Le second exploite la manière dont les propositions des segments de texte définis par la segmentation apparaissent dans les textes les uns par rapport aux autres pour déterminer la façon dont ces segments s'articulent les uns avec les autres.

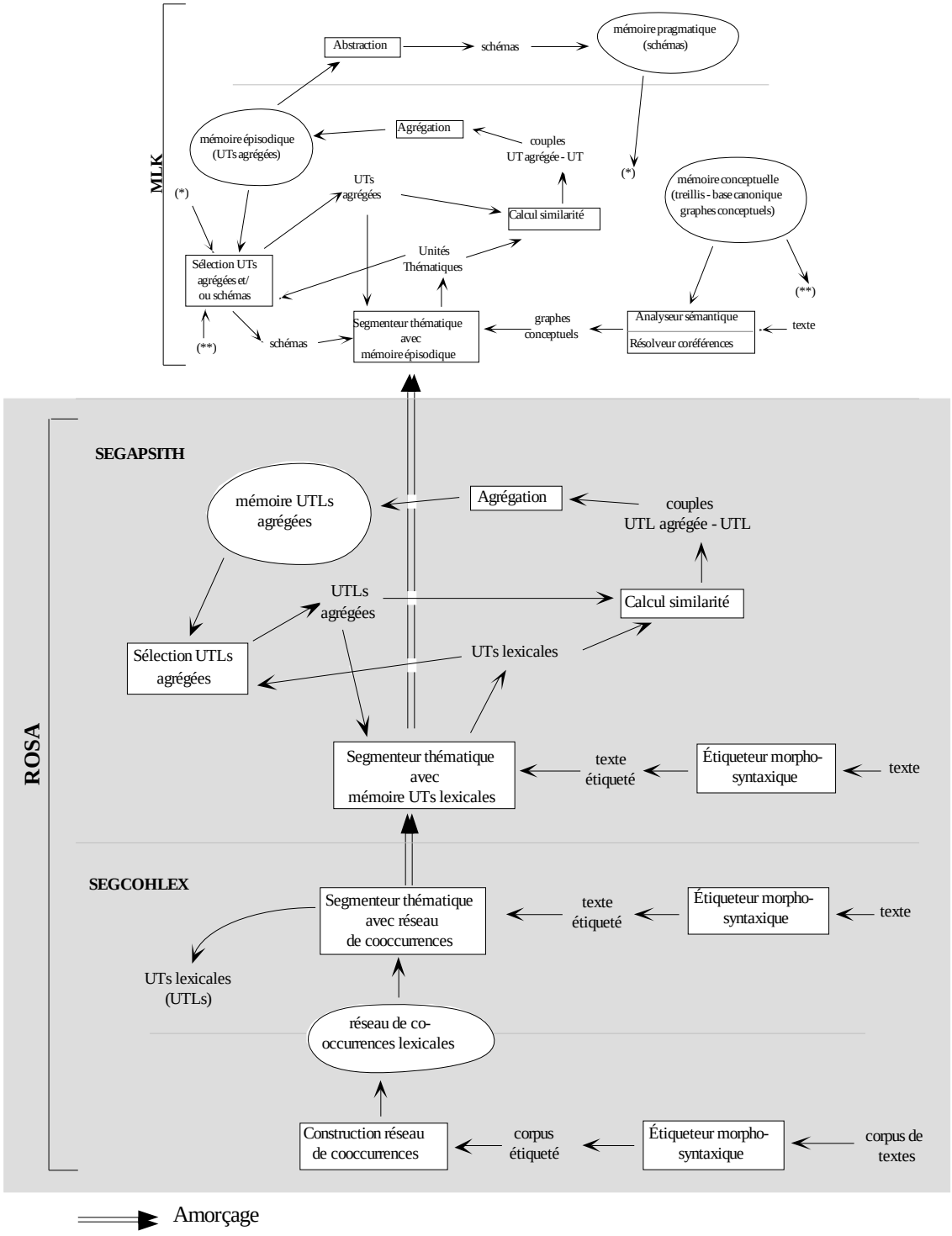
Afin de contrebalancer les caractéristiques des UTs agrégées représentant les situations dans la mémoire épisodique, la segmentation thématique de MLK fait le choix de les exploiter sous forme de configurations et non individuellement comme cela se ferait avec des connaissances plus sûres. Cette exploitation se manifeste plus précisément par l'attribution d'un contexte à la fois aux UTs en cours de construction et aux propositions des textes. Par l'entremise du mécanisme de rappel de la mémoire épisodique, le traitement de chaque proposition donne lieu à la sélection d'un ensemble d'UTs agrégées lui étant supposées proches. Le contexte de la proposition est constitué par cet ensemble d'UTs agrégées. Celui d'une UT en construction est le résultat de la fusion des contextes associés aux différentes propositions qui lui ont été rattachées. Les propositions et les UTs en construction étant caractérisées par des contextes de même type, on évalue la possibilité de rattacher une proposition à une UT en construction en calculant la similarité entre leurs deux contextes respectifs.

Le traitement d'une nouvelle proposition consiste à tenter de la rattacher par le biais de cette similarité entre contextes à l'une des UTs en construction existantes. Une nouvelle UT en construction est créée à partir de cette proposition lorsque cette tentative de rattachement échoue. La faible réactivité de la mémoire épisodique permet par ailleurs de détecter les cas d'incomplétude de cette mémoire vis-à-vis de certaines situations évoquées par les textes. Cette propriété est utilisée afin d'éviter la confusion entre un changement de thème et la présence de propositions relatives à une situation inédite du point de vue de la mémoire épisodique.

La segmentation thématique constitue pour le moment l'essentiel de l'analyse thématique de MLK. Pour le suivi thématique, nous nous limitons à proposer quelques heuristiques de base, fondées sur la distribution dans le texte des différentes UTs les unes par rapport aux autres, observée par le biais de la position relative de leurs propositions. On caractérise ainsi le fait que des UTs dont les propositions s'entremêlent ont plus de chances d'entretenir une relation de déviation thématique que deux UTs parfaitement séparées.

Partie III

ROSA


MLK
Abstraction
schémas
mémoire pragmatique
(schémas)
mémoire épisodique
(UTs agrégées)
Agrégation
couples
UT agrégée - UT
(*)
UTs
agrégées
Calcul similarité
mémoire conceptuelle
(treillis - base canonique
graphes conceptuels)
Unités
Thématiques
Sélection UTs
agrégées et/
ou schémas
(**)
schémas
Segmenteur thématique
avec
mémoire épisodique
graphes
conceptuels
Analyseur sémantique
Résolveur coréférences
texte
ROSA
SEGAPSITH
mémoire UTLs
agrégées
Agrégation
couples
UTL agrégée - UTL
UTLs
agrégées
Calcul similarité
Sélection UTLs
agrégées
UTs lexicales
Segmenteur thématique
avec
mémoire UTs lexicales
texte
étiqueté
Étiqueteur morpho-
syntaxique
texte
SEGCOHLEX
Segmenteur thématique
avec réseau
de cooccurrences
texte
étiqueté
Étiqueteur morpho-
syntaxique
texte
UTs lexicales
(UTLs)
réseau de co-
occurrences lexicales
Construction réseau
de cooccurrences
corpus
étiqueté
Étiqueteur morpho-
syntaxique
corpus de
textes
Amorçage

ROSA

Nous abordons ici ROSA, la composante inférieure d'ANTHAPSI (cf. figure ci-contre). ROSA travaille à partir d'une représentation des textes très proche de leur forme de surface. Cette représentation n'est en effet que le produit d'une lemmatisation et d'une désambiguïsation morpho-syntaxique des mots. ROSA est plus précisément chargé de mettre en œuvre les principes présentés dans les chapitres 1 et 3 dans ce contexte de représentations très peu structurées. Il est à cet égard une sorte de symétrique de MLK.

ROSA se divise plus exactement en deux composants : SEGCOHLEX (SEGmentation thématique par utilisation de la COHésion LEXicale) et SEGAPSITH (SEGmentation et APprentissage de SIgnatures THématiques). Le premier constitue une amorce initiale nécessaire à la mise en place du processus d'amorçage entre niveaux. Il n'est donc pas fondé sur les mêmes principes d'amorçage intra-niveau que les autres composants d'ANTHAPSI que sont SEGAPSITH et MLK. Il se présente en fait comme un simple module de segmentation thématique des textes, caractérisé par une grande robustesse et prenant appui sur une forme de connaissances, la cohésion lexicale entre mots, ne faisant pas de différenciation explicite entre les différents types de connaissances (pragmatiques, sémantique, etc.).

SEGAPSITH, au contraire, est véritablement l'image de MLK dans le contexte de faible niveau de structuration et de précision des représentations propres à ROSA. Cette similitude se manifeste par la présence conjointe d'un processus d'émergence de représentations de thèmes et d'un processus de segmentation thématique des textes, tout deux en étroite interaction.

Outre son niveau de représentation, ROSA se différencie de MLK par une double vocation. D'une part, il est le premier niveau d'ANTHAPSI et constitue de ce fait le point de départ de l'amorçage de l'ensemble du système. D'autre part, en raison de son caractère pleinement opérationnel, sans aucune restriction d'échelle comme il peut en exister pour MLK, il constitue un terrain privilégié d'expérimentation et de validation des principes mis en avant au sein d'ANTHAPSI. C'est pourquoi un accent tout particulier a été mis sur les expérimentations aussi bien pour ce qui est de SEGCOHLEX que de SEGAPSITH.

Du point de l'organisation du manuscrit, cette partie ne recouvre que deux chapitres. Le premier est dédié à la description de SEGCOHLEX. On peut lui associer étroitement

l’annexe I pour les lecteurs désireux d’avoir une vue assez générale des méthodes quantitatives de segmentation thématique. Le second chapitre rassemble la présentation de SEGAPSITH ainsi que celle des mécanismes d’amorçage entre SEGCOHLEX et SEGAPSITH d’une part, et entre SEGAPSITH et MLK d’autre part.

Chapitre 9

SEGCOHLEX

Nous commençons la description de ROSA par sa composante de plus bas niveau, SEGCOHLEX. Celle-ci a pour rôle d'assurer une segmentation thématique de nature quantitative, c'est-à-dire applicable à un grand nombre de textes. Cette segmentation est destinée à amorcer la seconde composante de ROSA, SEGAPSITH. L'analyse de l'existant ouvrant ce chapitre montre que l'objectif d'une large couverture suppose d'opérer au niveau lexical, seul niveau pour lequel des ressources largement applicables sont disponibles ou peuvent être apprises. SEGCOHLEX se fonde ainsi sur la cohésion lexicale afin d'estimer la cohérence thématique des différentes parties d'un texte et placer les ruptures thématiques aux endroits de plus faible cohérence. Cette cohésion lexicale est incarnée par une ressource formant le cœur de SEGCOHLEX : un réseau de cooccurrences lexicales constitué à partir d'un vaste corpus.

1. Introduction

1.1. Objectifs et contraintes de SEGCOHLEX

Ainsi que nous l'avons vu au chapitre 3, l'objectif principal de SEGCOHLEX (SEGmentation thématique par utilisation de la COHésion LEXicale) est de réaliser une analyse thématique des textes en vue d'amorcer celle de SEGAPSITH. Cet amorçage consiste à proposer une segmentation des textes lorsque l'analyse de SEGAPSITH ne dispose pas des connaissances, en l'occurrence une représentation des thèmes sous forme d'Unités Thématiques Lexicales (UTL) agrégées, nécessaires à son intervention. Le résultat de cette segmentation est utilisé afin de constituer les UTLs agrégées qui permettront ensuite au mécanisme d'analyse de SEGAPSITH d'être opérationnel sur les textes abordant le thème en question. De ce fait, l'objectif essentiel de la segmentation réalisée par SEGCOHLEX est donc de découper les textes en segments permettant de construire le plus rapidement possible les UTLs agrégées les plus représentatives d'un thème.

Pour être mené à bien, cet objectif ne suppose pas que les segments construits soient forcément les plus précis possibles par rapport à une segmentation de référence, par exemple établie à partir du recoupement de plusieurs jugements humains. En revanche, il impose de définir des segments les plus homogènes possibles sur le plan thématique. Ceci peut amener en particulier à construire des segments plus petits que ceux qui seraient

délimités par un humain et à laisser de côté de petits passages de texte situés entre ces segments et servant uniquement à faire la transition entre un thème et celui qui le suit. Toujours dans la perspective de l'émergence rapide d'une représentation des thèmes la moins bruitée possible, il est par ailleurs préférable de ne pas retenir forcément tous les segments construits à partir des textes mais au contraire de choisir ceux présentant une grande spécificité sur le plan lexical[1] par rapport aux thèmes abordés par les textes en question. Pour cela, il faut donc que la segmentation de SEGCOHLEX soit capable de produire une estimation du degré de cohésion de chaque segment construit.

À ces spécifications provenant de l'objectif général dévolu à SEGCOHLEX, il faut ajouter la contrainte importante de la nature des textes traités. Notre but étant finalement de faire émerger une représentation des situations prototypiques du monde, nous nous sommes intéressé à ce niveau, comme au chapitre 5, à des textes présentant la double caractéristique d'évoquer de telles situations et de le faire dans un style suffisamment narratif pour que la trame événementielle des textes soit bien explicitée. Il nous a semblé encore une fois que les dépêches d'agence de presse constituent la forme de texte répondant globalement le mieux à ces deux critères. Ils présentent également l'avantage d'une relative abondance. Compte tenu de l'absence dans SEGCOHLEX d'intervention manuelle au cours de l'analyse, nous avons pu en outre considérer des textes plus longs que dans MLK : une taille moyenne est d'environ 250 mots avant tout pré-traitement, certains d'entre eux dépassant même assez largement cette moyenne (cf. annexe G pour un exemple de dépêche traitée, sous sa forme originelle).

La contrainte de la nature des textes traités est importante dans la mesure où l'étude des méthodes existantes, ainsi que nous le verrons dans les paragraphes suivants, met en évidence une dépendance entre la nature de la méthode et le type de texte traité. Cette dépendance a été plus particulièrement examinée dans [Ferret et alii 1998]. L'étude réalisée confirme qu'une méthode fondée sur la réitération des mots donne de bons résultats sur les textes techniques mais voit ses performances se dégrader fortement sur des textes tels que ceux pris pour objet ici. Ce phénomène s'explique par l'absence de variabilité d'expression des concepts techniques, trait qui n'est pas partagé par des concepts plus quotidiens[2]. Il semble donc que la méthode de segmentation de

---

[1] Rappelons que les UTLs agrégées sont des ensembles de mots se référant à un même thème et pondérés en fonction de leur importance vis-à-vis de ce thème. On cherche donc à retenir les segments dont le thème apparaît le plus nettement possible au travers des mots qui les composent.

[2] De façon un peu plus surprenante, l'étude en question a montré qu'une méthode utilisant la cohésion lexicale comme source de connaissances donne de moins bons résultats sur des textes techniques qu'une méthode fondée seulement sur la réitération des mots. En effet, les termes techniques ne sont pas représentés au sein de ces connaissances lexicales et se trouvent donc défavorisés comparativement aux autres termes plus communs présents dans ces textes. Le renforcement inadéquat de ces termes communs engendre de fait un bruit qui a tendance à dégrader les performances globales.

SEGCOHLEX doive se situer parmi les méthodes utilisant une source de connaissances permettant au moins de mettre en évidence les relations de cohésion entre les mots.

Nous allons examiner cette nécessité plus en détail en donnant d'abord une vue d'ensemble des méthodes quantitatives de segmentation thématique puis en décrivant deux méthodes spécifiques. La première, TextTiling, est représentative des méthodes fondées sur la seule réitération des mots et présente l'intérêt d'avoir été évaluée finement. Nous exposerons le protocole suivi pour réaliser cette évaluation, protocole que nous avons partiellement repris pour évaluer la segmentation de SEGCOHLEX. La seconde méthode, Lexical Cohesion Profile, illustre pour sa part l'utilisation de connaissances dans le cadre de méthodes quantitatives de segmentation et nous a servi de source d'inspiration pour élaborer le mécanisme de SEGCOHLEX. On pourra trouver par ailleurs à l'annexe I un panorama plus détaillé des méthodes quantitatives de segmentation thématique.

1.2. Vue d'ensemble des méthodes quantitatives de segmentation thématique

La caractéristique des approches dites quantitatives de l'analyse thématique est la possibilité de les appliquer à un large ensemble de textes. Large ensemble ne signifie pas n'importe quel texte. Une méthode est généralement adaptée à un type de textes particulier. Mais en l'occurrence, il s'agit de grands types textuels – textes techniques, expositifs ou narratifs – au sein desquels existe malgré tout une diversité importante. La spécificité des méthodes quantitatives est de pouvoir faire face à cette diversité et donc, de faire preuve par là même d'une grande robustesse.

Cette robustesse est garantie en grande partie par le fait que ces méthodes interviennent au niveau du mot et ne s'appuient pas sur des pré-requis qui ne pourraient être satisfaits que dans un cadre restreint : les opérations utilisées sont souvent plus proches de l'appariement entre chaînes de caractères que de la mise en œuvre d'une analyse syntaxique complète, suivie d'une analyse sémantique.

Un survol rapide des travaux existant dans ce champ de recherche permet de dégager deux courants. Le premier s'affranchit de toute connaissance autre que des connaissances morpho-syntaxiques et réalise la segmentation des textes en se fondant uniquement sur la distribution des mots au sein de ceux-ci. Ce type de méthode présente l'avantage d'être facile à mettre en œuvre, efficace, indépendant de tout domaine et même aisément transposable d'une langue à une autre. En revanche, on peut lui reprocher d'avoir un champ d'application trop étroit. Une telle méthode n'est en effet applicable qu'à des textes

possédant un vocabulaire suffisamment spécifique. Or ce trait est surtout caractéristique des textes techniques.

Cette spécificité du vocabulaire est une condition nécessaire pour qu'une analyse de la répartition des mots ne reposant que sur des relations d'identité lexicale puisse être opérationnelle. Au travers des mots, on cherche en effet à accéder aux concepts et à étudier leur répartition dans un texte. Or, plus un terme est spécifique et plus la relation qu'il entretient avec le niveau conceptuel s'apparente à une relation bi-univoque. Autrement dit, plus un concept est spécifique et moins sa variabilité d'expression est grande. Un concept appartenant à un domaine technique très spécifique est ainsi exprimé par un seul terme technique sans chercher à utiliser un synonyme ou un hyperonyme pour le remplacer. Le terme technique est alors utilisé comme une sorte de sigle. En étudiant sa distribution dans un texte, on étudie par là même celle du concept qu'il désigne. L'analyse thématique s'effectue en s'intéressant plus largement à des configurations de concepts et en s'attachant aux évolutions de ces configurations au sein des textes.

Lorsque les notions abordées par un texte sont plus générales, elles peuvent être exprimées de façon beaucoup plus diversifiée. Dans un article du journal *Le Monde* par exemple, les attaques de paysans français contre des routiers britanniques transportant des moutons sont successivement désignées par les termes "coups de main", "opérations", "attaques", "guérilla" et "guets-apens". Il devient alors nécessaire de disposer de connaissances pour rendre compte des relations entre des mots faisant référence à un même concept. C'est la voie suivie par le second courant de recherche concernant les méthodes quantitatives de segmentation thématique des textes. Suivant la source de connaissances utilisée, les relations entre les mots peuvent recouvrir un champ plus ou moins large. Avec un dictionnaire des synonymes, on se contente de lier les mots faisant référence à un même concept. Avec un dictionnaire traditionnel, on cherche plutôt à rendre compte des relations entre un concept et des concepts plus élémentaires ou plus généraux intervenant dans sa définition, toujours néanmoins en passant par les mots. Un thesaurus renferme quant à lui des relations plus diversifiées : synonymie, hyperonymie, méronymie et jusqu'aux relations d'appartenance à un même contexte. Leur type n'étant pas souvent explicité, ces relations ne sont néanmoins pas toujours faciles à exploiter.

L'utilisation d'une source de connaissances ne représente pas cependant la solution universelle pour réaliser une segmentation thématique pertinente dans la mesure où cette source ne peut pas être elle-même universelle. C'est particulièrement le cas lorsque la segmentation opère dans des domaines spécialisés. Le choix est alors à faire entre des performances peu intéressantes et un investissement important. La première possibilité s'identifie à l'utilisation d'une source de connaissances assez générale. Les performances de la segmentation risquent dans ce cas de ne pas être bonnes à la fois du fait d'un déficit

d'information et d'un excès de bruit. Le déficit vient de ce que les termes utilisés les plus intéressants ne sont pas présents au sein de la source de connaissances du fait de leur trop grande spécificité. Le bruit provient au contraire de mots plus généraux ayant des liens avec les connaissances utilisées mais n'ayant pas de signification particulière par rapport au domaine abordé. La segmentation s'effectue dans ce cas sur la base des informations que ces mots généraux apportent, lesquelles sont non pertinentes vis-à-vis du domaine, voire génératrices de relations tout à fait erronées.

Le second terme du choix correspond à la constitution d'une source de connaissances spécifique. Ce travail étant réalisé le plus souvent de façon manuelle, il demande un investissement important qu'il faut recommencer pour chaque nouveau domaine.

Deux stratégies au moins sont envisageables pour sortir du dilemme exposé. L'une d'elle consiste à faire le deuil d'une méthode universelle et à porter plutôt l'effort sur le choix de la méthode, parmi celles qui existent, la plus adaptée à un texte donné. C'est le parti pris du travail exposé dans [Ferret et alii 1998]. En fonction de la présence ou non des termes les plus représentatifs d'un texte au sein de la source de connaissances utilisée, le système de segmentation thématique choisit d'appliquer une méthode fondée uniquement sur la distribution des mots ou bien une méthode dérivée de la première mais reposant sur cette source de connaissances pour rendre plus évidente la répartition des mots thématiquement liés.

La seconde stratégie consiste à utiliser une source de connaissances adaptable par une procédure d'apprentissage automatique au domaine particulier que l'on considère. Elle est moins souple que la précédente dans la mesure où cette adaptation n'est pas dynamique : elle est réalisée par l'équivalent d'une phase d'entraînement. Cette stratégie est néanmoins plus générale puisque la méthode utilisée est toujours la même et ne nécessite pas la mise en place d'un contrôle que l'on doit modifier à chaque intégration d'une nouvelle méthode spécifique. Nous verrons dans ce qui suit que cette seconde stratégie est celle que nous avons retenue pour SEGCOHLEX.

1.3. TextTiling : une segmentation thématique sans utilisation de connaissances

TextTiling [Hearst 1997] est une méthode, ou plus précisément un ensemble de principes concrétisés sous la forme de plusieurs variantes, développées par Hearst afin de segmenter des textes en unités rendant compte de leurs différents thèmes. Les textes considérés sont des textes expositifs, de type article de magazine ou rapport. Ils sont plutôt longs puisque les plus petits d'entre eux ont une taille d'une quinzaine de paragraphes, soit environ trois pages. Les unités thématiques mises en évidence sont elles

aussi de taille assez importante : le degré de granularité minimal s'établit aux alentours du paragraphe et des unités rassemblant trois à quatre paragraphes sont tout à fait courantes. Le type de texte considéré conjugué à la taille importante des unités thématiques ont une influence sur la nature de celles-ci. Les segments de texte délimités par TextTiling se distinguent les uns des autres sur le plan thématique mais également sur le plan fonctionnel. Certains d'entre eux sont en effet étiquetés "Introduction", "Conclusion" ou "Synthèse".

Quelle que soit la méthode d'analyse de la distribution des mots appliquée par TextTiling, la segmentation se traduit par l'exécution en séquence des trois opérations suivantes :

- normalisation du texte;
- calcul d'une courbe caractérisant l'évolution de la distribution lexicale au travers du texte;
- identification des bornes des segments.

La première opération permet de transformer le texte en une séquence de pseudo-phrases. L'adoption d'une unité arbitraire de taille fixe plutôt que l'utilisation des phrases initiales du texte se justifie par la grande variabilité dans la taille de ces dernières et la nécessité de disposer d'unités de taille fixe pour réaliser des comparaisons pertinentes. Cette transformation commence par la segmentation du texte, suivie par le filtrage des mots grammaticaux et des mots les plus fréquents (stop-list de 898 mots). Les mots sont ensuite lemmatisés puis regroupés en séquences de 20 mots.

La deuxième opération consiste à parcourir la liste des pseudo-phrases construites précédemment et à calculer un score pour chaque frontière entre deux pseudo-phrases adjacentes en fonction des mots que celles-ci contiennent respectivement. Ce score est destiné à rendre compte de la similarité entre les pseudo-phrases. L'hypothèse sous-jacente est que les variations de ce score sont le reflet des variations thématiques présentes dans les textes. On suppose en effet que la récurrence d'emploi de termes identiques est plus forte lorsque l'on se trouve à l'intérieur d'un segment thématiquement homogène que lorsque l'on se situe à cheval entre deux segments.

Trois scores sont proposés. Celui à partir duquel la méthode a été développée originellement repose sur une mesure de type cosinus. Les deux autres sont des extensions : l'un transpose dans le cadre de TextTiling le principe développé par Youmans concernant l'introduction des mots nouveaux tandis que le dernier se fonde sur la notion de chaîne lexicale en adaptant le travail décrit dans [Morris & Hirst 1991]. Les

principes de calcul de ces trois scores sont illustrés par la figure 9.1. Celle-ci fait apparaître huit pseudo-phrases consécutives. Chacune d'entre elles est représentée par une colonne. Les lettres formant chaque colonne figurent quant à elles les mots composant la pseudo-phrase correspondante.

Dans le cas du premier score, les comparaisons ne se font pas directement entre pseudo-phrases mais en adoptant une unité de taille supérieure : le bloc. De même que les pseudo-phrases sont assimilables à des phrases normalisées, les blocs définissent des paragraphes de taille normalisée. Un bloc est formé d'un ensemble de k pseudo-phrases. Dans les expérimentations décrites dans [Hearst 1997], k = 10 mais cette valeur peut être modifiée de manière à s'adapter à des textes plus ou moins longs. Dans l'exemple de la figure 9.1.a, k = 2. Le score considéré est obtenu en évaluant la similarité de deux blocs adjacents de la manière suivante. Soit une frontière f_i entre deux pseudo-phrases. *b1* est le bloc défini par les k pseudo-phrases précédant f_i tandis que *b2* est le bloc défini par les k pseudo-phrases suivant f_i . Le score associé à cette frontière est donné par le cosinus des vecteurs représentant chacun des deux blocs :

$$score(f_i) = \frac{\sum_t w_{t,b1} \; w_{t,b2}}{\sqrt{\sum_t w^2_{t,b1}} \; \sqrt{\sum_t w^2_{t,b2}}}$$

où t énumère les termes retenus à l'issue de la première étape de pré-traitement pour les deux blocs et $w_{t,b1}$ ($w_{t,b2}$) correspond au poids du terme t dans le bloc *b1* (*b2*). Le poids d'un mot dans un bloc est égal ici à son nombre d'occurrences dans le bloc. Cette mesure simple du poids des mots donne en pratique de meilleurs résultats que l'emploi d'un facteur de mise en évidence de la représentativité des mots comme *tf.idf*[1] (utilisé initialement dans [Hearst 1993]).
Ce score est calculé pour toutes les frontières entre pseudo-phrases du texte. Compte tenu de la définition dynamique des blocs, une même pseudo-phrase intervient donc dans le calcul de k*2 scores. La figure 9.1.a illustre donc le résultat de ce calcul seulement pour une moitié des frontières existant entre les huit pseudo-phrases prises comme exemple. Il est à noter que les valeurs données sont celles du produit scalaire des vecteurs. Pour obtenir le cosinus, il suffit de diviser par le produit des normes des vecteurs.

Le deuxième score reprend le principe du VMP (Vocabulary Management Profile) de Youmans (cf. §1.1 de l'annexe I). Pour chaque frontière entre deux pseudo-phrases, on détermine le nombre de mots nouveaux introduits par rapport à l'ensemble du texte déjà traité dans chacune des pseudo-phrases entourant la frontière considérée. Le score associé à cette frontière est donné par la somme de ces deux valeurs, divisée par le nombre de mots rassemblés par les deux pseudo-phrases entourant la frontière. La taille des

[1] cf. paragraphe 1.3 de l'annexe I sur Nomoto et Nitta pour une définition de *tf.idf*

pseudo-phrases étant de 20 mots, on obtient un espace de comptage de 40 mots pour chaque frontière, ce qui est proche des 35 mots de la fenêtre de Youmans. Le pas de déplacement est cependant plus important ici puisqu'il est égal à la taille d'une pseudo-phrase, à comparer au pas de 1 mot de Youmans. L'augmentation de ce pas permet cependant d'obtenir une courbe plus régulière sur laquelle les grandes tendances sont plus facilement analysables. La figure 9.1.b illustre le calcul de ce score pour les mêmes frontières que celles retenues pour la figure 9.1.a.

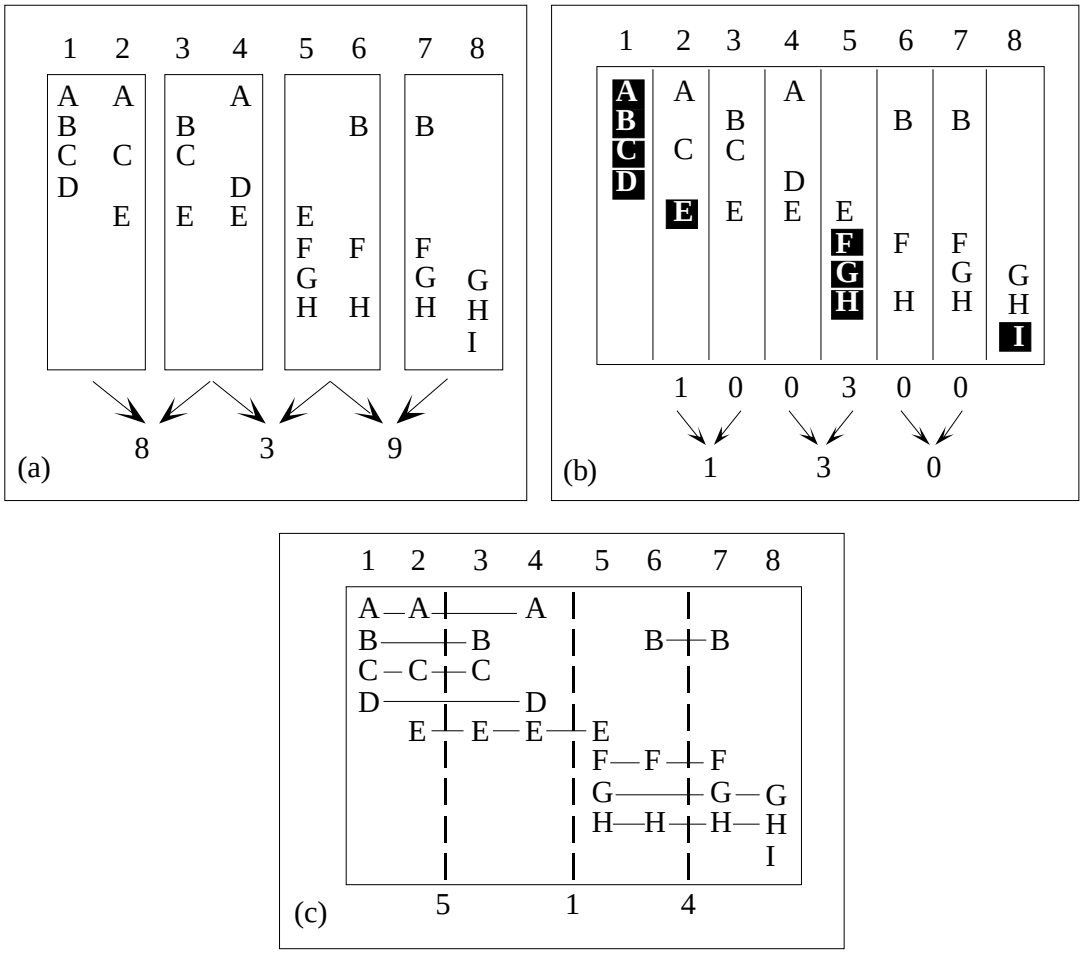


Fig. 9.1 - Principes du calcul des trois scores (adapté de [Hearst 1997])
(a) : produit scalaire; (b) : VMP; (c) : chaînes lexicales

Le dernier score enfin s'inspire des chaînes lexicales de Morris et Hirst[1]. Une chaîne lexicale correspond dans le cas présent à une suite d'occurrences d'un même terme au sein d'un texte, chacune de ces occurrences n'étant pas éloignée de plus d'une distance donnée de celle qui la précède et de celle qui la suit. L'intérêt de ces chaînes du point de vue de la segmentation thématique réside dans l'hypothèse selon laquelle les bornes des segments de discours se situent à des endroits minimisant le nombre de chaînes

[1]cf. paragraphe 2.1 de l'annexe I pour un développement de la notion de chaîne lexicale.

interrompues. Le score calculé par cette méthode s'inspire directement de cette hypothèse. Chaque frontière entre pseudo-phrases est en effet pondérée par le nombre de chaînes lexicales traversant cette frontière. La figure 9.1.c en donne un exemple pour les mêmes frontières que précédemment. La distance maximale entre deux occurrences d'une chaîne n'est pas explicitement mentionnée par Hearst mais il semblerait logique de reprendre une distance égale à 2*k pseudo-phrases (en incluant les bornes). Dans l'exemple donné, elle correspond à 4 pseudo-phrases.

La troisième et dernière opération vise quant à elle à analyser la courbe obtenue à partir de l'ensemble des scores pour détecter les bornes des segments thématiques. Pour obtenir cette courbe, on reporte en abscisse les positions des frontières entre pseudo-phrases et en ordonnée le score associé à chacune de ces frontières. La procédure d'analyse est la même quelle que soit la méthode adoptée pour calculer le score[1].

La première opération de cette procédure est un lissage de la courbe, destiné à supprimer les variations de score trop locales. On utilise pour ce faire une fenêtre de moyennage local des valeurs que l'on centre successivement sur toutes les frontières entre pseudo-phrases. La taille de cette fenêtre est ici de 3 : elle inclut une frontière de part et d'autre de celle que l'on considère. Le lissage s'effectue en affectant un nouveau score à la frontière située au centre de la fenêtre, égal à la moyenne des scores de toutes les frontières contenues dans la fenêtre. La procédure de lissage est appliquée une seule fois.

La deuxième opération consiste à pondérer chaque frontière correspondant à un minimum de la courbe des scores en fonction de son importance. Les minima les plus importants renvoient en principe aux ruptures thématiques les plus importantes tandis que les autres rendent compte de la structuration en sous-thèmes. Pour un minimum donné, appelé *f*, cette pondération est réalisée relativement aux maxima qui l'entourent. En parcourant la courbe de part et d'autre de *f*, on retient comme maxima les deux frontières, *f1* et *f2*, ayant la valeur la plus élevée avant inflexion de la courbe. Le poids de *f* est alors donné par :

$$poids(f) = (y_{f1} - y_f) + (y_{f2} - y_f)$$

La troisième et dernière opération est la formation proprement dite des segments. Elle consiste à effectuer un choix parmi les minima mis en évidence et donc, à déterminer combien de segments seront distingués. Pour ce faire, les minima pondérés par la

[1] Bien que l'application de la même procédure d'analyse pour toutes les méthodes de calcul des scores soit explicitement mentionnée dans [Hearst 1997], cette option paraît difficilement explicable dans le cas de la méthode utilisant le VMP. Pour les deux autres méthodes, les bornes de segment coïncident avec des minima de la courbe alors que dans le cas de cette dernière méthode, il semble plus évident de rechercher des maxima puisque les changements de thèmes doivent se traduire par une augmentation des mots nouveaux. Il y a là un point que nous n'avons pu éclaircir étant donné que [Hearst 1997] est la seule publication exposant l'extension de TextTiling utilisant le VMP.

deuxième opération sont triés selon l'ordre décroissant de leur poids. La fixation du seuil au-dessus duquel les minima sont sélectionnés en tant que bornes de segment s'effectue par rapport à la distribution des poids de ces minima : la valeur la plus restrictive de ce seuil est égale à $\bar{p} - \sigma/2$, avec $\bar{p}$, la moyenne des poids et σ, leur écart -type. Elle permet d'obtenir une meilleure précision mais un moins bon rappel que la valeur $\bar{p} - \sigma$. Pour garantir que les segments formés soient significatifs, on impose enfin que leur taille ne soit pas inférieure à 3 pseudo-phrases.

TextTiling a globalement été évalué de deux manières : d'une part en comparant la segmentation qu'il réalise sur un ensemble de textes avec celle effectuée par des lecteurs humains sur le même ensemble; d'autre part en recoupant ses résultats avec des frontières thématiques marquées dans les textes telles que les frontières de texte.

La première évaluation a été menée sur un ensemble de 12 textes, analysés par 7 lecteurs. Pour mettre en correspondance les jugements humains et les décisions de TextTiling, Hearst s'est inspirée des mesures proposées dans [Passonneau & Litman 1993], [Rosé 1995] et [Carletta 1996], en particulier pour tenir compte de la variabilité entre les jugements humains. Une segmentation de référence a été construite à partir de ces jugements en ne retenant que les bornes corroborées par au moins 3 lecteurs. Le jugement moyen des lecteurs humains, qui correspond à la moyenne des résultats des différents lecteurs par rapport à la segmentation de référence, a été utilisé comme point de référence supérieur pour situer les résultats de TextTiling. Suivant les conseils prodigués dans [Gale et alii 1992], Hearst a également fixé une limite inférieure en prenant les résultats moyens (sur 10 000 lancements) d'un processus fixant au hasard le même nombre de bornes que celles présentes dans la segmentation de référence.

Les résultats obtenus pour les trois variantes de TextTiling se situent globalement toujours au-dessus de la référence inférieure mais pour la presque totalité également en dessous des résultats obtenus par les lecteurs humains. Les indicateurs utilisés ont été la précision et le rappel[1]. La précision se définit ici par le nombre de bornes trouvées qui sont des bornes de référence divisé par le nombre de bornes trouvées. Le rappel correspond quant à lui au nombre de bornes trouvées qui sont des bornes de référence divisé par le nombre total de bornes de référence. Pour la référence inférieure, on obtient une précision et un rappel respectivement égaux à 0,5 et 0,51. Le jugement humain moyen possède en ce qui le concerne une précision de 0,83 et un rappel de 0,71. Les résultats de TextTiling, enfin, s'échelonnent entre des valeurs de 0,52 et 0,71 pour la

[1]Hearst a également introduit le cœfficient Kappa [Carletta 1996], mais seulement pour la méthodefondée sur le cosinus et celle utilisant le VMP. Les chiffres donnés pour la précision et le rappel sont extraits de [Hearst 1997]. Ces résultats ne concernent que les deux méthodes pré-citées. Les résultats relatifs à la troisième méthode, qui peuvent être trouvés dans [Hearst 1994], se situent néanmoins dans les mêmes intervalles, ce qui permet de parler à un niveau général.

précision et entre 0,59 et 0,78 pour le rappel. Ces intervalles recouvrent les valeurs obtenues pour les différentes méthodes de calcul des scores entre pseudo-phrases ainsi que pour différentes valeurs des paramètres de choix des bornes. Un examen un peu plus détaillé montre que la méthode fondée sur le cosinus obtient les meilleurs résultats. La différence est très faible avec les résultats obtenus au moyen de la méthode à base de chaînes lexicales et un peu plus importante pour celle utilisant le VMP.

La seconde évaluation de TextTiling rapportée dans [Hearst 1997] consistait pour sa part à retrouver les frontières d'un ensemble de textes concaténés les uns à la suite des autres. Bien qu'il s'agisse d'un test moins discriminant sur le plan thématique que le premier, il présente l'avantage de pouvoir être réalisé facilement puisqu'il ne nécessite pas l'intervention d'experts humains, dont il faut par ailleurs recouper les avis. L'expérimentationa été menée en l'occurrence sur 44 articles de journaux d'une taille moyenne d'une quinzaine de paragraphes. La mise en correspondance d'une frontière de texte et d'un minimum de la courbe des scores était acceptée avec une tolérance de 3 phrases. Seule la méthode fondée sur la mesure du cosinus a été appliquée. La précision obtenue pour cette tâche a été de 0,59 tandis que le rappel s'est élevé à 0,95.

1.4. Lexical Cohesion Profile : une approche de la segmentation thématique à base de connaissances

La notion de cohésion lexicale sur laquelle repose les chaînes lexicales sous-tend également, mais de manière différente, le travail de Kozima décrit dans [Kozima 1993, Kozima 1993]. Une de ses particularités est de capturer la cohésion lexicale au travers d'une source de connaissances construite automatiquement. Nous commencerons donc par présenter cette source de connaissances avant d'aborder le mécanisme de segmentation qui l'exploite.

1.4.1. La source de connaissances Paradigme

L'objectif spécifiquement dévolu à *Paradigme* est de fournir pour tout couple de mots (m,m´) une mesure de leur cohésion lexicale σ(m,m´), interprétable également en termes de proximité sémantique. L'estimation de cette proximité est réalisée sur la base des informations présentes dans un dictionnaire accessible sous forme électronique. Le dictionnaire est transformé pour ce faire en un réseau lexical auquel on adjoint un mécanisme de propagation d'activation.

Compte tenu de ce mode d'exploitation, il est important que la connectivité du réseau soit forte, donc que la circularité du dictionnaire soit la plus importante possible. Autrement dit, la définition des mots doit être réalisée en ne faisant appel qu'à des mots eux-mêmes définis par le dictionnaire. *Paradigme* a ainsi été construit à partir d'un dictionnaire, appelé *Glossème*, présentant la particularité d'être strictement circulaire. Ce dictionnaire est un sous-ensemble du dictionnaire LDOCE (*Longman Dictionary Of Contemporary English*) [LDOCE 1987], lequel comporte 56000 entrées mais n'utilise pour les définir qu'un vocabulaire de 2851 mots (on compte les racines et non les mots fléchis), formant ce que l'on appelle le *Longman Defining Vocabulary* (LDV). Ce sous-ensemble de la langue anglaise a été délimité à la fois sur le critère du caractère fondamental ou non des mots considérés (en particulier du point de vue de l'enseignement de la langue anglaise aux étrangers) ainsi sur le critère de leur fréquence. *Glossème* est tout simplement le sous-ensemble du LDOCE rassemblant les entrées de ce dernier correspondant aux mots du LDV. Chaque entrée de Glossème est composée d'une liste de définitions, classées dans l'ordre décroissant de leur fréquence d'usage. Une définition est une simple liste de références vers d'autres entrées du dictionnaire.

Le réseau lexical formé par *Paradigme* comporte autant d'unités que de mots du LDV, soit 2851. Chacun d'entre eux est lié aux autres par deux types de liens. Les liens de type *référant* rendent compte des références faites par un mot à d'autres entrées du dictionnaire par l'intermédiaire de ses définitions. Plus précisément, chaque définition d'un mot est incarnée par un *sous-référant*, lequel est lié aux unités du réseau figurant les entrées référencées par cette définition. On a ainsi le schéma suivant :

$$U_{définie} \rightarrow (p1) \rightarrow SousRef \rightarrow (p2) \rightarrow U_{définissante}$$

avec

$U_{définie}$: unité du réseau représentant une entrée E_{Gloss} du dictionnaire *Glossème*;

SousRef : sous-référant incarnant une des définitions, dénommée *Def*, de E_{Gloss} ;

$U_{définissante}$: unité du réseau représentant une des entrées de *Glossème* intervenant dans *Def*;

p1 : poids de la relation entre une unité et l'un de ses sous-référants. Ce poids est calculé en fonction de la position de *Def* dans la liste des définitions de E_{Gloss} (cette position reflétant la fréquence d'usage de la définition). La somme des poids des relations entre une unité et ses différents sous-référants est normalisée pour être égale à 1;

p2 : poids de la relation entre un sous-référant et une unité du réseau représentant une entrée de *Glossème* présente dans la définition associée à ce sous-référant. Ce poids est calculé en tenant compte en particulier de la fréquence de cette entrée au sein de *Glossème*. La somme des poids des relations de ce type pour un sous-référant est également normalisée pour être égale à 1.

Les flèches illustrent le sens de parcours de l'activation.

Le second type de lien existant dans le réseau lexical rend compte de la notion de *référé*. Les référés d'une unité U1 sont toutes les unités U2 du réseau faisant référence à U1 dans le cadre de leur définition. Il s'agit en fait simplement d'un changement de point de vue par rapport à la notion de référant. Une unité U2 ayant pour référant U1 est ainsi elle-même un référé de cette même U1. Le poids de la relation entre U1 et U2 est égal au poids de la relation entre U2 et U1 en appliquant toujours une normalisation telle que la somme des poids des référés d'une unité soit égale à 1.

Paradigme se présente donc comme un réseau de mots liés les uns aux autres par des relations pondérées. L'estimation de la cohésion $\sigma(m,m')$ entre deux mots *m* et *m´* est réalisée de la façon suivante. Le nœud du réseau correspondant à *m* est d'abord activé. Un mécanisme de propagation d'activation est ensuite mis en œuvre pendant un certain nombre de cycles. Ce nombre a été fixé expérimentalement à 10 et correspond au temps moyen de stabilisation de l'activation dans le réseau. Au terme de cette phase de propagation, la valeur de la cohésion est donnée par la valeur de l'activation de *m´*, modulée par la significativité de *m´*.

La mesure n'est donc pas symétrique, d'une part à cause de la modulation, et d'autre part parce que le résultat de la propagation n'est pas forcément le même suivant que c'est *m* ou *m´* qui est initialement activé ($\sigma(m,m') \neq \sigma(m',m)$). La significativité d'un mot correspond quant à elle à son information par rapport à un corpus de référence. Elle est égale à :

$$signif(m) = \frac{-\log(nbOcc(m)/tailleCorpusRéférence)}{-\log(1/tailleCorpusRéférence)}$$

où *nbOcc*(m) est le nombre d'occurrences du mot m dans le corpus de référence.

La fonction d'activation de chaque nœud se contente de sommer les activations entrantes venant à la fois des référants et des référés. La moitié de cette somme est ajoutée à l'activation précédente du nœud considéré et le résultat est seuillé à 1. La mesure de cohésion de deux mots est donc comprise dans l'intervalle [0,1]. La valeur de l'activation initialement injectée dans le réseau est égale à *signif*(m), la significativité du mot *m*.

Lorsqu'un mot ne figure pas dans *Glossème*, on le remplace par l'ensemble des mots constituant son entrée dans le LDOCE. Il n'est plus alors représenté par une seule unité mais par le vecteur d'unités du réseau *M* (m_1, ... m_i ...,m_n). La mesure de cohésion devient ainsi $\sigma(M,m')$ si *m* est absent de *Glossème* et les unités composant l'entrée de *m* dans le LDOCE sont activées avec une force égale à $s(m_i)^2 \Big/ \sum_k s(m_k)$.

Si *m´* est absent de *Glossème*, on mesure l'activation de chacun des mots composant son entrée dans le LDOCE et $\sigma(m,M')$, où *M´* est le vecteur d'unités représentant *m´*, est

alors donnée par la somme, seuillée à 1, de ces activations, chacune d'entre elles étant modulée par la significativité du mot correspondant à l'unité activée. En les conjuguant, ces deux principes permettent également de traiter le cas où à la fois *m* et *m´* sont absents de *Glossème*. La mesure de cohésion s'écrit alors σ(M,M´) et s'applique plus largement à l'évaluation de la cohésion de deux regroupements de mots.

1.4.2. Le Lexical Cohesion Profile et la segmentation des textes

Le paragraphe précédent nous a montré comment la cohésion entre deux mots, et même plus généralement la cohésion d'un ensemble de mots (cas traité de la même façon que le cas d'un mot absent de *Glossème*), peut être évaluée grâce au réseau *Paradigme.* Muni de cet outil, il est alors possible d'évaluer la cohésion lexicale des différentes parties d'un texte. En faisant l'hypothèse que cette cohésion est en rapport avec la cohérence thématique, on peut avancer l'idée que les parties d'un texte présentant une forte cohésion lexicale sont caractérisées par une forte homogénéité thématique tandis que les parties dotées d'une faible cohésion lexicale sont faiblement cohérentes sur le plan thématique.

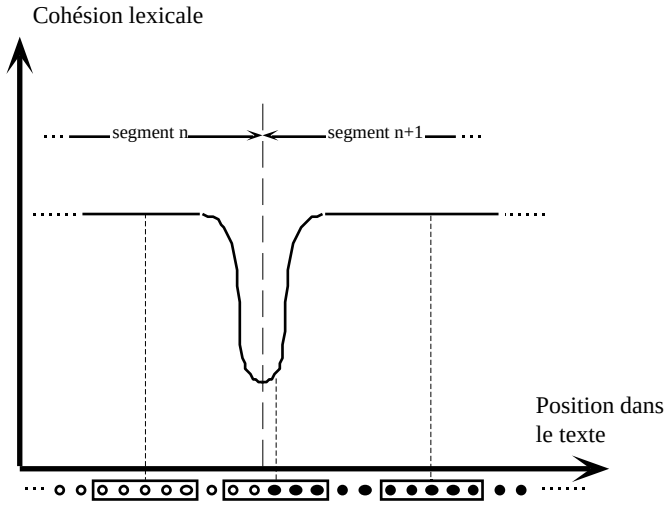


Fig. 9.2 - Corrélation entre la cohésion lexicale d'un texte et son découpage en segments thématiquement homogènes (adaptée de [Kozima 1993])

Si l'on postule d'autre part qu'un texte est formé d'une succession de segments thématiquement homogènes, il est raisonnable de penser qu'une partie de texte prise au sein d'un segment présentera une forte cohésion lexicale tandis qu'une partie de texte prise à cheval entre deux segments sera caractérisée par une cohésion plus faible (cf. figure 9.2). C'est en s'appuyant sur cette idée que Kozima a proposé une méthode de repérage des bornes des segments dans les textes. Plus précisément, son objet était de

découper des textes narratifs[1] en scènes, lesquelles sont équivalentes aux situations auxquelles nous nous intéressons ici.

La méthode proposée repose donc principalement sur la détermination de la cohésion lexicale en tout point d'un texte. Kozima utilise pour ce faire une fenêtre d'une taille fixe Tf qu'il déplace sur tout le texte selon un pas de 1 mot. Le déplacement suit le sens de lecture du texte. À chaque station de la fenêtre, on évalue la cohésion des mots se trouvant à l'intérieur de celle-ci en utilisant le réseau *Paradigme*. Si W représente le vecteur des mots présents au sein de la fenêtre pour une position donnée de celle-ci alors la valeur de cohésion pour cette position est donnée par $\sigma(W,W)$. Autrement dit, on évalue la cohésion des mots de la fenêtre par rapport à eux-mêmes. En pratique, on injecte donc de l'activation dans *Paradigme* à partir des mots constituant W et après propagation, on mesure l'activation obtenue pour ces mêmes mots. La courbe donnant la valeur de cette cohésion pour chaque position de la fenêtre est appelée *Lexical Cohesion Profile* (LCP).

La valeur du LCP pour une position du texte est obtenue lorsque la fenêtre est centrée sur cette position (la fenêtre recouvre donc toujours un nombre impair de mots). Pour les positions allant de 1 à $(Tf - 1) / 2$, on complète la partie gauche de la fenêtre se trouvant avant le texte par le premier mot du texte. Symétriquement, pour les positions allant de $N - ((Tf - 1) / 2) + 1$ à N, N étant la position du dernier mot du texte, on complète la partie droite de la fenêtre se trouvant après le texte par le dernier mot du texte. Kozima a par ailleurs testé différentes tailles de fenêtre (entre 11 et 121 mots) en prenant comme référence une segmentation réalisée par des sujets humains. La valeur de 51[2] mots a été retenue comme celle conduisant aux résultats les plus proches du jugement humain. La taille de la fenêtre est un facteur important puisqu'elle détermine le degré de granularité des segments trouvés. Si la fenêtre est grande, il ne sera ainsi pas possible de détecter des segments de petite taille. À l'inverse si elle est trop petite, le LCP risque d'être noyé dans le bruit par la faute du trop peu de mots significatifs sur le plan thématique à l'intérieur de la fenêtre à chaque station de celle-ci.

Même avec une taille de fenêtre suffisamment grande, le LCP est caractérisé par la présence d'un bruit important qu'il est souhaitable de filtrer afin de faire apparaître de manière plus nette les grandes évolutions. Pour ce faire, Kozima a utilisé une technique

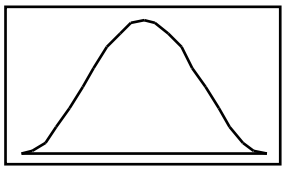
---

[1] Le type de texte retenu justifie l'emploi d'une méthode s'appuyant sur la cohésion lexicale étant donné que les textes narratifs sont généralement caractérisés par une variabilité assez grande de l'expression d'une même notion ainsi que par une faible présence de marques linguistiques telles que les connecteurs.

[2] Kozima ne précise pas les pré-traitements éventuels qu'il réalise sur les textes. Compte tenu de l'utilisation d'un dictionnaire, il est pratiquement certain qu'il procède à une lemmatisation. En revanche, il semble que tous les mots soient conservés.

classique de filtre passe-bas[1] consistant à diminuer l'influence des mots les plus extérieurs de la fenêtre pour le calcul de la cohésion. Le changement du contenu de la fenêtre intervenant lors de son déplacement se traduit ainsi par une variation moins brutale des valeurs de cohésion. Différentes fonctions sont applicables pour réaliser cette atténuation. Une atténuation nulle donne une fenêtre dite rectangulaire. Une atténuation linéaire conduit pour sa part à une fenêtre triangulaire. À la suite de tests réalisés en prenant toujours comme référence une segmentation faite par des humains, Kozima a retenu la fenêtre de Hanning, produisant une atténuation de type cosinus :

$$p(i,j) = \frac{1}{2}(1 + \cos(\frac{|i-j|}{\Delta}\pi))$$



où i est la position de la fenêtre dans le texte, j est la position dans le texte d'un mot de la fenêtre, p(i,j) est la valeur de l'atténuation du mot j lorsque la fenêtre se trouve à la position i et Δ = (Tf - 1) / 2.

En utilisant une telle fonction d'atténuation, l'activation initialement injectée dans le réseau n'est plus calculée à partir de la significativité des mots concernés mais sur la base de leur significativité modulée par p(i,j).

L'observation du LCP montre que comme attendu, ses principaux minima sont corrélés à des changements de segment thématique. Kozima ne propose toutefois pas véritablement de méthode pour réaliser le découpage du texte en segments à la suite du calcul du LCP. Pour comparer un peu formellement les jugements humains aux résultats de l'algorithme, il se contente de découper la courbe en espaces réguliers de taille δ et de retenir la rupture de phrase la plus proche du minimum du LCP dans chacun de ces espaces comme borne de segment. La précision et le rappel (au sens où nous les avons définies en présentant les travaux de Hearst) ont été calculés pour différentes valeurs de δ. Le rappel chute de 93% à 25% à mesure que δ croit de 6 à 100 mots alors que la précision passe dans le même temps de 33% à 66%[2]. Ces résultats ne sont néanmoins qu'indicatifs dans la mesure où ils n'ont été obtenus que sur deux textes. Pour le premier, le jugement humain de référence était le résultat du recoupement de l'avis de 16 juges mais dans le cas du second texte, il n'était constitué que de l'avis de Kozima lui-même.

[1] Le bruit prend la forme de micro-variations du LCP et se situe donc dans les hautes fréquences (cf. §1.4 de ce chapitre pour avoir un exemple du type de courbe que l'on obtient)

[2] Les chiffres donnés ne sont qu'approximatifs, ayant été mesurés sur une courbe et non fournis explicitement.

2. La méthode de segmentation thématique de SEGCOHLEX

2.1. Principes

Les contraintes définies en préambule de ce chapitre imposent à l'analyse thématique de SEGCOHLEX de segmenter des textes caractérisés par une faible réitération du vocabulaire en adoptant le meilleur compromis possible entre la taille des segments formés et leur cohérence thématique, sachant qu'en tout état de cause, la cohérence thématique de ces segments doit être évaluée. Pour ce faire, on reprend l'hypothèse stipulant que la cohésion lexicale est un reflet de la cohérence thématique des textes et on cherche à évaluer la cohésion des textes par rapport à des connaissances de référence sur la cohésion entre mots.

L'existence de telles connaissances, en particulier pour la langue française, est néanmoins un problème. Il n'existe pas en effet de thesaurus librement accessible sous forme électronique qui pourrait s'apparenter au *Roget's Thesaurus* pas plus qu'il n'existe sous la même forme de dictionnaire comparable au LDOCE et à *Glossème*. Les méthodes utilisant des connaissances sur la cohésion lexicale comme celle de Morris et Hirst (cf. annexe I) ou celle de Kozima ne sont donc pas directement applicables dans notre contexte de travail.

En revanche, à condition de disposer d'un corpus suffisamment important (représentant plusieurs millions de mots), il est possible d'enregistrer les cooccurrences existant entre les mots de cet ensemble de textes et d'obtenir, en appliquant une mesure adéquate, des informations significatives et d'une certaine généralité sur la force de cohésion entre deux mots. Si l'espace séparant deux mots qui cooccurrent est suffisamment large, cette cohésion peut refléter la présence de relations sémantiques et pragmatiques entre ces mots, donc l'existence d'une cohérence thématique sous-jacente.

En combinant les valeurs de cohésion issues des relations de cooccurrence présentes entre les mots d'un même ensemble, on peut évaluer la cohésion de cet ensemble de mots et donc celle d'un bloc de texte. Il devient alors possible de reprendre les principes développés par Kozima et que nous avons exposés au §1.4.2.

C'est le parti pris que nous avons retenu pour la méthode de segmentation de SEGCOHLEX [Ferret 1998, Ferret 1998]. On calcule la valeur de cohésion des mots se situant à l'intérieur d'une fenêtre passant par toutes les positions du texte. On obtient ainsi une courbe rendant compte de la cohésion de l'ensemble du texte. Les zones de faible cohésion représentent des groupes de mots peu homogènes, donc n'appartenant a priori pas au même domaine. Ce sont donc les zones présumées de changement de thème. Cette

propriété est exploitée afin de segmenter la courbe de cohésion de façon automatique en utilisant une succession de traitements très simples. On forme de cette manière un ensemble de segments présentant en outre la propriété, imposée par les contraintes pesant sur SEGCOHLEX, de se voir dotés chacun d'un niveau de cohésion caractérisant en quelque sorte sa "qualité thématique".

En adoptant un réseau de cooccurrences lexicales comme source de connaissances sur la cohésion lexicale, notre objectif était d'adapter la méthode proposée Kozima afin de produire une méthode de segmentation à la fois plus simple, plus efficace et imposant des pré-requis moins importants. Les deux premiers points font référence à l'utilisation d'un mécanisme de propagation d'activation au sein d'un réseau lexical pour mesurer la cohésion entre les mots. Même si le nombre de cycles est volontairement limité, la propagation d'activation est en effet un processus complexe à mettre au point et d'un fonctionnement coûteux. Le dernier point renvoie quant à lui à l'utilisation par Kozima d'un réseau lexical construit à partir d'un dictionnaire afin de capturer la notion de cohésion lexicale. Il nous a semblé en effet à la fois matériellement plus faisable mais également plus général de concevoir une méthode de segmentation capable d'utiliser une source de connaissances construite de manière automatique qu'une méthode fondée sur une source de connaissances qui est le produit d'un très lourd investissement humain et qui est donc difficilement adaptable à de nouveaux domaines.

2.2. Construction du réseau de cooccurrences lexicales

2.2.1. Pré-traitement des textes

Compte tenu de la tâche finale considérée, il est important de caractériser les textes par leurs mots significatifs sur le plan de la différenciation thématique. Nous n'avons ainsi retenu que la forme canonique des mots dits pleins de chaque texte, c'est-à-dire les noms, les verbes et les adjectifs. Nous avons laissé de côté les adverbes et tous les mots grammaticaux. Bien que certains adverbes puissent être significatifs sur le plan thématique (en particulier ceux formés à partir d'un adjectif et du suffixe -ment), nous avons estimé que cette catégorie est globalement trop hétérogène pour ne pas engendrer plus de bruit qu'elle n'apporte d'information.

Nous avons également laissé de côté les noms propres et les sigles, à la fois pour des raisons pratiques et des raisons de fond. Sur le plan pratique, on constate que le nombre de noms propres et de sigles présents dans les textes journalistiques – notre réseau de cooccurrences a été construit à partir du journal *Le Monde* – est très important (de l'ordre

de plusieurs dizaines de milliers au moins sur 2 ans du journal *Le Monde*) mais qu'ils n'apparaissent le plus souvent qu'une seule fois ou bien dans un seul texte. La grande majorité d'entre eux ne peuvent donc pas donner de cooccurrences significatives et se verraient de ce fait éliminés lors de la construction finale du réseau de cooccurrences. Il est donc préférable de les éliminer avant le calcul des cooccurrences afin de limiter le plus possible le nombre intrinsèquement très important de celles-ci.

Sur le fond, nous considérons que les noms propres et les sigles sont importants lorsque l'on cherche à repérer des événements précis mais qu'il est possible d'en faire abstraction pour la mise en évidence de thèmes plus larges. Bien entendu, un nom propre ou un sigle peut être très informatif sur le plan thématique : des sigles comme "SNCF", "IRA", "OTAN" ou des noms de personnes ou de lieux comme "Einstein", "Blériot" ou "Auschwitz" renvoient à des contextes spécifiques. Mais dans le même temps, les mots relevant de ces catégories ont souvent une pérennité d'usage assez faible. Une entreprise comme la SNCF peut très bien changer de nom tout en restant attachée au domaine ferroviaire et une organisation terroriste comme l'IRA peut disparaître si le conflit de l'Irlande du Nord se dénoue sans pour autant que le terrorisme de façon plus générale disparaisse. Des textes écrits à l'issue de ces changements ne feront ainsi plus référence aux noms "SNCF" et "IRA" tout en continuant à aborder les domaines pré-cités.

L'opération de sélection des mots est réalisée par une chaîne de traitement ayant pour point de départ des textes sous forme ASCII, accompagnés d'un balisage SGML. Ce balisage, résultat du processus décrit dans [Adda et alii 1997], permet de séparer les textes, de repérer leurs constituants particuliers, tels que les titres par exemple, et de mettre en évidence leurs paragraphes[1]. Les textes sont dans un premier temps segmentés à l'aide du segmenteur *Mtseg*, développé dans le cadre du projet MULTEXT [Véronis & Khouri 1995]. Celui-ci permet en particulier d'identifier des éléments spécifiques tels que les dates et les nombres qui seront filtrés par la suite. Il permet également de marquer les expressions composées de plusieurs mots. Il est ainsi possible d'éliminer plus facilement les locutions adverbiales et d'identifier les noms composés, lesquels sont particulièrement significatifs puisqu'a priori peu polysémiques. Une liste des 2300 noms composés les plus fréquents, en l'occurrence ceux possédant au moins 350 occurrences sur un corpus rassemblant un peu plus de 10 ans du journal *Le Monde* (de janvier 1987 à mars 1997) et 5 ans du journal *Le Monde Diplomatique* (de 1991 à 1995), a été constituée à l'aide de l'outil INTEX [Silberztein 1993] et a été incorporée au segmenteur. La limite fixée s'explique par la volonté de ne pas engorger le processus de construction du réseau de cooccurrences lexicales par un vocabulaire rare dans le corpus traité qui ne serait la source que de cooccurrences de fréquence trop faible pour être retenues par la suite.

[1] Nous remercions plus particulièrement Gilles Adda pour avoir mis à notre disposition 11 mois du journal *Le Monde* sous cette forme.

Après segmentation, les textes sont soumis à l'étiqueteur *TreeTagger* [Schmid 1994]. Nous avons utilisé les paramètres fournis pour le Français sans procéder à un réapprentissage spécifique pour notre corpus. Cet étiqueteur possède en effet un dictionnaire de large couverture complété par un mécanisme de traitement des mots inconnus sur la base des infixes. Par ailleurs, bien qu'une évaluation du *TreeTagger* sur notre corpus n'ait pas été réalisée, on peut penser que ses performances pour la tâche considérée, c'est-à-dire le repérage des noms, des adjectifs et des verbes, sont globalement satisfaisantes dans la mesure où ses erreurs concernent essentiellement des points qui ne sont pas en liaison avec notre tâche : distinction entre "tout" adverbe et "tout" pronom, entre "que" pronom relatif et "que" conjonction de subordination ou encore entre les occurrences de "des" correspondant à l'article indéfini et celles relevant de la contraction de la préposition "de" avec l'article défini "les" [Stein & Schmid 1995].

Seule une certaine faiblesse dans la reconnaissance des noms propres et des sigles, assimilés parfois à des noms, est à noter mais elle se trouve partiellement compensée par le filtrage final des cooccurrences de faible fréquence. Le traitement des noms composés est quant à lui un peu particulier dans la mesure où les 2300 qui sont reconnus par le segmenteur ne figurent pas dans le vocabulaire du *TreeTagger*, vocabulaire que nous ne pouvions pas augmenter sans procéder à un nouvel entraînement. Les constituants de chaque occurrence de nom composé ont donc été séparés avant de passer par l'étiqueteur, étiquetés et lemmatisés individuellement puis rassemblés pour reconstituer le nom composé originel. Bien entendu, il en résulte quelques erreurs qui n'interviendraient sans doute pas si les noms composés étaient intégrés au vocabulaire de l'étiqueteur.

Le pré-traitement des textes se termine par la sélection des mots jugés thématiquement représentatifs et la transformation des textes au format requis pour le calcul des cooccurrences ou la segmentation thématique. On pourra se reporter à l'annexe G pour avoir un descriptif précis de ces formats. La sélection des mots est réalisée directement sur la base de l'étiquetage réalisé par le *TreeTagger*. Outre la possibilité de repérer les noms, les verbes et les adjectifs, celui-ci nous permet également de ne pas sélectionner les auxiliaires *être* et *avoir* ainsi que plus généralement un certain nombre d'auxiliaires de modalité tels que *devoir, pouvoir* ou *falloir* (cf. contenu exact de cette stop-list à l'annexe G). Ceux-ci apparaissent en effet avec une fréquence élevée mais n'apportent de fait aucune information sur le plan thématique.

2.2.2. Le réseau de cooccurrences lexicales

Pour capturer la notion de cohésion, nous avons choisi de construire un réseau de cooccurrences lexicales calculées à partir d'un ensemble très important de textes. Le

corpus utilisé se compose de 24 mois du journal *Le Monde* répartis entre les années 1990 et 1994. Il représente un ensemble originel d'un peu plus de 39 millions de mots. Dans la mesure de nos possibilités, nous avons sélectionné ces mois sur un intervalle de temps suffisamment large pour minimiser la dépendance vis-à-vis d'événements spécifiques.

Le calcul des cooccurrences a été réalisé à partir de la méthode décrite dans [Church & Hanks 1990]. Le corpus a été pré-traité au préalable selon la procédure présentée au paragraphe précédent, ce qui a conduit à ne retenir qu'environ 37% des mots. L'évaluation des cooccurrences s'effectue en faisant glisser selon un incrément de 1 mot une fenêtre d'une taille de 20 mots sur les textes du corpus. À chaque position de la fenêtre, on enregistre les cooccurrences entre le mot de tête et les autres mots la fenêtre. Ce processus est illustré pour deux positions de la fenêtre par la figure 9.3. La fenêtre respecte la délimitation des textes dans la mesure où deux textes adjacents n'abordent généralement pas les mêmes thèmes. Après que le 20ème mot précédant la fin d'un texte a occupé la première position de la fenêtre, la taille de celle-ci diminue de ce fait d'un mot à chacune de ses progressions. Contrairement à Church et Hanks, nous ne sommes pas intéressé par la conservation de l'ordre au sein des cooccurrences. La cooccurrence mot1-mot2 est donc strictement équivalente à la cooccurrence mot2-mot1 et les deux ne sont pas différenciées lors de l'enregistrement.

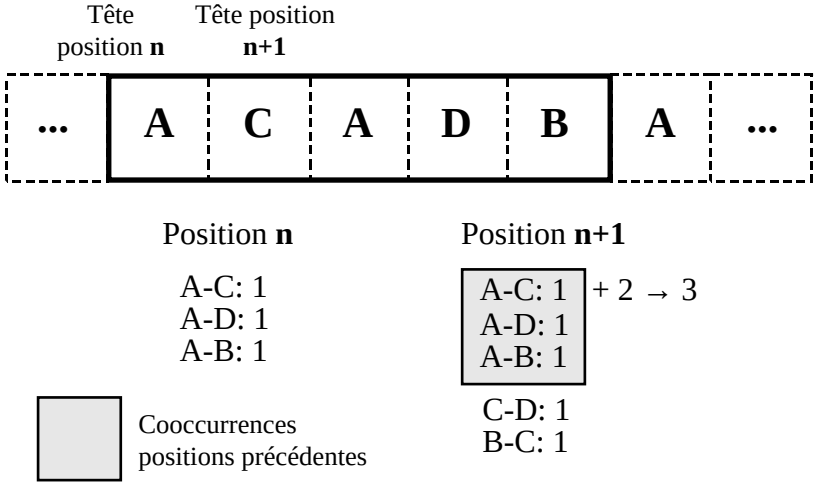


Fig. 9.3 - Calcul des cooccurrences au sein d'une fenêtre

Ces deux derniers points ont été dictés par notre tâche finale. Celle-ci nous a également guidé dans le choix de la taille de la fenêtre. Pour segmenter des textes sur un critère thématique, il est nécessaire de disposer de connaissances à la fois sémantiques et pragmatiques. Ces connaissances rendent compte en particulier de l'agencement des propositions au sein d'un texte et des relations existant entre ces propositions. Pour capturer ces connaissances au travers d'un réseau de cooccurrences lexicales (on utilisera également le terme de "collocation" dans ce qui suit), il est donc nécessaire que la fenêtre

au sein de laquelle ces cooccurrences sont comptabilisées couvre au moins l'équivalent de deux propositions. La limite supérieure est quant à elle imposée à la fois par des limites techniques, le nombre de collocations croissant très fortement avec la taille de la fenêtre, et par la granularité des unités que l'on souhaite distinguer. Une taille de 20 mots, après pré-traitement, s'est avérée le meilleur compromis dans le contexte présent[1].

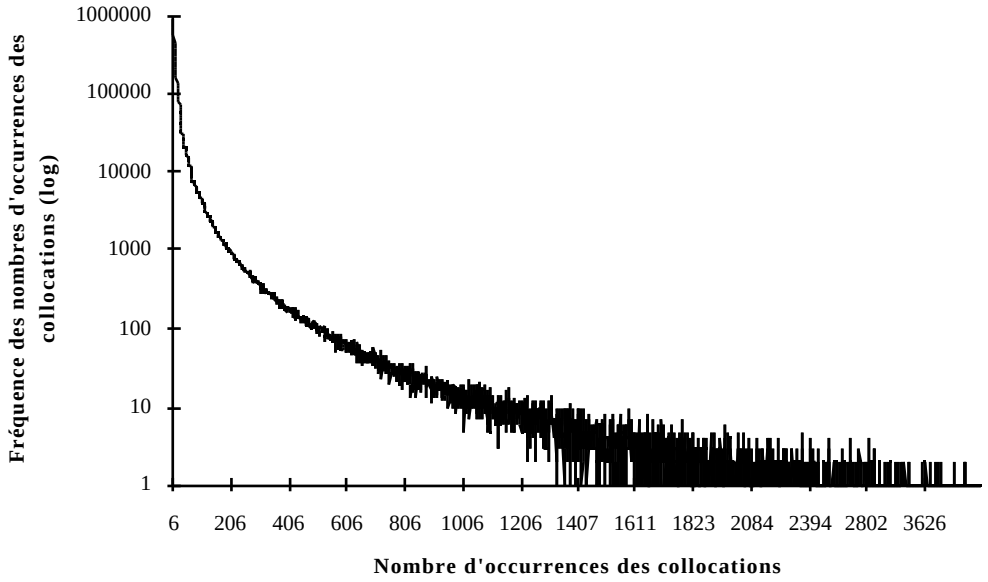


Fig. 9.4 - Distribution des collocations en fonction de leur nombre d'occurrences

À la suite du calcul des cooccurrences, nous opérons une sélection afin de ne retenir que les collocations les plus stables, i.e. les plus récurrentes, parmi les textes, donc les plus significatives, et éviter ainsi un bruit potentiellement gênant pour les processus exploitant le réseau. Seules les cooccurrences de fréquence supérieure à 5 sont ainsi conservées, ce qui représente à peu près 1/3 de celles initialement comptabilisées. On aboutit à un réseau formé de 31000 mots[2] liés par quelques 7 millions de relations[3].

---

[1] Le choix de la taille de la fenêtre s'accompagne d'une part d'arbitraire dans la mesure où la lourdeur du processus de construction d'un tel réseau permet difficilement de réaliser plusieurs expérimentations. À supposer que la chose soit réalisée, il resterait encore à trouver des critères permettant de juger qu'un réseau est plus intéressant qu'un autre. Une façon de faire serait d'observer l'impact d'un changement de réseau au niveau de la tâche finale, en l'occurrence la segmentation thématique. Outre que cela impose l'existence d'un cadre d'évaluation assez large de la tâche en question (il faut disposer d'un ensemble important et suffisamment diversifié de résultats de référence), une telle évaluation indirecte se heurte à la nécessité d'adapter les paramètres de la tâche aux caractéristiques de chaque réseau, ce qui constitue une difficulté supplémentaire dans la comparaison des résultats.

[2] Le filtrage des cooccurrences de faible fréquence entraîne également une diminution de 2/3 environ de la taille du vocabulaire puisque celui-ci comprenait initialement de l'ordre de 100000 mots. La plupart des mots ainsi supprimés correspondent à des noms propres ou des sigles ayant échappé au TreeTagger.

[3] Le chiffre de 7 millions correspond plus précisément au nombre de collocations. Chaque collocation doit en théorie donner lieu à deux relations, une relation de *mot1* vers *mot2* et une autre dans le sens opposé, ce qui conduit à un total de 14 millions de relations. Cependant, l'ordre des mots n'étant pas conservé dans les collocations considérées, on peut se contenter de faire référence aux mêmes données lorsqu'on

Chaque mot est donc impliqué en moyenne dans 439 collocations, ce chiffre recouvrant un spectre de valeurs assez important puisque certains mots “possèdent” moins d’une dizaine de collocations tandis que d’autres sont présents dans plus de 10000. La figure 9.4 montre la distribution des collocations obtenues en fonction de leur nombre d’occurrences. On observe sans surprise que les collocations peu fréquentes rassemblent la plus grande part des occurrences et que la fréquence des collocations chute très brutalement lorsque le nombre d’occurrences qu’elles regroupent s’élève. La stabilisation de cette tendance n’intervient qu’au delà d’un nombre d’occurrences d’environ 1000. Les fréquences sont alors d’une dizaine de collocations seulement et en dépit du spectre de valeurs couvert, le nombre de collocations se trouvant dans cette plage de nombre d’occurrences est donc assez faible.

Comme dans [Church & Hanks 1990], nous avons adopté une estimation de l’*information mutuelle* [Fano 1961] comme mesure de la cohésion entre deux mots. L’information mutuelle $I(x,y)$ de deux mots x et y ayant les probabilités $P(x)$ et $P(y)$ se définit par :

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)\ P(y)}$$

où $P(x,y)$ est la probabilité d’observer la cooccurrence entre les mots x et y [1].

La mesure de cohésion n’est qu’une estimation de l’information mutuelle étant donné que les probabilités $P(x)$, $P(y)$ et $P(x,y)$ ne sont elles-mêmes que des estimations réalisées en divisant les fréquences $f(x)$, $f(y)$ et $f(x,y)$[2] par N, la taille du corpus de référence. On obtient ainsi la mesure de cohésion $coh(x,y)$:

$$coh(x,y) = \log_2 (N\ \frac{f(x,y)}{f(x)\ \ f(y)})$$

Cette mesure est bornée inférieurement par $-log_2 N$ puisque $f(x)$ et $f(y) \leq N$ et $f(x,y) \geq 1$. Elle est également bornée supérieurement. On a en effet $f(x)$ et $f(y) \geq 1$ et compte tenu de la taille nécessairement limitée du corpus, il existe un nombre théorique maximal de cooccurrences entre deux mots. Cette valeur est obtenue dans le cas d’un texte composé de la répétition du même mot. $f(x,y)$ est alors égal $N\ \ (T_f - 1)$ et la borne supérieure de la mesure cohésion, appelée information mutuelle maximale relative au corpus, est donc donnée par la formule suivante :

---

présente la collocation *mot1-mot2* comme une collocation de *mot1* et la collocation *mot2-mot1* comme une collocation de *mot2*. Le nombre de relations s’identifie donc au nombre de collocations.

[1] Le fait de ne pas tenir compte de l’ordre au niveau des collocations permet de conserver la symétrie initiale de l’information mutuelle.

[2] Ces fréquences correspondent respectivement au nombre d’occurrences du mot x, du mot y et au nombre de fois où les mots x et y cooccurrent.

$$I_{max} = \log_2 N^2 (T_f - 1)$$

où T_f représente la taille de la fenêtre de calcul des cooccurrences.

Comme le font remarquer Church et Hanks dans [Church & Hanks 1990], il est très rare dans les faits d'avoir des valeurs négatives de l'information mutuelle dans le cas du calcul de cooccurrences car, pour obtenir des collocations significatives présentant cette caractéristique, il serait nécessaire de disposer de corpus beaucoup plus gros que ceux que nous manipulons. On considérera donc en pratique que *coh(x,y)* est bornée inférieurement par 0. En outre, on normalise les valeurs de cette mesure en les divisant par l'information mutuelle maximale relative au corpus. On obtient donc une mesure comprise dans l'intervalle [0,1]. Le seul inconvénient de cette normalisation vient de ce que l'information mutuelle maximale étant assez supérieure aux valeurs obtenues couramment, les valeurs normalisées de la mesure de cohésion n'exploitent qu'une partie du spectre des valeurs permises. Ces valeurs dépassent ainsi rarement 0,25 et ne vont presque jamais au delà de 0,3.

abbé	occ	coh	frégate	occ	coh	prendre	occ	coh
réifier	6	0,286	destroyer	24	0,295	(...) contre-pied	129	0,163
breuil	8	0,253	aviso	17	0,293	décision	6	0,160
sans-logis	19	0,240	ravitailleur	29	0,288	relai	9	0,159
aumônier	21	0,238	corvette	20	0,284	tenaille	58	0,156
mal-logé	13	0,238	(...) artimon	7	0,270	pouls	45	0,156
libertinage	7	0,233	lynx	9	0,270	(...) relève	189	0,146
sevrer	7	0,230	exocet	14	0,268	escampette	11	0,146
chanoine	6	0,229	crotale	9	0,264	géothermie	6	0,145
soutane	7	0,229	dragueur	11	0,263	subreptice	8	0,145
bénédictin	16	0,225	pièce_d'artillerie	16	0,253	mutandis	8	0,145
monastique	10	0,225	navire_de_guerre	11	0,251	bielle	6	0,145
candeur	9	0,221	croiseur	6	0,249	(...) précaution	379	0,141
bonté	10	0,221	(...) marine	96	0,227	inexploitable	6	0,141
évêché	6	0,220	porte-avions	20	0,225	ombrage	30	0,141
prieur	9	0,219	semonce	6	0,224	(...) relais	474	0,140
sans-abri	22	0,214	ravitaillement	13	0,223	recracher	6	0,139
abbaye	27	0,212	coque	12	0,222	coccinelle	10	0,139
buisson	12	0,209	(...) bateau	34	0,188	relayeur	10	0,139
monastère	28	0,208	(...) mine	23	0,182	(...) donner	3201	0,102
curé	20	0,206	maritime	15	0,181	saisir	383	0,099
ecclésiastique	11	0,205	(...) envisager	7	0,119	offrir	682	0,096
(...) musulman	7	0,09	franc	50	0,119	offre	204	0,093

Fig. 9.5 - Extraits du réseau de cooccurrences lexicales
occ : nombre d'occurrences; coh : valeur de la mesure de cohésion normalisée

La figure 9.5 ci-dessous donne un aperçu du contenu du réseau de cooccurrences lexicales en montrant un certain nombre des collocations, classées par ordre décroissant de leur valeur de cohésion, dans lesquelles se trouvent impliqués trois mots de ce réseau[1].

[1] Du fait du nombre élevé de collocations associées à un mot, la liste donnée pour chacun des trois mots par la figure 9.4 est en fait formée d'un ensemble de fragments, les interruptions étant signalées par le signe "(...)".

Ces exemples apportent en particulier un éclairage sur les principaux types de relations sous-tendant les collocations présentes dans un tel réseau. Ces collocations se répartissent de ce point de vue entre les quatre principaux cas de figure suivants :

- aucune relation systématique ne réunit les deux mots de la collocation. Leur présence simultanée au sein de la fenêtre de comptage des collocations n'est que l'effet du hasard ou bien est liée à un contexte extrêmement restreint que l'on ne retrouve que dans un ou deux textes. Cette situation s'applique malheureusement à une grande partie des collocations, ce qui revient à dire que la plupart d'entre elles ne représentent en fait que du bruit. Celui-ci peut être filtré pour partie en supprimant toutes les collocations ne possédant qu'un faible nombre d'occurrences. Le niveau de ce filtre est néanmoins délicat à ajuster : il faut trouver le point d'équilibre permettant de supprimer le plus de bruit possible tout en limitant au maximum la perte de collocations intéressantes pour notre application. Or, ces dernières n'ont pas toutes nécessairement une fréquence très élevée. C'est le cas par exemple des collocations *abbé–mulsuman* ou *prendre–saisir*. En choisissant un niveau de filtrage initial assez faible, comme dans le cas présent (seuil égal à 5), on conserve une part de bruit assez importante en faisant la supposition que l'on dispose, au sein de l'application spécifique dans laquelle ces collocations ont utilisées, de moyens plus fins pour éliminer ce bruit. Les collocations *abbé–sevrer, frégate–franc* ou encore *prendre–inexploitable* sont des exemples de ce bruit résiduel;

- les deux mots de la collocation sont réunis par une relation que l'on peut qualifier de lexico-syntaxique. Ces collocations illustrent la présence d'unités lexicales composites, que l'on désigne fréquemment sous le vocable de "mots composés". Un mot composé est un regroupement lexical plus ou moins figé présentant une certaine unité de sens. Un tel regroupement est plus ou moins stable, i.e. récurrent, et plus ou moins flexible, c'est-à-dire caractérisé par une plus ou moins grande variété de formes. La notion de mot composé regroupe donc aussi bien des formes assez figées comme "chemin de fer" que des formes plus flexibles comme "donner un aperçu". Compte tenu de la taille de fenêtre adoptée et des pré-traitements réalisés, on peut raisonnablement penser que tous ces types de mots composés donnent lieu ici à des collocations significatives. Les collocations *prendre–contre-pied* et *prendre–relais* témoignent d'ailleurs de leur présence au sein du réseau de cooccurrences.

  L'intérêt de ce type de collocations du point de vue de la segmentation thématique n'est cependant pas évident. Le fait de trouver par exemple que "chemin" est lié à "fer" du fait de leur présence simultanée dans "chemin de fer" n'est en effet aucunement significatif sur le plan thématique et ne permet pas en particulier de faire le lien avec le domaine ferroviaire. Seul le cas où l'un des termes assez général d'un mot composé permet d'accéder à un autre de ses termes, plus spécifique et relié au thème courant,

serait potentiellement intéressant. Cette possibilité n'apparaît cependant guère exploitable. Un terme général est par nature lié à beaucoup de mots, mais faiblement et uniformément. Il ne constitue donc pas un bon indice de rappel pour des mots plus spécifiques. Par ailleurs, dans un domaine donné, il est plus probable de rencontrer le terme d'un mot composé le plus lié à ce domaine que les autres de ses termes. Dès lors, la fonction d'évocation mentionnée précédemment risque simplement de ne pas servir.

- il existe une relation de nature sémantique entre les deux mots de la collocation. Cette relation peut aussi bien faire partie des relations sémantiques intervenant au niveau lexical, comme les relations de synonymie ou d'antonymie, que des relations sémantiques prenant place davantage au niveau des concepts évoqués par les mots, comme les relations d'hyponymie, d'hyperonymie ou de méronymie. Ces deux catégories sont en effet représentées dans le réseau de cooccurrences. Les collocations *abbé–curé* (synonymie) et *prendre–offrir* (antonymie) par exemple sont représentatives du premier type de relations tandis que les collocations *abbé–écclésiastique, frégate–bateau* (hyperonymie ou hyponymie suivant le sens considéré) et *frégate–coque* (méronymie) illustrent la présence de relations du second type.
  Du point de vue de la segmentation thématique, ces collocations présente l'intérêt d'élargir le nombre de termes utilisés pour parler d'une même chose, ce qui accroît les chances de mettre en évidence les relations pragmatiques existant entre ces mots et d'autres mots. Supposons par exemple que les mots *m1* et *m2* n'interviennent pas dans une collocation significative mais fassent référence à des entités impliquées dans une même situation. Si dans le réseau de cooccurrences, *m1* est liée assez fortement à *m3* du fait d'une relation sémantique et que *m3* est lui-même lié à *m2* sur un critère pragmatique, on pourra tout de même faire apparaître la relation pragmatique existant entre *m1* et *m2* en utilisant la médiation de la relation entre *m1* et *m3*.

- une relation de nature pragmatique unit les deux mots de la collocation. Cette relation traduit l'appartenance de ces deux mots, ou plutôt de deux concepts auxquels ceux-ci font référence, à une même situation ou plus généralement à un même thème. La collocation *frégate–artimon* provient ainsi du fait que les termes "frégate" et "artimon" appartiennent tout deux au domaine de la marine tandis que c'est plus particulièrement la présence de mines dans les situations de guerre sur mer qui sous-tend la collocation *frégate–mine*.

Il est admis que les ressources constituées à partir d'un corpus entretiennent une dépendance vis-à-vis de ce dernier et héritent de ses spécificités. Le réseau de cooccurrences lexicales n'échappe pas à cette règle. Bien qu'il ait été construit à partir du journal *Le Monde*, qui peut être considéré comme une source textuelle assez généraliste, on peut remarquer que certaines des collocations qu'il abrite, même parmi

les plus significatives, sont directement le fruit d'une dépendance vis-à-vis de son corpus d'origine. Ce phénomène s'observe en particulier au niveau des collocations sous-tendues par une relation pragmatique.

La collocation *abbé–sans-logis* en est une illustration. Sur un plan général, il n'y a aucune raison de lier les mots "abbé" et "sans-logis". Le premier appartient au domaine de la religion tandis que le second fait référence à la situation sociale des individus. Néanmoins, compte tenu de l'existence de l'abbé Pierre et de son action pour les sans-abris, on est confronté à la présence rapprochée des mots "abbé" et "sans-logis" dans beaucoup d'articles du *Monde.*, ce qui explique la présence et la force de cette collocation. On trouve donc au sein du réseau de cooccurrences à la fois des collocations étayées par des relations pragmatiques très générales – le mot "abbé" est également fortement lié aux mots "abbaye" et "soutane" par exemple – et des collocations au contraire liées à un contexte plus étroit mais fortement représenté dans le corpus utilisé.

Or, le réseau de cooccurrences ne fournit guère de moyen, en propre, de faire la distinction entre les deux types de collocations. Autant que permet de le constater une rapide analyse manuelle, aucun indice touchant au nombre d'occurrences ou à la valeur de cohésion n'offre en effet le moyen d'opérer cette discrimination. La présence de ces différents types de collocations conduit à relativiser la notion de bruit. Si l'on traite un texte portant sur une question religieuse, la collocation abbé–sans-abris représentera une forme de bruit, en dépit du fait qu'elle possède une valeur de cohésion élevée. En revanche, elle pourra être utile si l'on aborde un texte plus directement lié au corpus de référence et traitant de l'action de l'abbé Pierre. C'est en cela que la dépendance du réseau vis-à-vis du corpus influe sur la nature des textes qu'il permet de traiter.

Dans ce cadre, les collocations qualifiées de bruit rassemblent les collocations intervenant dans un contexte à la fois très spécifique et peu récurrent. Le degré de récurrence reste cependant un paramètre dont la valeur est fixée de façon arbitraire, ce qui rend en définitive assez ténue la frontière entre bruit et information[1]. Précisons que la détection des collocations relevant du bruit doit s'appuyer à la fois sur le nombre d'occurrences et la valeur de cohésion. Avec l'information mutuelle, une collocation peut en effet avoir une valeur de cohésion élevée avec un faible nombre d'occurrences si les mots qui la composent possèdent eux-même un faible nombre d'occurrences.

Compte tenu de la taille du réseau de cooccurrences lexicales et de l'absence de moyen automatique pour caractériser le type d'une collocation, il est bien entendu impossible de

[1] Dans [Church & Hanks 1990], Church stipule que les collocations dont le nombre d'occurrences est inférieur ou égal à 5 doivent être laissées de côté du fait de leur trop grande instabilité. Dans [Church & Mercer 1993], ce seuil passe à 10 mais dans un cas comme dans l'autre, aucune justification n'est véritablement apportée à l'appui de la valeur retenue.

donner une estimation de la proportion des collocations se répartissant entre ces quatre grandes catégories. Nous verrons au chapitre 10 un premier moyen de structurer un tel réseau sur le plan thématique, ce qui peut contribuer indirectement à réaliser une partie de cette estimation.

2.3. Méthode de segmentation

Rappelons que la méthode de segmentation que nous proposons comporte deux étapes. Tout d'abord, nous évaluons la cohésion des différentes parties du texte à segmenter. Nous exploitons ensuite les ruptures significatives de cette cohésion afin de détecter les changements thématiques.

2.3.1. Évaluation de la cohésion d'un texte

Principes généraux

L'évaluation de la cohésion d'un texte est le résultat de l'évaluation de la cohésion de ses différentes parties. Une partie de texte est définie ici de façon arbitraire par une fenêtre rassemblant un nombre donné de mots du texte, après que celui-ci a été pré-traité suivant le processus décrit au §2.3.1. Au sein de cette fenêtre, nous caractérisons la cohésion entre les différents mots qui la composent en utilisant comme référence le réseau de cooccurrences lexicales présenté ci-dessus. Nous faisons l'hypothèse que le nombre et la force des liens que ces mots entretiennent au niveau d'un tel réseau constituent un bon indicateur de leur cohésion mutuelle. Plus spécifiquement encore, nous estimons que cette cohésion, que l'on peut qualifier de globale compte tenu de la diversité des relations sous-tendant les collocations (cf. §2.3.2), est fortement corrélée avec la cohésion thématique de la portion de texte considérée.

Cette corrélation repose sur le raisonnement suivant. Lorsque la fenêtre d'évaluation de la cohésion se trouve à cheval entre deux segments faisant référence à des thèmes distincts, on trouve peu de liens entre les mots de la fenêtre : d'une part ces mots se répartissent implicitement en deux sous-ensembles qui n'ont que peu de liens l'un avec l'autre (un sous-ensemble par thème), et d'autre part chacun d'entre eux est trop petit pour contenir un nombre suffisamment important de mots spécifiques du même domaine, les seuls à être la source de relations significatives. Ce faible nombre de liens a pour conséquence directe une valeur faible de la cohésion. À l'inverse, lorsque la fenêtre d'évaluation de la cohésion se situe à l'intérieur d'un segment thématiquement homogène, ses mots ne forment qu'un seul ensemble, faisant référence à un sujet unique. Le nombre

potentiel de liens est de ce fait beaucoup plus important et par voie de conséquence, les valeurs de la cohésion sont plus élevées.

En renversant le sens d'exploitation des règles, on fait l'hypothèse que des valeurs de cohésion élevées sont le signe d'une continuité thématique alors que des valeurs de cohésion faibles marquent une zone de rupture thématique.

La fenêtre d'évaluation de la cohésion est déplacée sur tout le texte à segmenter selon un pas de 1 mot. Ce déplacement suit le sens de lecture habituel des textes. Une valeur de cohésion est calculée à chaque station de la fenêtre, c'est-à-dire pour chaque position du texte. Cette position est définie par le numéro du mot sur lequel la fenêtre est centrée lors du calcul de cohésion. En raison du pas adopté, les portions de texte correspondant à deux positions successives de la fenêtre se recouvrent fortement, ce qui permet de saisir les évolutions de la cohésion du texte de façon fine. Au début et à la fin du texte, un mécanisme de recopie du premier, respectivement du dernier, mot du texte assure que la fenêtre comporte toujours le même nombre de mots. Soit F_i, le vecteur représentant le contenu de la fenêtre lorsqu'elle est centrée sur le $i^{ème}$ mot du texte :

$$F_i = (m_g, m_{g+1}, \ldots, m_i, \ldots, m_{d-1}, m_d)$$

Pour F_1, on a donc $m_g = m_{g+1} = ... = m_1$ et pour F_N, N étant le nombre de mots du texte, on a $m_d = m_{d-1} = \ldots = m_N$. *Tf* correspond à la taille de la fenêtre, toujours égale à un nombre impair. Ce mécanisme de recopie intervient pour :

$\begin{cases} i \text{ de } 1 \text{ à } \dfrac{Tf-1}{2} & \text{au début du texte,} \\ i \text{ de } N - \dfrac{Tf-1}{2} + 1 & \text{à la fin du texte.} \end{cases}$

Calcul de la cohésion au sein de la fenêtre glissante

Le calcul proprement dit de la cohésion de la partie du texte délimitée par la fenêtre glissante se déroule en trois étapes. La première consiste à définir l'ensemble des mots présumés les plus en adéquation avec le thème actif, que ce soit des mots de la fenêtre ou des mots du réseau de collocations liés à ceux de la fenêtre. La deuxième étape a pour but, quant à elle, de pondérer l'ensemble des mots ainsi sélectionnés en fonction de leur importance supposée. La troisième et dernière étape, enfin, réalise le calcul de la valeur de cohésion associée à la position courante de la fenêtre à partir des poids établis lors de l'étape précédente.

Sélection des mots intervenant dans le calcul de la cohésion

Fonctionnellement, la première étape du calcul de la cohésion se divise en deux opérations successives. La première est chargée de définir l'ensemble des mots susceptibles d'intervenir dans le calcul de la cohésion du fait de leur proximité avec le contexte établi par le contenu de la fenêtre glissante. Cet ensemble se compose de l'intégralité des mots de cette fenêtre ainsi que de tous les mots du réseau de collocations directement liés aux mots de la fenêtre. Compte tenu de l'importance du nombre de collocations par mot et de la proportion de bruit parmi celles-ci, nous avons délibérément choisi de limiter à une profondeur d'un seul lien la recherche de mots du réseau de collocations liés à ceux de la fenêtre glissante. Une autre façon de procéder pourrait être de mener une recherche de profondeur variable en fonction d'une première évaluation de l'intérêt des mots trouvés. L'utilisation de la significativité des mots (cf. présentation des travaux de Kozima au §1.4.2) est un moyen possible de procéder à cette évaluation.

La seconde opération est la sélection dans ce premier ensemble des mots les plus liés au thème actif. Dans le contexte de SEGCOHLEX, aucune représentation explicite des thèmes n'existe. Il n'est donc pas possible de juger directement de l'adéquation d'un mot par rapport à un thème en général, et par rapport au thème actif en particulier. Le jugement conduisant à retenir ou au contraire à écarter un mot lors de cette première étape est rendu par un moyen plus indirect. À ce titre, il est entaché d'une certaine incertitude, partiellement compensée par le nombre total de mots impliqués. Ce moyen repose sur la seule information de nature thématique disponible à ce niveau, en l'occurrence celle contenue dans le réseau de collocations, et plus spécifiquement dans certains de ses liens.

N'étant pas capable de discerner la nature des liens du réseau (cf. §2.3.2), nous nous contentons de les exploiter de façon quantitative. On fait plus précisément le choix de définir un mot intéressant, c'est-à-dire en adéquation avec le thème actif, comme un mot entretenant un nombre minimal de relations avec les autres mots retenus à l'issue de la première opération. Ce choix s'appuie sur l'hypothèse que parmi ces mots candidats, on compte toujours un nombre substantiel de mots liés au thème actif et que ceux-ci, du fait justement de cette unité thématique, entretiennent entre eux un plus grand nombre de liens que les mots non spécifiques de ce thème. En retenant les mots à forte connectivité, on a donc plus de chances de retenir des mots relevant du thème actif. Le résultat attendu est une configuration de mots assez fortement liés les uns aux autres par l'entremise du réseau de collocations.

La connectivité d'un mot vis-à-vis de l'ensemble des mots candidats est donc un paramètre de cette seconde opération. En l'occurrence, nous différencions le cas des mots

présents dans la fenêtre glissante de celui des mots sélectionnés à partir du réseau de collocations. Les premiers sont retenus s'ils sont liés à au moins un autre mot candidat, que celui-ci soit un mot de la fenêtre ou un mot du réseau. Les conditions pour les seconds sont plus strictes étant donné qu'ils ne bénéficient pas, au contraire des premiers, du statut que leur confère leur présence explicite dans le texte. Pour être sélectionnés, ils doivent ainsi être liés à au moins 3 mots faisant partie de la fenêtre[1]. La figure 9.6 donne les mots de plus fort poids (cf. étape suivante) sélectionnés à partir du réseau de collocations pour quatre positions du texte de la figure 9.8 (la position est indiquée par le mot du texte qui l'occupe; celui-ci est en outre souligné au niveau de la figure 9.8). On constate ainsi que les mots du réseau sélectionnés sont majoritairement en accord thématique avec le thème du segment dans lequel se trouve la position considérée : le thème de la guerre pour la première position (avec un seuil de sélection fixé à trois mots, le thème de la santé est quasi-absent des mots sélectionnés), celui de la santé pour la deuxième, le thème de la guerre pour la troisième et pour la dernière position. Le degré d'accord est bien entendu variable en fonction de la plus ou moins grande spécificité du segment en question vis-à-vis du thème qu'il évoque.

explosion (pos. 4)	an (pos. 23)	sauter (pos. 37)	aérien (pos. 51)
pentagone (1,81)	hospitalier (0,72)	casque_bleu (1,16)	contingent (1,64)
embuscade (1,68)	médical (0,68)	assaillant (0,89)	casque_bleu (1,00)
blessé (1,63)	médicament (0,64)	coup_de_feu (0,87)	force_multinationale (0,81)
grenade (1,59)	thérapeutique (0,64)	patrouille (0,87)	déploiement (0,80)
patrouille (1,54)	infirmier (0,56)	gaza (0,84)	pentagone (0,66)
évacuer (1,52)	dispensaire (0,56)	force_multinationale (0,83)	envoi (0,62)
tuer (1,51)	médecin_généraliste (0,56)	fusil (0,82))	nations_unies (0,61)
somalien (1,47)	assistance_publique (0,55)	convoi (0,80)	interposition (0,54)
séparatiste (1,44)	hôpital_public (0,55)	israélien (0,80)	onusien (0,52)
tir (1,42)	praticien (0,54)	colon (0,80)	mortier (0,52)

Fig. 9.6 - Mots du réseau de collocations sélectionnés lors du calcul de la cohésion pour quatre positions du texte de la figure 9.8

Dans le cas présent, faire porter l'évaluation de la cohésion non seulement sur les mots de la fenêtre glissante mais également sur un certain nombre de mots sélectionnés à partir

[1] Nous avions au préalable retenu le chiffre de 2 mots et l'évaluation de la méthode présentée au §3 ainsi que les expérimentations menées au chapitre 10 ont été réalisées avec cette valeur. Quelques tests complémentaires nous ont néanmoins montré que la valeur de 3 mots semble donner de meilleurs résultats : on supprime du bruit parmi les mots sélectionnés sans perdre pour autant de mots significatifs vis-à-vis du thème actif. Au delà de 3 mots en revanche, la sélection devient globalement trop sévère et entraîne la perte de trop de mots intéressants.

du réseau de cooccurrences lexicales répond au souci de rendre la détection d'un thème plus stable. Un segment de texte n'est pas nécessairement homogène du point de vue de l'explicitation du thème qu'il véhicule. En général, le vocabulaire qui est spécifique de ce thème n'est donc pas distribué uniformément dans le segment. En ne faisant porter le calcul de la cohésion que sur les mots du texte, on risque de voir celle-ci soumise à des variations importantes, variations qui ne sont pas significatives d'un changement de thème mais seulement indicatives de changements dans la façon dont le thème courant est explicité.

La figure 9.6 montre qu'il est possible d'extraire du réseau de collocations un ensemble de mots en rapport avec le thème courant. Cet ensemble présente l'avantage d'être à la fois beaucoup plus large que le peu de mots spécifiques présents dans les textes et surtout beaucoup plus stable tout au long d'un segment de texte relatif au même thème. De ce fait, on comprend aisément que le calcul de la cohésion d'un texte devienne beaucoup moins sensible aux micro-variations non significatives lorsqu'il intègre dans son champ cet ensemble de mots. Cette moindre sensibilité accroît par ailleurs la fiabilité et la facilité d'exploitation de ses résultats.

Pondération des mots intervenant dans le calcul de cohésion

La pondération des mots vise à caractériser leur importance vis-à-vis du thème actif. Les principes adoptés lors de la première étape pour détecter l'adéquation d'un mot par rapport au thème actif conduisent naturellement à faire l'hypothèse suivante :
soit $MS = \{ms_i\}$, l'ensemble des mots sélectionnés à l'issue de la première étape; un mot ms_i est d'autant plus important dans le contexte courant que le nombre de liens qu'il entretient au sein du réseau de collocations avec les autres mots de *MS* est grand. Le poids d'un mot est donc fonction du nombre de liens qu'il possède avec les autres mots de *MS*. Plus précisément, ce poids est la somme des contributions provenant des autres mots de *MS* avec lesquels il est lié. La contribution d'un mot *m1* au poids d'un mot *m2* est égal au poids initial de *m1*, modulé par la valeur de la mesure de cohésion associée au lien entre *m1* et *m2*. Cette modulation prend en l'occurrence la forme d'un produit.

Le poids initial d'un mot de la fenêtre glissante est égal au nombre d'occurrences de ce mot dans la fenêtre. Le poids initial d'un mot venant du réseau de collocations est quant à lui égal à 0. Un mot de la fenêtre ne reçoit donc de contribution que d'autres mots de la fenêtre. C'est le cas par exemple des mots *m1* et *m2* de la figure 9.7. Du fait du lien existant dans le réseau de collocations entre les deux mots, *m1* envoie une contribution de 0,14 (produit de son poids initial, égal à 1,0, et de la valeur de la cohésion associée au lien entre *m1* et *m2*, égale à 0,14) à *m2* tandis que réciproquement, *m2* envoie une contribution de même valeur à *m1* puisque leurs poids initiaux sont identiques. Il est à

noter qu'un mot de la fenêtre ne faisant pas partie de *MS,* comme *m6,* peut être considéré comme ayant un poids final égal à 0.

Un mot du réseau de collocations ne reçoit lui aussi de contribution que de mots de la fenêtre. Le mot *mr2* de la figure 9.7 est dans cette situation. Son poids est ainsi égal à la somme de son poids initial (égal à 0) et des contributions provenant des mots de la fenêtre *m3* (produit du poids initial de *m3*, égal à 1,0, et de la valeur de cohésion associée au lien entre *mr2* et *m3*, égale à 0,18), *m4* (0,13) et *m5* (0,17) auxquels il est lié par l'intermédiaire du réseau de collocations.

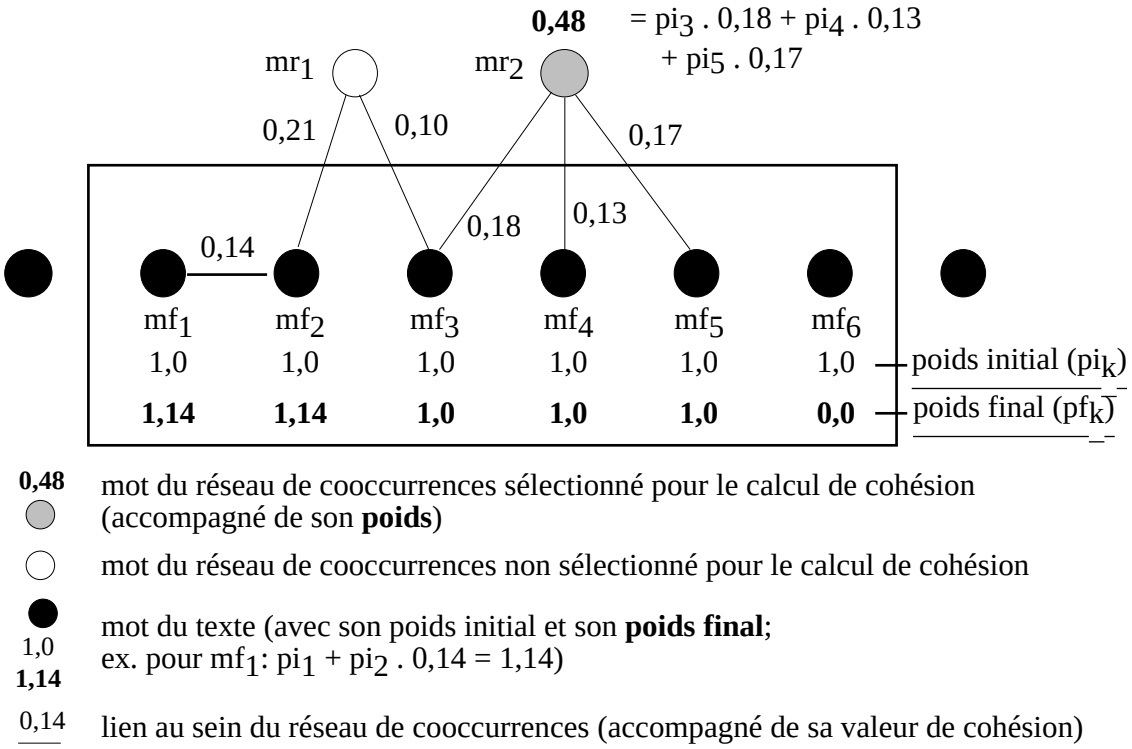


Fig. 9.7 - Calcul du poids des mots

La valeur du poids initial des mots reflète la différence de statut inhérente à leur type : les mots de la fenêtre sont considérés comme premiers et constituent le point de départ d'une forme de diffusion d'activation. De ce point de vue, la pondération des mots de *MS* est une traduction globale du degré d'interaction des mots de la fenêtre, donc du nombre de liens que ceux-ci entretiennent si l'on veut se ramener au critère d'importance énoncé au début de ce paragraphe. Lorsque cette interaction intervient directement entre deux mots de la fenêtre, elle prend la forme d'un renforcement direct du poids des mots en interaction. Elle peut également passer la médiation d'un mot du réseau de collocation, auquel cas c'est le poids de ce mot qui rend compte de l'importance de l'interaction en question. Au total, ces deux modes d'évaluation se trouvent mêlés dans le calcul final de la valeur de cohésion.

Calcul final de la valeur de cohésion associée à une position de la fenêtre glissante

La troisième et dernière étape de l'évaluation de la cohésion consiste à agréger les poids calculés à l'étape précédente afin de former un indicateur unique représentatif de la cohésion de la portion de texte délimitée par la fenêtre glissante. Cet indicateur est égal en l'occurrence à la simple somme pondérée de ces poids :

$$cohésion(p) = \sum_i signif(m_i) \quad poids(m_i)$$

où *poids(m_i)* est le poids du mot m_i appartenant à *MS* calculé selon les principes exposés ci-dessus,
signif(m_i) est la significativité du mot m_i par rapport au corpus de référence du journal *Le Monde* utilisé pour la construction du réseau de collocations (cf. §1.4.2).

Au final, la cohésion résulte donc de la combinaison, pour chaque mot de *MS*, d'un poids dépendant de son contexte d'apparition et d'une mesure générale de sa capacité discriminante sur le plan thématique.

<*ST*> Le soldat américain de l'IFOR blessé par l'explosion d'une mine en Bosnie est arrivé mercredi à l'hôpital militaire américain de Landstuhl (ouest), près de Francfort, a indiqué le **porte-parole** de l'hôpital. </*ST*>
<*ST*> "Son état est stable, il est très fatigué par le voyage" a précisé Marie Shaw. Elle a indiqué ignorer combien de temps Martin John Begosh, 23 ans, devait rester à Landstuhl, soulignant que les médecins devaient encore faire une évaluation et qu'un bulletin de **santé** serait publié jeudi.
"Il se peut qu'il rentre ensuite aux Etats-Unis", a-t-elle dit. </*ST*>
<*ST*> Begosh a été grièvement blessé à la jambe lorsque son véhicule a sauté sur une mine en Bosnie samedi. Il est le premier soldat américain de la force de maintien de la paix de l'OTAN (IFOR) à avoir été blessé. </*ST*>
<*ST*> Le soldat **américain** est arrivé à Landstuhl vers 12H30 (11H30 GMT), via la base militaire américaine aérienne de Ramstein, a précisé Mme Shaw, en provenance de Taszar (Hongrie). </*ST*>

Fig. 9.8 - Exemple de texte considéré (dépêche de l'AFP de janvier 1996)

La figure 9.9 montre le résultat du calcul de cette courbe de cohésion pour le texte exemple de la figure 9.8 en utilisant une fenêtre glissante d'une taille de 19 mots[1]. Une analyse manuelle de cette courbe laisse apparaître trois grandes zones. La première d'entre elles couvre le début du texte jusqu'à approximativement le mot "hôpital". Elle s'identifie assez bien au premier segment thématique mis en évidence par l'analyse manuelle du

[1] Précisons également que seules les collocations d'une fréquence supérieure à 15 et d'une cohésion supérieure à 0,15 sont retenues afin de limiter l'importance du bruit (ces paramètres sont modulables, au contraire du seuil initial de filtrage de 5 occurrences). Avec de telles valeurs, la taille du réseau considéré est donc assez nettement inférieure à celle du réseau dans son entier.

texte[1]. Ce premier segment expose globalement l'information motivant la dépêche. Il regroupe à ce titre les deux thèmes principaux abordés par ce texte : le thème de la guerre et celui de la santé. Le passage du premier au second dans le segment explique la décroissance de la cohésion dans cette zone mais les deux thèmes sont néanmoins suffisamment liés pour que le tout forme un ensemble discernable du point de la cohésion.

La seconde zone, qui s'arrête aux alentours du mot "publié", relève du thème de la santé. Elle est un peu plus petite que le deuxième segment délimité manuellement. Cette imprécision peut s'expliquer par l'absence d'un vocabulaire très marqué. Seuls les termes "médecin" et "santé" sont en effet réellement spécifiques du thème de la santé. Ce manque de spécificité s'observe à la fois qualitativement au travers des mots du réseau de collocations sélectionnés (cf. deuxième colonne de la figure 9.6) et quantitativement sur la courbe de la figure 9.9 : les zones séparant le premier segment du deuxième et le deuxième du troisième forment deux bassins assez larges (de l'ordre de 5 à 6 positions) au milieu desquels le segment n'est identifié que par un pic assez étroit, induit par la présence des quelques mots spécifiques du thème.

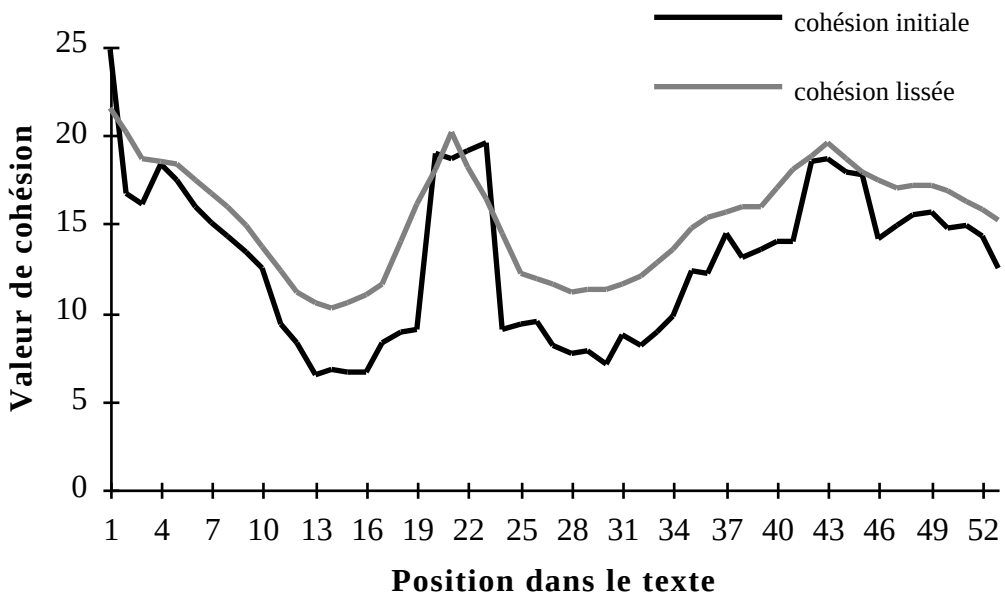


Fig. 9.9 - Courbes de cohésion initiale et lissée pour le texte de la figure 9.8

La troisième zone est plus massivement marquée au niveau de la courbe de cohésion. En revanche, elle ne permet pas de faire aisément la distinction entre les deux derniers

[1] Au niveau de la figure 9.9, les segments thématiques mis en évidence par l'analyse manuelle du texte sont délimités par les balises <ST> et </ST>. Il faut préciser que cette analyse ne résulte pas comme dans [Passonneau & Litman 1993] d'une expérimentation rigoureuse recoupant le jugement de plusieurs sujets.

segments : la chute de cohésion aux alentours de la position 46 n'est pas suffisamment marquée pour qu'une décision certaine puisse être prise. Cette insuffisance est le résultat de la conjugaison de trois facteurs. Tout d'abord, le troisième segment est thématiquement assez fortement marqué. On y trouve des mots comme "blessé", "soldat", "mine", "paix", qui évoquent sans ambiguïté le thème de la guerre qui caractérise ce segment. Au contraire, le dernier segment est très neutre quant à son contenu en ne faisant que rendre compte du voyage d'un militaire. Enfin, les seuls mots un peu évocateurs de ce dernier segment sont des mots relevant du thème de la guerre ("soldat", "base militaire"), donc renvoyant plutôt au segment précédent. En se fondant sur des critères qui, en dépit de l'usage du réseau de collocations, ne sont que des critères de surface, il semble de fait assez peu évident de détecter un changement de thème net entre les deux segments. C'est en l'occurrence une limite intrinsèque de ce type de méthodes.

Sur un plan plus technique, la mise en œuvre de l'évaluation de la cohésion nécessite d'apporter deux précisions par rapport au processus décrit ci-dessus. Tout d'abord, il s'est avéré plus simple et plus efficace d'inverser les deux premières étapes, inversion qui s'accompagne en fait d'une fusion partielle. On considère les mots de la fenêtre les uns après les autres et pour chacun d'entre eux, on récupère dans le réseau de cooccurrences lexicales l'ensemble de ses collocations. On conserve tous les mots ainsi récupérés ainsi que les mots de la fenêtre dans une structure de dictionnaire associant à chaque mot son poids ainsi que le nombre de mots auquel il est lié. Pour un mot donné, ces deux variables sont mises à jour à chaque fois qu'une collocation ramenée par un mot de la fenêtre contient le mot en question. La pondération des mots s'effectue donc de façon progressive en même temps que sont rassemblées les informations permettant de déterminer les mots devant former l'ensemble *MS*. La décision finale quant au contenu de *MS* est rendue lorsque tous les mots de la fenêtre ont été passés en revue de cette façon. À ce moment, la pondération des mots de *MS* a déjà été réalisée.

La seconde précision concerne une optimisation rendue possible par l'absence de filtre (filtre linéaire ou de type cosinus comme dans le cas de Kozima) associé à la fenêtre glissante et par la réversibilité de la fonction d'agrégation des poids des mots. Étant donné la faiblesse du rapport (pas de déplacement / taille de la fenêtre), il est en effet beaucoup plus intéressant de supprimer l'influence des mots sortant de la fenêtre et d'ajouter celle des mots qui y entrent plutôt que de refaire l'évaluation de la cohésion à partir de l'ensemble des mots de la fenêtre. Cet intérêt est particulièrement fort ici dans la mesure où cette optimisation permet de limiter le nombre d'accès au réseau de collocations, accès qui constitue en pratique l'essentiel du temps de traitement.

2.3.2. Segmentation de la courbe de cohésion

Compte tenu de notre approche de la segmentation, segmenter un texte implique de segmenter sa courbe de cohésion en fonction de ses principaux minima, supposés s'identifier aux bornes de segments thématiquement homogènes. Nous faisons appel pour ce faire à une succession de traitements simples. Le premier d'entre eux consiste à lisser la courbe initiale de cohésion afin d'en supprimer les micro-variations, assimilables à du bruit, et faciliter ainsi la détection des minima et des maxima les plus significatifs. Cette opération est réalisée à nouveau en déplaçant une fenêtre sur toutes les positions du texte. À chaque position, la valeur de cohésion associée au centre de la fenêtre est réévaluée pour devenir la moyenne des valeurs de cohésion de toutes les positions incluses dans la fenêtre.

L'effet de lissage obtenu est d'autant plus grand que la taille de la fenêtre est elle-même grande, d'où l'importance de ce paramètre. Il contribue en effet, au même titre que la taille de la fenêtre de calcul de la cohésion (son action est même plus directe et plus forte), à définir le degré de finesse de l'analyse thématique et donc, la taille moyenne des segments dégagés. La seconde courbe de la figure 9.9 montre le résultat de l'application d'un tel lissage sur la courbe de cohésion brute d'un texte, en l'occurrence celui de la figure 9.9. La taille de la fenêtre de lissage est dans ce cas égal à 5 mots. Dans les expérimentations que nous rapportons au chapitre 10, elle a été globalement fixée à 9 mots, sachant que les textes traités étaient généralement plus longs que celui de la figure 9.8. Cette différence met en lumière la nécessité de prévoir ultérieurement une adaptation automatique de la taille de la fenêtre de lissage en fonction de la taille des textes, de façon à coller au plus près à leur structure réelle. La même remarque s'applique à la fenêtre de calcul de la cohésion, bien que l'influence de sa taille sur celle des segments soit dans ce cas plus difficile à cerner.

On notera que contrairement à Kozima mais de façon similaire à Hearst, nous avons choisi de découpler le filtrage du calcul de la cohésion proprement dit. Outre les optimisations qu'elle autorise (cf. ci-dessus), cette option présente l'avantage de laisser plus de latitude quant à la manière de contrôler le lissage. La même préoccupation devrait à l'avenir nous pousser à examiner comment combiner le moyennage réalisé ici avec des filtres moins brutaux (filtre linéaire, filtre de Hanning, ...) afin d'obtenir un spectre plus large de modulation du lissage.

Dans la chaîne de traitement conduisant à la segmentation du texte, l'opération de lissage est suivie de la localisation des minima et des maxima de la courbe de cohésion. Cette localisation permet de réduire la représentation de cette courbe à la seule liste de ses points caractéristiques. Cette liste est en l'occurrence suffisante pour élaborer la version

segmentée de la courbe. En vertu de l'interprétation donnée à la cohésion calculée, les minima s'identifient en effet à des changements thématiques. Un segment est donc caractérisé par la séquence *minimum – maximum – minimum*. La détermination des extrema est réalisée selon les principes classiques de l'analyse, c'est-à-dire en calculant la dérivée de la courbe de cohésion, en notant les passages de cette dérivée par 0 et en supprimant ceux correspondant à des points d'inflexion (cas où la dérivée ne change pas de signe). Bien entendu, il ne s'agit dans le cas présent que d'une estimation assez rudimentaire de la dérivée de la courbe de cohésion. Soient x1, x2 et x3, trois positions du texte successives. La dérivée de cohésion(x2) est donnée par l'expression :

$$\frac{cohésion(x3) - cohésion(x1)}{2}$$

L'étape finale consiste à transformer la courbe de façon à ce que chaque segment soit représenté par un plateau dont le niveau est déterminé par la valeur du maximum situé entre les deux minima encadrant le segment. On obtient de cette manière non seulement une segmentation du texte mais également une évaluation du niveau de cohésion thématique de chacun des segments construits. Nous verrons au chapitre 10 que cette évaluation est utilisable en tant qu'indicateur permettant de se concentrer sur les parties du texte particulièrement homogènes sur le plan thématique.

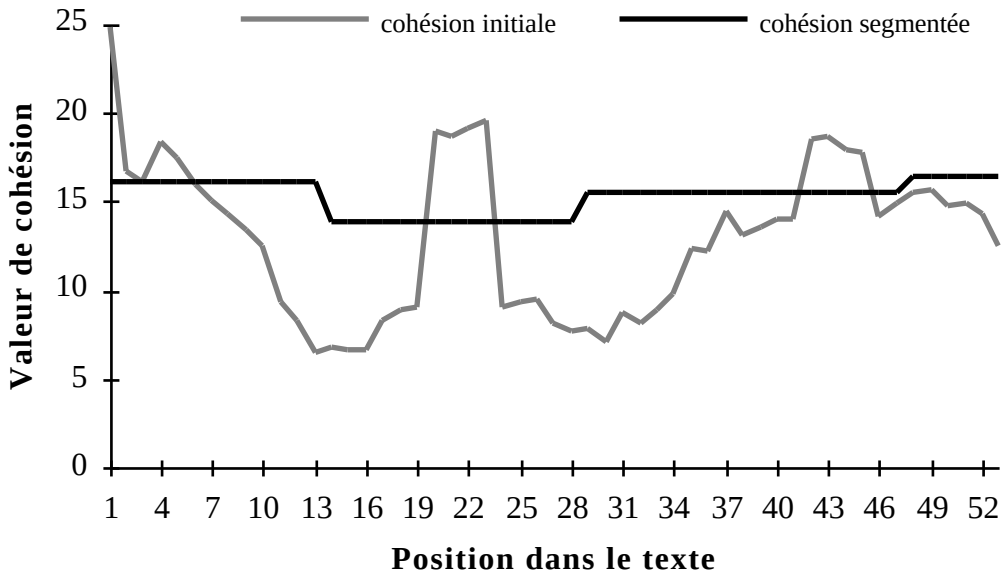


Fig. 9.10 - Courbes de cohésion initiale et segmentée calculées pour le texte de la figure 9.8

La figure 9.10 montre le résultat de l'ensemble du processus de segmentation pour le texte de la figure 9.8. La mise en correspondance des bornes des segments ainsi définis avec le texte est montrée sur la même figure par le formatage en gras des mots occupant la

position de ces bornes. La segmentation automatique confirme pour l'essentiel l'analyse manuelle de la courbe de cohésion initiale que nous avions réalisée ci-dessus. Les trois grandes zones mises en évidence sont matérialisées par trois grands segments : segments 1–13, 14–29 et 30–47. Un dernier segment (48–53), plus petit, se rattache également à la troisième zone et correspond au changement de thème étiqueté précédemment comme incertain. La segmentation de la courbe a dans ce cas permis d'opter pour la solution la plus proche de la segmentation manuelle. Compte tenu de l'ambiguïté attachée à ce type de cas, il faut néanmoins souligner le caractère un peu aléatoire de ce choix à l'échelle d'un ensemble important de textes. Le résultat est alors fortement dépendant de la valeur des paramètres, au contraire des cas dans lesquels les bornes sont plus nettement marquées.

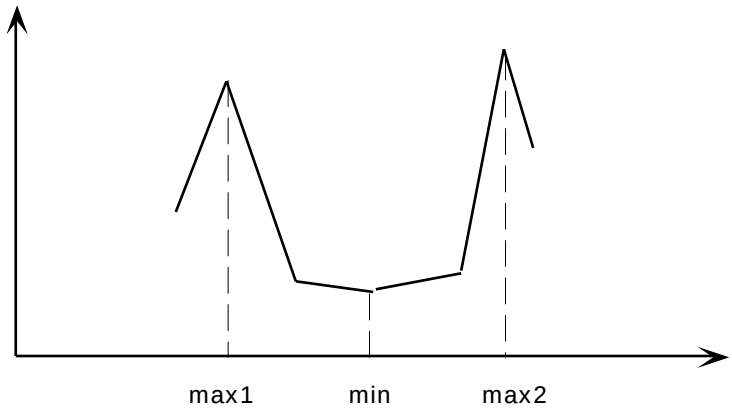


Fig. 9.11 - Configuration nécessitant l'application d'une heuristique spécifique

La chaîne de traitement décrite jusqu'à présent dessine les grandes lignes du processus de segmentation. À ce cadre initial, peuvent venir se greffer des heuristiques destinées à prendre en compte certains problèmes spécifiques. La figure 9.11 illustre l'un de ces problèmes : un minimum se trouve situé au milieu d'un bassin assez large et plat alors qu'il est précédé et/ou suivi d'une cassure assez brutale conduisant au maximum le plus proche. Dans une telle configuration, il peut sembler plus cohérent de situer la borne du segment à l'endroit de la cassure plutôt qu'à l'emplacement du minimum, en particulier si l'objectif est de mettre en évidence les segments les plus homogènes possible sur le plan thématique. Pour ce faire, on applique l'heuristique suivante : on considère chaque maximum comme le point central d'un segment et l'on parcourt les positions de part et d'autre de celui-ci jusqu'à localiser les bornes du segment considéré. Une borne est alors une position du texte répondant à l'une des deux conditions suivantes :

- la valeur de cohésion associée à la position est inférieure à un pourcentage fixé de la valeur du maximum (établi ici à 65%) et la position se situe au moins à une certaine distance (égale dans le cas présent à 5 positions) du prochain minimum;

- la position correspond au minimum le plus proche du maximum considéré dans ce sens de parcours des positions.

Cette heuristique est la seule que nous ayons mise en place pour le moment mais il est assez évident que des heuristiques permettant de prendre en compte d'autres configurations particulières pourraient être développées. Le problème essentiel est alors d'évaluer leur efficacité réelle, ce qui entre dans le champ plus général de l'évaluation des méthodes de segmentation thématique.

3. Évaluation

3.1. Méthodes d'évaluation de la segmentation thématique

À la faveur de la présentation réalisée au paragraphe 1 de ce chapitre des différentes méthodes de segmentation thématique existantes, nous avons eu également l'occasion de présenter deux méthodes utilisées pour évaluer cette tâche. L'une d'elles, qui a été employée par Hearst, Reynar ainsi que Nomoto et Nitta, peut être considérée comme un test de base. Elle consiste à recouper les résultats de la segmentation avec des marques de ruptures thématiques "naturellement" présentes dans les textes. Les marques les plus évidentes sont les frontières de texte et à un degré plus fin, les frontières de paragraphe. Les seuls travaux publiés concernent le recoupement avec des frontières de texte.

Les méthodes de segmentation ayant généralement une granularité inférieure au texte, ce type d'évaluation impose de faire un choix parmi les différentes ruptures trouvées afin de proposer uniquement celles devant correspondre au type de frontière considéré. Bien entendu, la mise en relation d'une frontière avec une rupture fournie par la méthode testée s'effectue toujours avec une certaine tolérance quant à la distance les séparant. Sur la base du résultat de ces mises en correspondance, on calcule une valeur de précision et une valeur de rappel exprimées comme suit. Soient nb_{front}, le nombre de frontières cherchées, nb_{rupt}, le nombre de ruptures assimilées à des frontières et nb_{corr}, le nombre de ruptures correspondant effectivement à des frontières. On a alors :

$$rappel = \frac{nb_{corr}}{nb_{front}} \qquad précision = \frac{nb_{corr}}{nb_{rupt}}$$

La seconde méthode d'évaluation rejoint la première dans la mesure où elle consiste également à essayer de mettre en correspondance des frontières connues avec des ruptures proposées par une méthode de segmentation. Dans ce cas néanmoins, les frontières ne sont plus des marques liées à la forme des textes mais correspondent au résultat d'un

jugement humain. Plus précisément, afin de limiter les effets dus à la subjectivité individuelle, on recoupe le jugement de plusieurs personnes auxquelles on a demandé de localiser les changements de thème au sein d'un ensemble de textes. On retient ensuite comme référence pour la segmentation automatique les frontières communes à une certaine proportion des sujets. Cette méthode a été appliquée par Hearst, par Okumura et Honda, par Kozima dans une moindre mesure et surtout par Litman et Passonneau.

Dans ce type d'évaluation, que la référence soit constituée de jugements humains ou de marques textuelles, il est intéressant de disposer d'un point bas de comparaison des performances. Ce point est généralement constitué par la moyenne (sur un nombre important d'exécutions) des performances d'un processus effectuant la tâche considérée de façon aléatoire. Dans le cas de la segmentation, un tel processus place au hasard, dans chacun des textes intervenant dans l'évaluation, un nombre de bornes identique au nombre de bornes constituant la segmentation de référence.

La troisième et dernière méthode d'évaluation que nous évoquerons n'a été utilisée dans aucun des travaux exposés au paragraphe 1 de ce chapitre. Elle a été initialement proposée dans [Beeferman et alii 1997] et sert maintenant de support à l'évaluation de la tâche de segmentation de l'action Topic Detection and Tracking. Son principe général consiste à évaluer la probabilité que deux mots d'un flot de texte séparés par une distance donnée (en relation avec la taille des segments à mettre en évidence) se trouvent bien classés au regard de leur appartenance à un même segment. Autrement dit, on cherche à déterminer si deux mots appartenant effectivement à un même segment se retrouvent ou non classés dans le même segment par le système de segmentation automatique à évaluer.

	rattachement au même segment	rattachement à des segments différents
appartenance au même segment	(1) classement juste	(3) erreur de détection
appartenance à des segment différents	(2) fausse alarme	(4) classement juste

Fig. 9.12 - Inventaire des mises en correspondance entre les décisions et les situations possibles

Le tableau de la figure 9.12 fait apparaître les quatre cas possibles, considérés du point de vue de la théorie de la détection. Le principe adopté présente l'avantage de s'affranchir indirectement du problème de la mise en correspondance des bornes trouvées avec des bornes de référence. En pratique, on compare les systèmes sur la base des erreurs qu'ils commettent. Les indicateurs retenus sont donc la probabilité du cas (2) et

celle du cas (3). Ces probabilités sont estimées à partir d'un ensemble important de couples de mots.

3.2. Évaluation de la segmentation thématique de SEGCOHLEX

Des trois méthodes d'évaluation présentées ci-dessus, nous n'avons véritablement appliqué que la première. Nous avons implicitement appliqué la deuxième lors de la mise au point de la méthode mais on ne peut pas parler d'évaluation dans ce cas car le jugement humain de référence n'est pas allé au delà de notre propre intuition, de surcroît non formalisée. Il s'agit plutôt de la confrontation qualitative des résultats de la méthode présentée avec notre jugement sur une vingtaine de textes. Cela nous a permis en particulier d'ajuster les divers paramètres de la méthode. Nous avons ainsi pu constater que les résultats obtenus, même s'ils évoluent un peu en fonction des réglages adoptés, restent globalement assez stables. La taille de la fenêtre de calcul de la cohésion ainsi que la taille de la fenêtre de lissage influent bien évidemment sur la taille et le nombre des segments formés mais dans une plage de valeurs allant de 9 à 21 mots, on n'observe que de faibles variations. Même en dehors de cet intervalle, les grandes tendances restent assez marquées. Afin d'obtenir néanmoins des indications plus objectives, nous avons appliqué notre méthode au problème "classique" de la redécouverte des frontières d'un ensemble de textes concaténés, problème relevant de la première méthode d'évaluation présentée au paragraphe précédent.

Le protocole adopté pour cette évaluation reprend celui présenté dans [Hearst 1997], en dépit de la différence entre les deux approches. L'algorithme de segmentation fournit un ensemble de ruptures que l'on classe dans l'ordre décroissant de leur importance, en faisant l'hypothèse que les frontières de textes sont plus marquées que les autres ruptures. On fixe ensuite un nombre nb_{rupt} de ruptures et l'on détermine le nombre de ruptures nb_{corr}, parmi ces nb_{rupt} plus importantes, qui correspondent à des frontières de texte. La comparaison avec le cadre présenté ci-dessus met en évidence que dans ce protocole, nb_{rupt} n'est pas établi par la méthode évaluée mais prend la forme d'un seuil arbitraire. La précision et le rappel conservent néanmoins la même définition : la précision juge toujours quelle proportion des ruptures les plus marquées sont des frontières de texte tandis que le rappel évalue ce que représentent les ruptures bien placées par rapport à l'ensemble des frontières.

L'évaluation que nous avons réalisée, qui reste pour le moment préliminaire, a porté sur un ensemble de 39 textes extraits du journal *Le Monde*, textes ayant en moyenne une

taille de 80 mots environ. Pour valuer l'importance des ruptures, nous avons également repris la procédure décrite par Hearst. Chaque rupture, qui correspond en fait à un minimum de cohésion, est pondérée par la somme des différences entre ce minimum et les deux maxima qui l'entourent. On mesure ainsi la profondeur du minimum non de façon absolue mais par rapport aux valeurs environnantes. La tolérance dans la mise en correspondance des frontières et des ruptures est un intervalle de plus ou moins 9 mots (après pré-traitement). Les résultats obtenus pour différentes valeurs de nb_{rupt} sont synthétisés dans le tableau de la figure 9.13. Ce tableau donne à la fois les résultats de la segmentation de SEGCOHLEX sur les 39 textes de notre jeu de test et les résultats de TextTiling de Hearst sur un jeu de test de 44 articles de journaux. nb_{rupt} *max* correspond au nombre total de ruptures thématiques délivrées par chacune des méthodes et la notation "38*" ou "43*" signale la valeur de nb_{rupt} correspondant au nombre réel de frontières de textes pour chacun des jeux de test.

nb_{rupt}	nb_{corr} (S/H)	Précision (S/H)	Rappel (S/H)
10	5 / 8	0,5 / 0,8	0,13 / 0,19
20	10 / 16	0,5 / 0,8	0,26 / 0,37
30	17 / 22	0,58 / 0,73	0,45 / 0,51
38*	19 / –	0,5 / –	0,5 / –
40	20 / 27	0,5 / 0,68	0,53 / 0,63
43*	– / 29	– / 0,67	– / 0,67
50	24 / 31	0,48 / 0,62	0,63 / 0,72
60	26 / 36	0,43 / 0,6	0,68 / 0,83
67 (nb_{rupt} max S)	26/–	0,39 / –	0,68 / –
70(nb_{rupt} max H)	– / 41	– / 0,59	– / 0,95

Fig. 9.13 - Précision et rappel de TextTiling (H) et de SEGCOHLEX (S) pour une tâche de redécouverte de frontières de textes[1]

Globalement, les résultats que nous obtenons sont moins bons que ceux de Hearst. Il faut préciser cependant que le champ d'application privilégié des deux méthodes n'est pas le même. Hearst élimine les textes de moins de 10 phrases dans la mesure où ils ne peuvent être traités par TextTiling. Notre évaluation porte au contraire sur des textes de cet ordre de grandeur. De ce point de vue, les deux évaluations ne sont donc pas directement comparables. Il n'existe cependant pas de travaux ayant mené ce type d'évaluation à un degré de granularité moindre que celui de textes assez longs. Ceci peut s'expliquer en partie par le manque de fiabilité des unités plus fines telles que le paragraphe. Nous n'avons donc pas eu d'autre choix que de prendre comme point de comparaison l'évaluation détaillée la plus proche, en l'occurrence celle de Hearst, en étant

[1] La notation "–" signifie simplement que la valeur n'a pas été mesurée pour la méthode correspondante

conscient que les différences de performance observées sont sans doute autant le fruit d'une différence intrinsèque de performance des deux systèmes que le produit de la différence des contextes d'application.

Nous estimons ainsi qu'il est plus difficile d'obtenir de bons résultats pour une segmentation fine que pour une segmentation de plus gros grain, comme celle consistant à retrouver des frontières de texte. Par ailleurs, une méthode fondée sur la simple réitération des mots est naturellement assez efficace pour accomplir la seconde tâche : à moins d'un fort recouvrement thématique, deux textes adjacents se caractérisent par un vocabulaire assez différent. On retrouve donc peu de liens de réitération franchissant les frontières de texte, ce qui facilite la détection de ces dernières. À l'inverse, une méthode telle que celle de SEGCOHLEX a pour objectif de trouver des liens échappant aux simples réitérations. Ces liens sont intéressants pour établir des liaisons entre des passages de texte lorsqu'une grande variabilité de l'expression induit une faible réitération des mêmes mots. Ils peuvent néanmoins constituer une gêne dans la détection des coupures franches. Tous les liens mis en évidence ne sont pas nécessairement pertinents au regard du contexte (cf. réseau de cooccurrences lexicales) et leur présence contribue dans ce cas à diminuer le contraste existant entre les deux parties de texte considérées.

On voit ainsi que chacune de ces deux méthodes est plus spécifiquement adaptée à un ordre de grandeur donné d'unités textuelles. Celle que nous présentons permet un suivi des évolutions fines des thèmes d'un texte mais se révèle moins efficace qu'une méthode fondée sur la simple réitération des mots pour la mise en évidence de changements plus radicaux, comme ceux existant entre les textes.

Les résultats de la figure 9.13 montrent par ailleurs que la valuation des ruptures de thème semble ici peu significative. L'évolution globalement assez lente et peu importante de la précision à mesure que le nombre de ruptures retenues augmente en est une indication. Même en introduisant quelques variations dans la politique de valuation de ces ruptures, nous n'avons pas réussi à obtenir une modification significative de ce phénomène. Une part de cet effet est imputable à la faiblesse des résultats globaux concernant la segmentation (il suffit pour s'en convaincre de prendre les valeurs de la ligne correspondant à $nb_{rupt}\,max$) : en tout état de cause, si aucune rupture thématique ne s'accorde avec une frontière de texte, peu importe la politique de choix des ruptures. Mais une part également importante de cet effet puise son origine dans l'inadéquation de cette politique de choix. La valeur de cohésion est en pratique un indicateur trop frustre pour juger de façon fiable de l'ampleur d'un changement thématique. Il serait plus significatif de s'appuyer directement sur les mots sélectionnés à partir du réseau de collocations et de s'intéresser en particulier à leur taux de renouvellement.

Nous en terminerons avec cette évaluation en remarquant qu'une façon d'améliorer les résultats obtenus serait de se ramener à des frontières plus naturelles, c'est-à-dire dans notre cas, de faire correspondre les ruptures de thème trouvées avec les frontières de phrase les plus proches d'elles.

Pour achever plus généralement cette discussion sur l'évaluation de la segmentation thématique de SEGCOHLEX, il faut évoquer la possibilité de mener une évaluation indirecte. Ce mode d'évaluation, dont nous n'avons pas parlé au §3.1, consiste à comparer les performances de différentes méthodes assurant la mise en œuvre d'une tâche au travers de leur impact sur une tâche plus globale à laquelle contribue la première. Il s'agit donc d'une évaluation in situ. Nous verrons au chapitre 10 qu'une telle évaluation a été esquissée pour juger de l'impact de la segmentation thématique de SEGCOHLEX sur une tâche de construction de représentations de thèmes. Ce jugement a été réalisé par comparaison avec une situation de référence dans laquelle les textes ne sont pas segmentés.

4. Implémentation

Du point de vue de son implémentation, SEGCOHLEX se divise en trois grandes parties : le pré-traitement des textes, la constitution du réseau de cooccurrences lexicales et la segmentation thématique. Nous avons vu au §2.3.1 que l'essentiel du pré-traitement est réalisé par des outils pré-existants, que ce soit le segmenteur *Mtseg* ou l'étiqueteur morpho-syntaxique *TreeTagger*. Sachant que ces outils sont écrits en C sous la forme de programmes autonomes et que Smalltalk ne se prête pas très bien à la réalisation de chaînes de traitements batch, il nous est apparu plus commode d'implémenter le pré-traitement des textes en dehors de Smalltalk. Celui-ci s'articule autour d'un script Cshell principal lançant les différentes opérations en séquence (cf. annexe G pour avoir le détail de cette chaîne). En dehors de la segmentation, de l'étiquetage morpho-syntaxique et de la sélection finale des mots pleins, ces opérations ont en charge tous les travaux de conversion de format permettant de rendre compatibles les uns avec les autres les formats d'entrée et les formats de sortie des outils pré-existants. Elles sont réalisées par de petits programmes écrits en Awk ou en Perl, encapsulés afin d'enregistrer la valeur d'un certain nombre d'indicateurs. Ceux-ci permettent de contrôler a posteriori que des défauts d'alignement n'ont pas été introduits par le dysfonctionnement d'un outil.

La construction du réseau de cooccurrences lexicales est implémentée quant à elle par une chaîne de traitements écrite en Smalltalk mais faisant aussi appel à des utilitaires Unix de base tels que le programme Sort. Le cœur de cette construction est constitué par

l'enregistrement des cooccurrences à partir d'un fichier texte. Compte tenu du nombre de cooccurrences produites (il est très importante du fait de la taille de la fenêtre) et de l'absence de filtrage des cooccurrences de faible fréquence en cours de traitement, il est cependant impossible d'enregistrer les cooccurrences sur l'intégralité du corpus en une seule fois. Seuls des fichiers équivalents à la moitié environ d'un mois du journal *Le Monde* peuvent être traités efficacement. Au delà, la plupart des collocations ne se trouvent plus en mémoire[1] mais sur des fichiers de débordement, ce qui ralentit le processus de façon inacceptable. Il est en fait plus efficace de découper le corpus en une série de fichiers de taille adéquate, d'enregistrer les cooccurrences pour chacun d'entre eux de façon indépendante et finalement de fusionner l'ensemble de ces résultats, ce qui est rendu possible par le caractère additif des cooccurrences entre deux mots. On pourra se reporter à l'annexe H pour avoir plus de précisions sur l'outil permettant de contrôler cette chaîne de traitements.

La segmentation thématique est implémentée pour sa part entièrement en Smalltalk. Elle s'appuie sur un gestionnaire du réseau de collocations assurant un accès transparent à ce dernier, quelle que soit la façon dont ce réseau est physiquement représenté. Compte tenu de sa taille, celui que nous avons construit ne peut être présent en mémoire sous la forme d'objets Smalltalk. Il est donc matérialisé par un ensemble de fichiers texte. Seul un index se trouve en mémoire. Le principal facteur limitant de la segmentation thématique est donc le temps d'accès au réseau. Nous avons vu précédemment (cf. 2.4.1) que dans le mécanisme même de segmentation, nous avons introduit des optimisations afin de restreindre le plus possible les accès au réseau. Au niveau du gestionnaire de réseau, la présence d'un cache va dans le même sens mais de façon plus générique. Enfin, au sein des fichiers contenant le réseau, les collocations d'un mot sont triées suivant l'ordre décroissant de leur valeur de cohésion. On peut ainsi ne charger qu'une partie de ses collocations lorsqu'un seuil minimal de cohésion est fixé. Il faut préciser que contrairement au cas de la construction du réseau, un changement de langage d'implantation permettrait sans doute de travailler avec le réseau présent dans son entier en mémoire et donc, d'accroître très significativement les temps d'accès.

Plus en aval, la segmentation thématique est supportée par un outil de test permettant de lancer ses différents traitements, de visualiser les courbes de cohésion associées, de les mettre en correspondance avec le texte traité et enfin, de contrôler ses nombreux paramètres. Cet outil est présenté lui aussi plus concrètement à l'annexe H.

---

[1] Cette affirmation dépend évidemment du cadre d'implantation utilisé. En l'occurrence, nous avons travaillé sur une Sun UltraSparc 1 sous Solaris dotée de 128 Mo de mémoire centrale. Il est par ailleurs indéniable que Smalltalk ne permet pas forcément de gérer aussi finement le volume mémoire occupé par les structures de données que le langage C. Certaines optimisations seraient possibles en changeant de langage mais elles n'ouvriraient pas plus la possibilité de traiter l'ensemble du corpus d'un seul tenant (sauf à disposer de mémoires de l'ordre du giga-octet)

5. Discussion et extensions possibles

Même sans évaluation très poussée, tout laisse à penser que la segmentation thématique de SEGCOHLEX apporte un plus dans une tâche de construction de représentations de thèmes comme celle décrite au chapitre 10 mais que ses performances concernant spécifiquement la tâche de segmentation ne sont pas excellentes. Bien que la démarche générale d'ANTHAPSI n'exige pas que les performances de SEGCOHLEX soient très bonnes, il est tout de même intéressant d'examiner toutes les améliorations possibles que l'on peut y apporter, dès lors qu'elles ne conduisent pas à sortir du cadre de définition de SEGCOHLEX.

À l'occasion de l'évaluation de la segmentation, nous avons déjà mis en avant le fait que la cohésion calculée nous semble un indicateur trop réducteur pour juger de la présence d'un changement thématique. Cette opinion est motivée par la constatation, faite à de nombreuses reprises, que les mots sélectionnés à partir du réseau de collocations peuvent être pertinents vis-à-vis du thème courant[1] alors même que les bornes de segment mises en évidence par la courbe de cohésion ne sont pas bien placées. Pour éviter cette apparente perte d'information, il apparaît donc préférable de travailler directement sur la base des mots sélectionnés. La solution la plus évidente dans ce cadre reprend le principe développé au chapitre 8 : on fonde la décision de changement de thème sur l'observation de l'évolution des éléments de la mémoire qui sont sélectionnés. Il s'agit ici de mots et non d'UTs agrégées mais cela suppose, selon une démarche identique, de développer une mesure de similarité entre deux configurations de mots du réseau de collocations, donc par voie de conséquence entre deux positions de la fenêtre glissante.

Outre une amélioration des principes de la segmentation, la manière la plus évidente d'améliorer la segmentation thématique de SEGCOHLEX est de se pencher sur le réseau de cooccurrences lexicales, qui est la seule source de connaissances exploitable par un mécanisme de segmentation de SEGCOHLEX. Les sources potentielles de perfectionnement sont nombreuses dans la mesure où les paramètres sont eux-mêmes nombreux : taille de la fenêtre d'enregistrement des collocations, configuration de la fenêtre (si l'on est plutôt intéressé par les relations à longue distance, on peut placer un cache dans la fenêtre tel que les cooccurrences avec les mots les plus proches ne sont pas enregistrées), choix des catégories de mots à retenir, traitement de normalisation des mots (ici, on ne fait que se ramener à la forme canonique des mots du point de vue de la

[1] Les résultats présentés au chapitre 10 sur la structuration du réseau de collocations vont également dans ce sens.

morphologie flexionnelle mais on pourrait imaginer de pousser plus loin la normalisation en faisant également intervenir la morphologie dérivationnelle par l'utilisation du stemming), considération plus ou moins étendue des mots composés, mesure de la cohésion, etc.

Néanmoins, tester la plupart de ces paramètres impliquerait de reconstruire le réseau de collocations pour chaque modalité testée. Compte tenu de la lourdeur du processus, on imagine aisément que les possibilités d'exploration de différentes valeurs sont assez limitées. Par ailleurs, le problème de l'évaluation est encore une fois posé. À moins de définir des propriétés du réseau en relation directe avec la segmentation thématique, la principale évaluation est indirecte : on compare deux réseaux de collocations par leur influence sur les performances de la segmentation thématique. La procédure est donc très pesante à mettre en œuvre.

Son application se trouverait bien entendu facilitée par une taille plus réduite du volume de textes à traiter pour construire un réseau contenant des collocations significatives. En toute généralité, la taille ne suffit sans doute pas. Il faudrait aussi caractériser d'une façon ou d'une autre le contenu du corpus; mais il s'agit là d'un problème très vaste et assez peu exploré (cf. [Biber 1993] tout de même). Pour le moment, la taille du corpus nécessaire est fixé de façon très arbitraire et plutôt en suivant une politique de majoration systématique. Un ajustement plus précis serait particulièrement intéressant, surtout s'il conduit à réduire significativement la taille des corpus requis.

Nous avons eu l'occasion en effet de voir qu'un réseau de collocations reste assez fortement attaché au corpus utilisé pour sa construction. Il est donc clair que des réseaux différents doivent être construits pour des domaines et des types de textes différents. Dans cette optique, la constitution d'un réseau de collocations n'est pas une opération exceptionnelle. La taille du corpus à former pour construire le réseau devient donc un élément important, qu'il convient de minimiser le plus possible.

Précisons tout de même que la construction complète d'un nouveau réseau pour chaque nouveau contexte n'est pas la seule solution envisageable. Une voie moins onéreuse d'adaptation à un domaine et un type de texte particuliers consiste à essayer de "teinter" un réseau généraliste, ou tout du moins supposé comme tel, avec des collocations extraites d'un corpus propre à ce domaine (l'opération est facilitée par le caractère additif des occurrences de collocations). La validation d'une telle approche reste encore à faire et là encore, on n'échappe pas au problème de la détermination du volume du corpus nécessaire à cette opération de "teinture".

La dernière piste d'amélioration que nous évoquerons à propos du réseau de collocations est la possibilité de réduire l'influence du bruit caractérisant un tel réseau en

le structurant davantage. Le chapitre 10 apporte un exemple d'une telle structuration, en l'occurrence sur le plan thématique. Le but de SEGCOHLEX n'étant pas de faire émerger un type spécifique de connaissances, on peut envisager à ce stade de réaliser une structuration assez générale et indifférenciée, à l'image du réseau lui-même. Dans ce contexte, chaque mot peut être vu comme une entité caractérisée par un ensemble d'attributs pondérés. Dans le cas présent, les attributs d'un mot sont les mots avec lesquels il est lié dans le réseau de collocations (au moins les plus significatifs) et le poids d'un de ces mots-attributs correspond à la valeur de cohésion associée à la collocation avec ce mot. En définissant une distance entre deux mots, représentés comme ci-dessus, il est tout à fait possible d'appliquer des algorithmes de classification usuels (k-noyaux, classification hiérarchique) afin de regrouper les mots du réseau en classes et doter ainsi celui-ci d'une structure de plus haut niveau que les simples relations entre mots[1].

Lors de l'évaluation de la cohésion, au cours de la segmentation, on peut s'appuyer sur l'appartenance des mots venant du réseau de collocations à l'une de ces classes ou à un petit groupe d'entre elles pour faire la distinction entre les mots significatifs pour le calcul de la cohésion, i.e. les mots appartenant à cette ou ces classes, et ceux qui ne constituent que du bruit, i.e. les mots n'appartenant pas à cette ou ces classes. Compte tenu du caractère générique du critère de formation des classes, il ne faut pas s'attendre à une discrimination parfaite entre mot significatif et bruit mais on peut raisonnablement penser qu'un tel procédé devrait permettre d'éliminer une part importante du bruit sans pour autant faire disparaître trop de mots intéressants pour l'évaluation de la cohésion.

Récapitulatif

Dans ce chapitre, notre attention s'est portée sur la description de SEGCOHLEX (SEGmentation thématique par utilisation de la COHésion LEXicale), dont l'objectif essentiel est de fournir une segmentation thématique des textes robuste en vue d'amorcer celle de SEGAPSITH lorsque celle-ci ne dispose pas des connaissances pour agir. Nous avons donc commencé par nous intéresser aux méthodes quantitatives de segmentation thématique des textes. Le tour d'horizon que nous avons réalisé (cf. annexe I pour l'intégralité de ce panorama) laisse apparaître que ce problème a été abordé sous différents angles mais qu'aucune évaluation globale ne permet pour le moment de déterminer si l'un d'entre eux est plus approprié que les autres. Par ailleurs, le champ d'application de chacune des méthodes n'est guère mieux défini. L'axe d'analyse privilégié reste donc celui des moyens mis en œuvre pour accomplir la segmentation. De ce point de vue, les

[1] Nous tenons à remercier Benoît Habert pour nous avoir suggérer cette idée et plus généralement pour nous avoir fait profiter de son expérience concernant les problèmes de constitution et de traitement de corpus.

méthodes considérées se partagent entre celles opérant uniquement à partir de la distribution des mots dans les textes et celles mobilisant des ressources extérieures aux textes. Les premières, dont le prototype est la méthode TextTiling de Hearst, s'inspirent assez largement des techniques utilisées en Recherche d'Informations en transformant des portions de texte en vecteurs de mots pondérés et en ramenant la comparaison de ces morceaux de texte au calcul d'une mesure de similarité entre les vecteurs chargés de les représenter. La forme des secondes est moins homogène dans la mesure où elle est dépendante de la nature de la source de connaissances utilisée : un thesaurus dans le cas de Morris et Hirst (cf. annexe I) et un dictionnaire dans celui de Kozima.

En l'absence de référence bien établie, notre choix pour le processus de segmentation thématique de SEGCOHLEX s'est porté sur le type de méthode répondant le mieux aux deux critères suivants : être adapté au type de textes que nous traitons ici et offrir une marge de progression potentiellement importante. La grande variabilité d'expression des textes à forte composante narrative ajoutée aux limites intrinsèques du premier type de méthodes[1] nous ont orienté vers une méthode fondée sur une source de connaissances. En l'occurrence, il nous est apparu que la possibilité d'adapter cette source de connaissances de façon automatique à un nouveau domaine constituait également un critère important. C'est pourquoi nous avons opté en faveur d'un réseau de cooccurrences lexicales, même si celui-ci s'avère moins précis et plus incertain qu'un thesaurus ou bien un dictionnaire. Le réseau que nous avons utilisé a été construit à partir d'un vaste ensemble de textes du journal *Le Monde* (environ 39 millions de mots). Ses paramètres de construction ont été ajustés de manière à privilégier les relations de nature sémantique et pragmatique. Le résultat est un réseau de près de 31000 mots et de 7 millions de relations.

La méthode de segmentation proprement dite s'inspire de celle élaborée par Kozima. Une fenêtre glissante est déplacée sur l'intégralité du texte à analyser et à chaque station de la fenêtre, une mesure de la cohésion des mots présents au sein de cette fenêtre est calculée. On obtient ainsi une courbe rendant compte de la cohésion de l'ensemble des parties du texte considéré. La mesure de cohésion est calculée sur la base du nombre et de la force des relations que les mots de la fenêtre entretiennent entre eux au sein du réseau de collocations. Au niveau de la courbe de cohésion globale, les zones de faible cohésion représentent des groupes de mots peu homogènes, donc n'appartenant a priori pas au même domaine. Ce sont donc les zones présumées de changement de thème. Cette propriété est exploitée afin de segmenter la courbe de cohésion de façon automatique en

[1] On peut toujours élaborer une nouvelle mesure de similarité ou une façon particulière d'analyser la distribution des mots mais en l'absence de connaissances externes, on reste toujours prisonnier de la forme des textes.

utilisant une succession de traitements simples : lissage par utilisation d'une fenêtre de moyennage, calcul de la dérivée, repérage des extrema et transformation en courbe à plateaux. On forme de cette manière un ensemble de segments présentant en outre la propriété intéressante de se voir dotés chacun d'un niveau de cohésion, caractérisant en quelque sorte sa "qualité thématique".

La méthode proposée a été évaluée sur une tâche "classique" de redécouverte des frontières d'un ensemble de textes ayant été concaténés. Les résultats obtenus, inférieurs à ceux de Hearst pour une tâche similaire mais un type de textes différent, montrent que cette méthode est plus adaptée au suivi fin des variations thématiques et au repérage des zones de texte particulièrement homogènes sur ce même plan qu'au découpage des textes en grandes sections thématiquement distinctes.

Nous avons vu par ailleurs qu'un certain nombre de pistes existent pour améliorer la méthode. La plus prometteuse en l'espèce consisterait à adapter dans le cadre de SEGCOHLEX les principes développés au chapitre 8 : au lieu de fonder la détection des changements de thèmes sur les seules variations d'une valeur de cohésion, qui est un indicateur assez réducteur, la segmentation s'appuierait directement sur l'évolution des mots sélectionnés à partir du réseau de collocations.

Chapitre 10

SEGAPSITH

Dans ce chapitre, nous présentons la seconde composante de ROSA, SEGAPSITH. Comme MLK, celle-ci s'articule autour de deux dimensions complémentaires. La dimension apprentissage est incarnée par l'extraction automatique de signatures thématiques à partir de textes préalablement segmentés en blocs thématiquement homogènes. Ces signatures prennent la forme d'ensembles de mots pondérés. Comme dans MLK, on fait appel pour leur constitution à un mécanisme de mesure de similarité et d'agrégation proche de la notion de regroupement conceptuel. La seconde dimension de SEGAPSITH est une analyse des textes consistant en une segmentation thématique similaire dans la forme de ses résultats à celle de SEGCOHLEX. À la différence de cette dernière toutefois, la segmentation de SEGAPSITH opère en utilisant comme connaissances de référence les signatures thématiques extraites dans le cadre du processus d'apprentissage. Nous terminons la présentation de cette composante de ROSA en montrant comment s'effectue son amorçage par SEGCOHLEX et comment elle-même amorce MLK.

Nous avons déjà fait apparaître au chapitre 3 que SEGAPSITH se présente comme la transposition de MLK dans un contexte où les pré-requis nécessaires à l'implantation complète du modèle proposé sont inclus dans l'état de l'art actuel. SEGAPSITH est donc composé de deux éléments en étroite interdépendance : une mémoire dotée d'un mécanisme d'apprentissage incrémental permettant de faire émerger des représentations de thèmes et un processus d'analyse thématique des textes utilisant ces représentations. Nous commencerons par détailler le processus d'émergence des représentations de thèmes.

1. L'extraction de signatures thématiques

1.1. Introduction

Comme dans le cas de la mémoire épisodique de MLK, l'objectif essentiel est ici de mémoriser des segments de texte thématiquement homogènes et d'en faire émerger progressivement, par un mécanisme de détection de similarité et d'agrégations successives entre segments, une représentation générale des thèmes abordés dans les textes traités. Les représentations ainsi construites sont à leur tour utilisées par la segmentation de SEGAPSITH pour analyser de nouveaux textes.

Nous rappelons que les segments de texte issus de cette analyse sont appelés des Unités Thématiques Lexicales (UTLs), par analogie avec les Unités Thématiques de MLK et par référence au fait que dans SEGAPSITH, elles ne sont composées que de mots. Toujours par analogie avec MLK, les représentations de thèmes sont appelées des Unités Thématiques Lexicales agrégées. Nous emploierons également le terme de *signature thématique* [1]. Le remplacement du mot "situation" par celui plus général de "thème" traduit la plus grande imprécision des représentations construites dans SEGAPSITH. Nous conservons l'idée de produire la représentation de situations prototypiques mais la faible structuration des UTLs ne permet pas forcément de maintenir cette ligne directrice de façon aussi précise que dans MLK.

1.2. Les travaux relatifs à la construction automatique de représentations de thèmes

1.2.1. Construction de représentations de thèmes et catégorisation de textes

La construction de représentations de thèmes n'apparaît pas, au même titre par exemple que la segmentation thématique, comme un champ de recherche autonome et clairement identifié. Un seul travail à notre connaissance, celui de Lin [Lin 1997], que nous évoquerons plus en détail au §1.2.2, fait d'ailleurs explicitement référence à ce problème. En revanche, la catégorisation de textes est un champ de recherche voisin et très actif. L'objectif de cette tâche est de classer une collection de textes suivant un ensemble de catégories pré-définies. Ces catégories sont généralement de nature thématique, comme dans le cas par exemple des dépêches d'agences de presse[2]. Mais la catégorisation de textes en tant que telle s'apparente essentiellement à une tâche d'identification thématique, c'est-à-dire à la détermination du thème d'un texte.

La proximité avec la construction de représentations de thèmes se situe concrètement au niveau de la tâche d'élaboration d'une caractérisation des différentes catégories considérées. Cette tâche vient en amont de la catégorisation et constitue un pré-requis indispensable à cette dernière puisque son résultat sert ensuite de support au classement des textes. Les caractérisations des catégories sont généralement construites automatiquement à partir d'un ensemble de textes de référence ayant été classés manuellement. Néanmoins, toutes les caractérisations de ce type ne peuvent pas être

[1] Dans [Ferret & Grau 1998] et [Ferret & Grau 1998], on trouve également le terme de domaine sémantique.

[2] Il existe notamment un corpus de dépêches de l'agence Reuters ayant servi de référence commune pour l'évaluation de différents travaux dans le domaine.

considérées comme des représentations des thèmes sous-jacents à ces catégories. Il est possible en effet de caractériser une catégorie soit par la donnée des traits la définissant intrinsèquement, soit par la mise en évidence des traits la distinguant des autres catégories en présence. Dans le premier cas, la caractérisation résultante est directement utilisable en tant que telle dans une tâche de reconnaissance et constitue une représentation autonome du thème lié à la catégorie. Dans le second cas au contraire, elle ne permet que d'opérer une discrimination parmi un ensemble déterminé de catégories et ne peut donc servir de représentation générale du thème de la catégorie. Bien entendu, le recoupement entre les caractérisations produites par les deux approches n'est bien souvent pas nul.

La seconde approche a particulièrement été explorée par des travaux cherchant à appliquer des algorithmes d'apprentissage automatique au problème de la catégorisation de textes [Apté et alii 1994, Lewis & Ringuette 1994, Moulinier et alii]. À un niveau de représentation plus élevé, il faut également citer [Riloff & Lehnert 1994], travail dans lequel la caractérisation des catégories n'est pas construite seulement sur la base des mots des textes mais prend la forme de véritables petits schémas construits par des techniques d'extraction d'information à partir des textes de référence destinés à cerner la nature des catégories. Le choix des informations extraites est néanmoins guidé par le souci de discriminer les différentes catégories plus que par celui de construire une représentation générale des thèmes abordés.

La première approche, qui retient plus spécifiquement notre intérêt ici, a été principalement développée par des travaux issus du domaine de la Recherche d'Information. Elle est donc dominée par le modèle lui-même dominant dans ce champ de recherche, autrement dit le modèle *vector-space*, qui est évoqué à l'annexe I à propos du travail de Nomoto et Nitta (cf. §1.3) : un texte est représenté par un vecteur possédant autant de dimensions que de termes. Chaque terme est pondéré en fonction de son importance dans le texte. La proximité entre deux textes est jugée par une mesure de similarité entre les vecteurs qui les représentent.

Dans ce contexte, il n'est pas surprenant que la représentation d'une catégorie, définie par un ensemble de textes de référence, soit elle-même un vecteur de termes pondérés. Plus précisément, ce vecteur s'identifie au vecteur centre de gravité des vecteurs représentant les textes de référence de la catégorie. Il est également appelé *vecteur centroïde* [Salton et alii 1994]. Ce vecteur centroïde présente l'avantage de notre point de vue de représenter la catégorie concernée, et donc son thème associé, indépendamment des autres catégories. La catégorisation des textes s'effectue pour sa part suivant la même logique globale. Un nouveau texte à classer est transformé en vecteur, lequel est ensuite comparé, au moyen d'une mesure de similarité, aux vecteurs représentant les différentes

catégories considérées. Le texte est finalement rattaché à la catégorie dont le vecteur est le plus proche de celui du texte.

Dans le schéma ci-dessus, les représentations des catégories n'ont pas de lien les unes avec les autres et ne possèdent pas de structure interne autre que celle d'un ensemble de mots. Ce n'est pas toutefois le cas pour tous les travaux portant sur ce problème. Tout en conservant les éléments de base du modèle vector-space, certains d'entre eux optent en faveur d'une structuration hiérarchique des vecteurs représentant les textes d'une catégorie, voire de plusieurs catégories si l'on s'intéresse aux liens entre celles-ci. La construction de la hiérarchie s'effectue le plus souvent suivant une procédure proche de la classification hiérarchique et aboutit ainsi à des arbres binaires. L'algorithme GAC (Groupe Average Clustering) [Cutting et alii 1992] est un exemple particulièrement efficace de ce type de méthode.

1.2.2. Les "topic signatures"

Dans le cadre de sa thèse [Lin 1997] portant sur les méthodes robustes d'identification des thèmes des textes, Lin a proposé une procédure automatique de construction de représentations de thèmes. Ces représentations sont appelées *topic signatures*, terme que nous traduirons par *signatures thématiques*. Le travail de Lin se situe dans la ligne directe des travaux sur la catégorisation de textes issus du domaine de la Recherche d'Information.

Le point de départ de la construction des signatures thématiques est un ensemble de thèmes définis chacun par une collection de textes reliés à ce thème. Soit ST_i, une signature thématique et $CT_i=\{Txt_{ij}\}$, la collection de textes dont elle est issue. À la fin du processus, chaque signature thématique ST_i se présente comme un ensemble de termes pondérés. Ces termes correspondent aux termes des textes de CT_i possédant les poids les plus forts. En vertu de l'application "classique" du facteur *tf.idf* (cf. définition donnée au §1.3 de l'annexe I), tout terme T_{ijk} d'un texte Txt_{ij} est pondéré au sein de ce texte par le résultat de la modulation de son nombre d'occurrences, t_{ijk}, par l'inverse de sa distribution parmi l'ensemble des textes de toutes les signatures. De manière similaire, le poids d'un terme au sein de la signature ST_i est donné par le nombre d'occurrences de ce terme dans cette signature (plus précisément la moyenne des t_{ijk}, pour j énumérant tous les textes de CT_i) modulé par l'inverse de la distribution de ce terme parmi l'ensemble des signatures. On ne conserve dans la représentation finale de chaque signature que ses 300 premiers termes de plus fort poids.

Avant la construction des signatures, les textes sont pré-traités au moyen d'un étiqueteur morpho-syntaxique afin de ramener les mots qui les composent à leur forme

canonique et obtenir ainsi les termes évoqués ci-dessus. WordNet est utilisé comme dictionnaire de référence, y compris pour les mots composés.

La procédure de construction des représentations de thème décrite précédemment forme le cœur de la méthode de Lin mais n'en constitue pas l'originalité puisqu'elle est identique aux principes adoptés généralement en Recherche d'Information pour réaliser ce type de tâche. La spécificité des propositions de Lin réside dans la prise en compte des cas d'ambiguïtés lors du processus d'identification thématique. Rien dans le processus de construction des signatures ne garantit en effet que celles-ci soient les plus disjointes possible au regard de la mesure de similarité utilisée pour décider de l'affectation d'un texte à un thème. Cette mesure est ici le cosinus des vecteurs à comparer (cf. §1.3 de l'annexe I).

Pour traiter ce problème, Lin définit des signatures thématiques à plusieurs niveaux. Celles-ci résultent de l'application de façon récursive de la procédure de construction des signatures décrite précédemment. Le processus est le suivant. Après que le premier niveau de signatures a été élaboré, on calcule la similarité de chaque signature avec toutes les autres. On détermine ainsi pour chacune d'elles, l'ensemble des signatures qui lui sont le plus proches, donc celles avec lesquelles une confusion est susceptible de se produire lors du processus d'identification thématique. Cet ensemble, appelé *ensemble de confusion* ("confusion set"), est donc spécifique de chaque signature et réciproquement, celle-ci est désignée comme la signature source de l'ensemble de confusion en question.

Le nombre de signatures de l'un de ces ensembles de confusion est bien entendu beaucoup moins important que celui de l'ensemble des signatures du niveau courant. Il est donc nécessaire de réévaluer les poids des termes au sein des signatures de ces ensembles. Ces poids dépendent en effet de la distribution des termes parmi les signatures considérées. On adapte donc la représentation des thèmes à ce nouveau contexte plus restreint, de la même façon que l'on augmente le grossissement d'un microscope pour mieux différencier certains détails. En procédant de cette façon, on se rapproche d'ailleurs d'une caractérisation des thèmes fondée sur leur différenciation, et par là même de la seconde approche exposée au §1.2.1.

Cette adaptation équivaut à relancer le processus de construction des signatures, mais cette fois-ci, au sein de chacun des ensembles de confusion non vides du niveau courant. Cette réévaluation des poids ayant été opérée, on peut réitérer le calcul de similarité entre la signature source de chacun des ensembles de confusion et les autres signatures de celui-ci, de façon à redessiner les contours de l'ensemble de confusion de cette signature source. À ce stade, on relance récursivement le processus décrit ci-dessus sur chaque nouvel ensemble de confusion jusqu'à ce que celui-ci devienne finalement vide. On

obtient ainsi pour chaque signature source une succession de niveaux de signatures tels que chaque niveau définit les critères permettant de faire la distinction, implicitement de façon de plus en plus fine, entre un nombre de signatures de plus en plus restreint.

Sur un plan plus technique, la détermination des signatures composant l'ensemble de confusion d'une signature source ST_i – elles sont appelées des *outliers* – s'effectue uniquement sur des bases statistiques à partir des valeurs de similarité calculées entre la signature source et l'ensemble des autres signatures du niveau considéré. Ces outliers se définissent plus précisément comme les signatures dont le niveau de similarité est supérieur au seuil θ_i, égal à la somme du troisième quartile de la distribution des valeurs de similarité et de 1,5 fois son écart interquartile. Le troisième quartile d'une distribution de valeurs correspond à la valeur telle que les trois quarts des valeurs lui sont inférieures. Suivant la même logique, le premier quartile est la valeur telle qu'un quart des valeurs lui sont inférieures. L'écart interquartile est donné par la différence entre le troisième et le premier quartile.

Compte tenu de la représentation des thèmes adoptée, la procédure d'identification thématique d'un nouveau texte se déroule de la façon suivante. Après la construction de la représentation vectorielle du texte, on calcule les valeurs de similarité entre ce texte et chacune des signatures de premier niveau. Pour chaque couple texte-signature (ST_i), on applique alors l'une de ces quatre règles :

- si max_i (valeur maximale de similarité de la signature ST_i avec les autres signatures du niveau) $\leq \theta_i$, i.e. l'ensemble de confusion de ST_i est vide, donc le thème considéré est bien séparé de ces voisins, et similarité texte-signature $> \theta_i$ alors assigner le texte au thème représenté par la signature ST_i;
- si $max_i \leq \theta_i$ et similarité texte-signature $\leq \theta_i$ alors le texte n'est assigné pas au thème considéré;
- si $max_i > \theta_i$, i.e. il peut y avoir confusion entre ST_i et une autre signature, et similarité texte-signature $> max_i$ alors assigner le texte au thème représenté par ST_i;
- si $max_i > \theta_i$ et similarité texte-signature $\leq max_i$ alors on se trouve dans le cas où l'on fait intervenir les niveaux supérieurs de signature afin de prendre la décision. On applique alors la même série de tests mais avec les nouvelles valeurs de max_i et de θ_i.

La méthode d'identification thématique proposée par Lin, et donc par voie de conséquence indirecte, celle de construction des représentations de thèmes, ont été évaluées sur un ensemble de 32 thèmes. Les signatures thématiques correspondant à ces thèmes ont été élaborées à partir d'une collection de 16137 articles du *Wall Street Journal*.

Le jeu de test des textes à classer était quant à lui constitué de 12906 textes de ce même journal, couvrant 31 des 32 thèmes représentés. Les performances ont été évaluées par les mesures de rappel et de précision appliquées relativement à chacun des thèmes. Le rappel pour un thème correspond au rapport entre le nombre de textes correctement rattachés à ce thème et le nombre de textes qui auraient dus y être rattachés. La précision est donnée quant à elle par le rapport entre le nombre de textes correctement rattachés au thème et le nombre de textes qui y sont effectivement rattachés. On obtient ainsi une valeur moyenne de 0,848 pour le rappel et de 0,764 pour la précision sur le corpus de constitution des signatures ainsi que des valeurs respectives de 0,802 et 0,729 sur le corpus de test.

1.2.3. Les travaux réalisés dans le cadre de TDT

Toutes les méthodes de construction de représentations de thèmes évoquées jusqu'à présent s'appuient sur l'hypothèse suivante : au stade initial, tous les textes intervenant dans la définition des thèmes sont connus. Cette hypothèse de travail se manifeste tantôt au travers même de la méthode utilisée, c'est le cas en particulier des algorithmes de type classification hiérarchique, qui procèdent à des comparaisons entre tous les éléments à classifier, tantôt au travers de points moins centraux, comme la pondération des mots des représentations de texte. L'utilisation dans ce dernier cas du facteur *tf.idf*, lequel nécessite de connaître la répartition des mots dans tous les textes, est un exemple de la dépendance vis-à-vis de cette hypothèse. Une seconde hypothèse, un peu moins systématique que la première, stipule que le rattachement de chaque texte à un thème est lui aussi fixé initialement. Or, ces deux hypothèses vont à l'encontre des principes fondateurs d'ANTHAPSI, en vertu desquels l'apprentissage s'y fait de façon incrémentale et non supervisée. Ce principe s'applique de fait à la construction des représentations de thème dans le cadre de SEGAPSITH. Les méthodes précédentes ne sont donc pas directement utilisables dans notre contexte.

En revanche, nous avons vu au §1.2 du chapitre 8 que la tâche de Détection de l'évaluation TDT (Topic Detection and Tracking) [Yang et alii 1997] impose des contraintes rendant les solutions adoptées pour la mener à bien assez proches de notre objet d'intérêt. Cette tâche comporte plus précisément deux parties : la détection rétrospective et la détection on-line. La première consiste à partitionner de façon non supervisée un ensemble de dépêches, initialement connues, afin de les regrouper en fonction de l'événement qui en est le sujet. Le fait de disposer au début du processus de l'intégralité des textes à classifier renvoie à certains travaux évoqués précédemment et nous conduit à ne pas approfondir davantage cette partie. Par ailleurs, les trois participants à l'étude pilote de TDT dont nous évoquons les travaux ici (l'université du Massachussetts (UMass), l'université Carnegie Mellon (CMU) et la société Dragon

Systems) ont globalement utilisé dans cette première partie les mêmes techniques que dans la seconde[1].

La détection on-line consiste pour sa part à détecter la présence d'un nouvel événement dans un flux de nouvelles arrivant en continu. La nouveauté d'un événement n'est pas définie par rapport à une référence extérieure mais par rapport aux événements déjà rencontrés dans le flux de dépêches. C'est pourquoi la technique la plus utilisée pour réaliser cette tâche a consisté à classifier les dépêches à la volée. Une nouvel événement s'identifie alors à une dépêche que l'on ne peut classer dans aucune des catégories construites jusqu'à présent. Cette technique suppose qu'une représentation de chaque événement soit construite à partir des dépêches qui y font référence, d'où le lien avec nos préoccupations. Seul UMass a opté en faveur d'une procédure ne nécessitant pas de créer une telle représentation : on conserve pour chaque dépêche rencontrée une sorte de signature (plus exactement un groupe de mots supposés importants qui s'identifient à une requête susceptible de retourner la dépêche en question); la nouveauté d'une dépêche est déterminée par confrontation directe, via des techniques de Recherche d'Information, entre le texte de la dépêche et l'ensemble des signatures déjà accumulées.

Dans leur parti pris de construire une représentation des événements rencontrés, CMU et Dragon se sont orientés vers des méthodes proches de celle adoptée dans MLK et donc, de celle pressentie pour SEGAPSITH. Le principe général en est le suivant. La représentation de chaque nouvelle dépêche est confrontée, par l'entremise d'une mesure de similarité, aux représentations des événements déjà rencontrés. Si la valeur de cette mesure dépasse un seuil fixé a priori, la dépêche est rattachée à l'événement pour lequel ce dépassement intervient et la représentation de ce dernier est mise à jour. Si au contraire, le seuil de similarité n'est franchi pour aucun des événements déjà rencontrés, on considère que la dépêche fait référence à un nouvel événement et celle-ci sert de point de départ à la construction d'une représentation de ce dernier.

L'application de ce principe s'est accompagnée aussi bien par CMU que par Dragon de la prise en compte des spécificités de la tâche dans TDT afin d'introduire certaines optimisations qu'il ne sera pas possible de reproduire dans SEGAPSITH. Ainsi, on sait que les dépêches sont classées dans le flot selon leur ordre chronologique et que les événements qu'elles relatent n'y sont représentés que pendant une certaine durée. On peut ainsi limiter le parcours des représentations d'événements présents en mémoire en fonction de ce critère. En supposant que le corpus total couvre un an, on pourra par exemple se limiter aux événements du mois courant ou du mois précédent.

[1] Seul CMU a utilisé dans l'une de ses expérimentations l'algorithme GAC, qui suppose de disposer de l'intégralité du corpus au départ.

Bien entendu, quelques différences existent entre les solutions de CMU et de Dragon, concernant notamment la façon dont sont représentées les dépêches ou la forme de la mesure de similarité entre la représentation d'une dépêche et celle d'un événement. CMU pondère ainsi les mots des dépêches par un facteur de type *tf.idf* alors que Dragon se contente d'adopter comme poids la probabilité d'apparition des mots dans un corpus de référence. Il est à noter que la pondération de type *td.idf* prend ici comme référence, pour juger de la répartition des mots parmi les dépêches, celles qui ont déjà été rencontrées, et non l'ensemble du corpus. La mesure de similarité de CMU est la "classique" mesure cosinus tandis que Dragon utilise pour sa part une distance de type Kullback-Leibler[1], avec un lissage du poids des mots dans les représentations des événements par la probabilité d'apparition de ces mots dans un corpus de référence.

Comme les deux autres tâches de TDT, la tâche Détection a fait l'objet d'une évaluation. Celle-ci a été réalisée à partir du corpus utilisé pour l'ensemble des tâches de TDT. Ce corpus est composé d'environ 16000 textes, une partie d'entre eux étant des dépêches de l'agence *Reuters* et l'autre partie provenant de la transcription de sujets télévisés de *CNN*. Un peu moins de 10% de ces textes ont été étiquetés en fonction de leur référence à l'un des 25 événements retenus pour les tests.

La détection rétrospective a été appliquée sur l'ensemble du corpus mais l'évaluation proprement dite ne porte que sur les 25 événements test. Pour ce faire, les textes étiquetés ont été utilisés afin de sélectionner les 25 représentations d'événement, parmi celles formées à partir de la totalité du corpus, qui sont les plus proches des 25 événements considérés. Un tableau de contingence répertoriant les différents cas possibles[2] (cf. exemple d'un tel tableau à la figure 9.12 du chapitre 9) a été établi pour chaque représentation d'événement et un tableau global réalisant la moyenne de ces différents tableaux a été calculé. C'est à partir de celui-ci qu'ont été évaluées les valeurs de précision, de rappel, de fausse alarme, d'erreur ainsi que la F-mesure. Rappelons que la F-mesure permet de combiner la précision et le rappel en un seul indicateur selon la formule suivante :

$$F_\beta(r, p) = \frac{(\beta^2 + 1)\ p\ r}{\beta^2\ p + r}$$

---

[1] $d = \sum_n s_n/S \ \log \frac{s_n/S}{c_n/C}$, avec $S = \sum_n s_n$, $C = \sum_n c_n$ et

s_n: poids du mot n dans la dépêche; c_n: poids du mot n dans la représentation de l'événement.

[2] Quatre cas sont possibles pour un événement E : (1) texte relatif à E et classé dans la représentation de E; (2) texte non relatif à E et non classé dans la représentation de E; (3) texte relatif à E et non classé dans la représentation de E (erreur de classement); (4) texte non relatif à E et classé dans la représentation de E (fausse alarme).

p étant la précision et r, le rappel. Dans le cas présent, le cœfficient β, qui permet de moduler les influences respectives de la précision et du rappel, est égal à 1. On mesure donc $F_1(r,p)$. À titre indicatif, le meilleur système obtient une précision de 82%, un rappel de 62% et une F-mesure de 0,71. Le moins bon obtient quant à lui une précision de 16%, un rappel de 33% et une F-mesure de 0,21.

Parmi les points particulièrement étudiés lors de cette évaluation, figure la dépendance entre le taux d'erreurs de classement et le taux de fausses alarmes. En modifiant le paramétrage des systèmes, on peut en effet favoriser l'un au détriment de l'autre ou vice versa. En systématisant ces modifications, il est possible d'obtenir un ensemble de points offrant une vue globale de ce phénomène d'échange, appelé DET (Decision Error Trade-off), vue que l'on peut ensuite matérialiser en reportant ces points dans un repère ayant le taux d'erreurs en ordonnée et le taux de fausses alarmes en abscisse. Ce DET a également été étudié dans le cadre de la tâche de détection on-line.

Pour sa part, celle-ci a été évaluée uniquement sur les mille textes environ du corpus TDT ayant été étiquetés avec un seul des 25 événements test. Puisque la tâche consiste à détecter un nouvel événement, les exemples positifs se trouvent limités aux seuls 25 textes ayant chacun marqué la détection d'un des 25 événements test. Afin d'augmenter le nombre de ces exemples positifs et assurer ainsi une plus grande représentativité de l'évaluation, une procédure en plusieurs passes a été conçue. À la suite de chaque passe, on élimine de l'ensemble de test le premier texte ayant donné lieu à la formation de chacune des représentations d'événement et on relance le processus de détection. L'évaluation a été réalisée en l'occurrence sur le résultat de 10 passes.

Le principe de calcul des différents indicateurs reste le même que dans le cas précédent puisque chaque texte est toujours caractérisé par un jugement binaire. Seul la signification de celui-ci change. Au lieu de déterminer si le texte est relatif ou non à l'événement considéré, ce jugement établit si le texte marque ou non l'arrivée d'un nouvel événement. Le meilleur système obtient une précision de 45%, un rappel de 50% et une F-mesure de 0,48. Le moins bon obtient quant à lui une précision de 28%, un rappel de 27% et une F-mesure de 0,28. Ces résultats montrent, comme on pouvait s'y attendre, que cette tâche est plus difficile que la précédente.

1.3. Construction de signatures thématiques par agrégation de segments de textes

1.3.1. Principes

Le principe général de la construction des signatures thématiques dans SEGAPSITH reprend comme nous l'avons indiqué en préambule le principe de la mémorisation des représentations de texte de MLK. Il est également très proche d'un certain nombre de solutions adoptées pour résoudre la tâche Détection on-line de l'action d'évaluation TDT (cf. §1.2.3). Ce principe est formalisé par l'algorithme détaillé ci-dessous. Celui-ci est une instanciation, dans le cadre de SEGAPSITH, de l'algorithme générique présenté au paragraphe 4.1 du chapitre 6. Il prend en entrée l'ensemble des UTLs d'une représentation de texte et produit un renforcement de certaines signatures et/ou de nouvelles signatures. Précisons que les UTLs sont traitées ici indépendamment les unes des autres.

```
Pour UTL énumérant toutes les UTLs de la représentation de texte considérée faire
    liste_signatures ← sélection(mémoire,UTL)
    Répéter
        signature ← élémentSuivant(liste_signatures)
        sim ← similarité(UTL,signature)
    Jusqua (sim = vrai) ou estVide(liste_signatures)
    Si (sim = vrai) alors
        agrégation(UTL,signature)
    Sinon
        créationSignature(UTL)
    Fin_si
Fin_pour
```

1.3.2. Représentation des textes et des signatures thématiques

Avant de détailler les différentes opérations caractérisant la mémorisation d'une représentation de texte, nous allons examiner de plus près la forme que revêtent à la fois les entrées du processus de mémorisation, c'est-à-dire les représentations de texte, et ses sorties, autrement dit les signatures thématiques.

Les représentations de texte et les UTLs

Une représentation de texte dans SEGAPSITH est beaucoup plus simple que dans MLK puisqu'elle n'est composée que d'une liste d'Unités Thématiques Lexicales (UTLs). Chacune de ces UTLs est construite à partir d'un segment mis en évidence par la segmentation thématique. Néanmoins, tous les segments d'un texte ne donnent pas lieu à

la construction d'une UTL. Les différents thèmes d'un texte sont parfois trop enchevêtrés les uns dans les autres au niveau de certaines de ses parties pour qu'un segment thématiquement homogène puisse être défini à ces endroits. Dans un contexte comme celui de SEGAPSITH ou de SEGCOHLEX dans lequel beaucoup de textes sont traités, y compris sur un même thème, on peut donc ne retenir que les segments de texte les plus thématiquement cohérents pour former des UTLs et éviter ainsi d'introduire du bruit dans les signatures thématiques ou de créer des signatures trop spécifiques.

L'évaluation de la cohérence thématique d'un segment est donnée par un niveau de cohésion associé à chaque segment à l'issue de la segmentation. Nous avons vu que dans le cas de SEGCOHLEX, ce niveau correspond au maximum, après lissage, des valeurs de cohésion calculées lors de la segmentation pour les différentes positions du texte situées à l'intérieur du segment. On se reportera au paragraphe 2 de ce chapitre pour une description du mécanisme équivalent dans SEGAPSITH. La sélection des segments les plus cohérents s'effectue finalement par comparaison de cette valeur de cohésion vis-à-vis d'un seuil adaptatif : on ne retient que les segments dont le niveau de cohésion dépasse le seuil déterminé par la somme de la moyenne de ces valeurs de cohésion et de leur écart-type.

Les UTLs sont elles-mêmes très simples comparativement aux Unités Thématiques de MLK. Elles ne sont formées en effet que de mots auxquels est associé un poids. Leur format est le même, qu'elles soient le produit de la segmentation thématique de SEGAPSITH ou celui de la segmentation de SEGCOHLEX. Une UTL regroupe deux catégories de mots, maintenus séparés en deux ensembles distincts. La catégorie la plus évidente est bien entendu celle des mots composant le segment à partir duquel l'UTL considérée a été construite. Les mots relevant de cette catégorie seront dénommés par la suite *mots des textes*.

La seconde catégorie rassemble pour sa part un ensemble de mots qui ne sont pas présents explicitement dans le segment mais qui sont suggérés par les mots de ce segment. C'est pourquoi ils seront appelés par la suite *mots inférés*. Ces mots inférés sont plus précisément des mots avec lesquels les mots du segment entretiennent des liens forts au sein de la source de connaissances utilisée pour la segmentation, en l'occurrence le réseau de collocations pour SEGCOHLEX et les signatures thématiques pour SEGAPSITH. En ce qui concerne SEGCOHLEX, nous avons vu au chapitre précédent que pour chaque position du texte, un ensemble de mots du réseau de collocations liés aux mots de la fenêtre de calcul de la cohésion sont sélectionnés. Les mots inférés d'un segment s'identifient alors aux mots qui sont sélectionnés le plus fréquemment au sein de ce segment. Plus précisément, on ne retient comme mot inféré que les mots du réseau de

collocations sélectionnés pour au moins 75% des positions du texte située à l'intérieur du segment considéré. On se reportera au paragraphe 2 de ce chapitre pour connaître la façon dont les mots inférés sont déterminés dans le cas de SEGAPSITH.

La présence des mots inférés dans les UTLs est un moyen de dépasser les simples similarités de surface, ce qui est particulièrement nécessaire dans le contexte de ROSA où les briques de base des représentations, i.e. les mots, sont à la fois faiblement structurées et assez ambiguës. Plus précisément, les mots inférés permettent de limiter l'influence des deux phénomènes suivants : d'une part, la variabilité d'expression d'un même thème au niveau lexical; d'autre part, le fait qu'un segment ne contient pas nécessairement beaucoup de mots spécifiques du thème auquel il fait référence. Même s'ils sont loin de représenter tout le vocabulaire spécifique d'un domaine, les mots inférés associés à un segment sont généralement assez nombreux pour élargir la définition du thème au point de lever une éventuelle ambiguïté. Ils conduisent donc selon les cas à renforcer une similarité faiblement établie entre deux segments faisant référence au même thème mais ne partageant pas beaucoup de mots ou au contraire à infirmer une similarité entre deux segments n'ayant pas le même thème mais un certain nombre de mots en commun. Ces mots inférés interviennent avec la même action lors de la sélection des signatures thématiques les plus proches d'une UTL.

Qu'ils viennent des textes ou bien qu'ils soient inférés, les mots des UTLs se voient associer un poids. Celui-ci a pour objectif de rendre compte de l'importance supposée de ce mot au sein de l'UTL. Le poids d'un mot m_i dans UTL est donné par la formule :

$$poids(m_i, UTL) = \sqrt{nbOcc(m_i, UTL)} \;\; signif(m_i) \qquad [1]$$

avec $nbOcc(m_i,UTL)$: nombre d'occurrences du mot m_i dans le segment,

$signif(m_i)$: significativité du mot m_i par rapport au corpus utilisé pour construire le réseau de collocations (selon la définition donnée au §1.4.1 du chapitre 9).

Le premier terme traduit l'importance du mot m_i dans le contexte du segment considéré, suivant l'hypothèse qu'un mot est d'autant plus important qu'il est fréquemment répété. L'atténuation produite par la racine carrée évite que ce terme ne prenne trop d'importance au niveau de la mesure de similarité. Le second terme reflète quant à lui le degré de spécificité du mot m_i, supposé évalué de façon générale. En réalité, cette mesure est relative au corpus du journal *Le Monde* utilisé pour construire le réseau de collocations. Même si ce corpus est assez varié, il ne s'agit donc que d'une estimation d'un degré de spécificité général. Le poids d'un mot se veut ainsi la combinaison de son importance en contexte et hors contexte. Il est à noter que dans le cas d'un mot inféré le premier terme est toujours égal à 1 puisque l'on ne peut pas parler de nombre de

répétitions. Dans le cas de SEGCOHLEX, il serait néanmoins possible de calculer un poids ayant une signification comparable à ce qui prévaut pour les mots des textes. La fréquence de répétition correspondrait alors à la fréquence d'apparition en tant que mot sélectionné parmi les différentes positions du segment, au delà bien sûr du seuil minimum de 75%.

La représentation des signatures thématiques

Les signatures thématiques sont le résultat de l'agrégation de plusieurs UTLs. L'élément de base des UTLs, i.e. le mot, ne subissant aucune transformation lors de l'opération d'agrégation, au contraire des concepts par exemple, la structure des signatures thématiques est de ce fait identique à celle des UTLs. Une signature thématique est donc composée de deux listes de mots pondérés, l'une correspondant au cumul des mots des textes venant des UTLs ayant été agrégées pour former cette signature et l'autre, au cumul des mots inférés de ces mêmes UTLs. La différence essentielle entre les UTLs et les signatures thématiques concerne la pondération adoptée pour leurs mots. Le poids du mot m_i appartenant à la signature ST est donné par la formule :

$$poids(m_i, ST) = \frac{nbOcc(m_i, ST)}{nbAgr(ST)} \; signif(m_i) \; \frac{nbAgr^4(ST)}{(nbAgr(ST)+1)^4} \qquad [2]$$

avec $nbOcc(m_i,ST)$: nombre d'occurrences du mot m_i dans la signature thématique ST,

$nbAgr(ST)$: nombre d'agrégations ayant conduit à la formation de la signature thématique ST.

En dépit de sa différence avec [1], on retrouve dans cette formule les trois composantes caractérisant [1] :

- une composante traduisant l'importance du mot par rapport à l'entité considérée, en l'occurrence la signature thématique. Elle est incarnée ici par le premier terme;
- une composante rendant compte de l'importance du mot en général. Il s'agit ici du deuxième terme. Cette composante est identique dans [1] et dans [2];
- un facteur d'atténuation parant à certains effets indésirables de la première composante vis-à-vis de la mesure de similarité et du processus de sélection des signatures. Dans [1], ce facteur est lié directement à la première composante et évite que le poids d'un mot fréquemment répété soit disproportionné, au regard du critère d'importance, par rapport au poids des autres mots. Dans [2], il prend la forme d'un terme à part entière (le troisième plus précisément) et contrebalance le poids exagéré des mots lors des premières agrégations. Le but est d'empêcher que les signatures nouvellement créées n'agissent comme des attracteurs.

Outre la façon de calculer le poids des mots, UTLs et signatures se distinguent également par le fait que les signatures, à l'instar des structures de MLK, conservent les moyens de retrouver les UTLs à partir desquelles elles ont été formées. Chaque mot présent dans une signature se voit ainsi associé l'ensemble des identifiants des UTLs dans lesquelles il était présent. Cette information n'est pas utilisée pour le moment mais pourrait être intéressante si l'on souhaite raisonner a posteriori sur les différences inter-individuelles entre les UTLs regroupées dans une même signature. C'est le cas en particulier lorsque l'on souhaite scinder une signature de trop grosse taille en un ensemble de signatures plus petites et surtout plus homogènes.

1.3.3. Sélection des signatures thématiques

En vertu de l'algorithme défini au §1.3.1, la première étape de la mémorisation d'une nouvelle UTL consiste à rechercher en mémoire les signatures thématiques susceptibles de s'agréger avec elle. Cette mémoire pouvant potentiellement contenir un grand nombre de signatures, il est d'abord nécessaire de sélectionner de manière efficace un sous-ensemble d'entre elles pour lesquelles une similarité plus profonde pourra être ensuite évaluée. Pour ce faire, nous réalisons l'équivalent d'un pas de propagation d'activité à partir des mots composant l'UTL considérée. L'activité d'une signature ST_i ayant au moins un mot en commun avec la nouvelle UTL est donnée par la fonction suivante :

$$activité\,(ST_i) = \sum_j poids(m_j, ST_i)\ \ poids(m_j, UTL)$$

où $poids(m_j,ST_i)$ renvoie à [1] et $poids(m_j,UTL)$ renvoie à [2].

On voit que chaque mot commun à l'UTL et à la signature apporte à l'activité de cette signature une contribution égale à la combinaison, par un produit dans le cas présent, de l'importance de ce mot au sein de la signature et de son importance au sein de l'UTL.

Il faut préciser que l'activation des signatures s'effectue aussi bien à partir des mots des textes que des mots inférés. Cependant, du fait de leur statut de mots incertains puisque non explicitement présents dans les textes, les mots inférés voient leur contribution à l'activation des signatures volontairement réduite de moitié comparée à celle des mots venant des textes. Par ailleurs, les mots dont le poids est trop faible (en l'occurrence inférieur à 0,1[1]), que ce soit au sein de l'UTL ou au sein d'une signature, ne sont pas pris en compte pour l'activation des signatures.

[1] En ce qui concerne les signatures, ce seuil a été empiriquement fixé sur la base de l'examen manuel de l'intérêt des mots présents dans les signatures vis-à-vis du thème de ces signatures. Nous avons ainsi pu déterminer qu'en moyenne les mots dont le poids est inférieur à 0,1 ne sont plus significatifs au regard du thème de la signature dont ils font partie.

La sélection des signatures thématiques les plus activées, donc à examiner de plus près quant à leur similarité avec la nouvelle UTL, s'effectue quant à elle par comparaison à un seuil fonction de la distribution de toutes les activités. Comme pour la sélection des segments de texte voués à devenir des UTLs, on retient les signatures possédant une activité supérieure à la somme de la moyenne de ces activités et de leur écart-type.

1.3.4. Similarité et agrégation

Le processus de sélection décrit au paragraphe précédent peut être vu comme une première mesure de similarité, peu élaborée mais applicable largement du fait de son coût raisonnablement faible. Après que le champ des possibilités a été restreint par cette première sélection, il devient possible d'appliquer une mesure de similarité plus complexe afin de déterminer si la nouvelle UTL s'agrège à l'une des signatures thématiques sélectionnées ou, lorsque la similarité est en dessous d'un seuil fixé a priori, si elle constitue le point de départ d'une nouvelle signature.

Similarité

La mesure de similarité appliquée ici se fonde uniquement sur les mots communs entre une signature et une UTL. La méthode d'apprentissage utilisée est en effet source d'un bruit important et de ce fait, les différences entre les deux entités ne sont globalement pas significatives de leur similitude ou de leur dissimilitude. Le phénomène est assez évident en ce qui concerne les mots des textes dans la mesure où le vocabulaire d'un segment est loin de ne contenir que des mots spécifiques du thème qu'il évoque. Mais c'est également le cas pour ce qui est des mots inférés, en particulier lorsque la représentation des thèmes abordés est encore naissante et que la segmentation des textes repose alorsessentiellement sur SEGCOHLEX. En effet, les relations présentes dans le réseau de cooccurrences lexicales ne sont pas seulement thématiques comme nous avons le voir au chapitre 9. Le type des cooccurrences restant implicite, des mots inférés sont donc retenus dans les UTLs sur la base d'autres critères que la proximité thématique et représentent aussi une source de bruit du point de vue de notre tâche.

Sur le principe, la mesure de similarité entre une UTL et une signature combine l'importance que revêtent du point de vue de chacune de ces deux entités les mots qu'elles ont en commun par rapport à l'ensemble des mots qui les forment respectivement. Ce principe est matérialisé au travers de deux ratios, l'un attaché à la signature thématique, $ratio_{ST}$, et l'autre à l'UTL, $ratio_{UTL}$. Chacun d'entre eux décline le rapport entre mots communs aux deux entités et mots propres à l'entité considérée à la fois selon le poids de ces mots et selon leur nombre d'occurrences. On cherche de cette manière à éviter d'avoir

une forte similarité entre une UTL et une signature ne partageant qu'un petit nombre de mots communs de fort poids de part et d'autre. La combinaison du facteur poids et du facteur nombre d'occurrences s'effectue grâce à une moyenne géométrique. Celle-ci présente l'intérêt de défavoriser davantage les forts écarts de valeur que la moyenne arithmétique. Par ailleurs, utiliser un simple produit aurait conduit à manipuler des valeurs généralement très petites au sein d'une échelle de valeurs assez large, ce qui ne facilite pas la fixation d'un seuil. Les ratios associés à l'UTL et à la signature, $ratio_{ST}$ et $ratio_{UTL}$, sont également combinés grâce à une moyenne géométrique. Plus formellement, on a les définitions suivantes :

$$ratio_{ST} = \sqrt{\frac{\sum_c poids(m_c, ST)}{\sum_t poids(m_t, ST)} \frac{\sum_c nbOcc(m_c, ST)}{\sum_t nbOcc(m_t, ST)}}$$

$$ratio_{UTL} = \sqrt{\frac{\sum_c poids(m_c, UTL)}{\sum_t poids(m_t, UTL)} \frac{\sum_c nbOcc(m_c, UTL)}{\sum_t nbOcc(m_t, UTL)}}$$

$$similarité(UTL, ST) = \sqrt{ratio_{UTL} \; ratio_{ST}}$$

où l'indice *c* énumère tous les mots communs à l'UTL et à la signature thématique *ST* et l'indice *t* énumère l'ensemble des mots composant respectivement l'UTL et la signature thématique *ST*.

Nous appliquons pour la mesure de similarité les mêmes principes concernant les mots inférés et les mots de faible poids que pour l'opération de sélection des signatures : les valeurs (poids et nombre d'occurrences) associées aux mots inférés sont divisés par 2 et les mots de faible poids (seuil également fixé à 0,1) ne sont pas pris en compte. Les expérimentations réalisées ont permis par ailleurs de fixer le seuil de création d'une nouvelle signature thématique à 0,25.

Agrégation

L'opération d'agrégation d'une UTL et d'une signature est quant à elle très simple du fait même de la simplicité de la structure de ces deux entités. Elle consiste pour l'essentiel à fusionner deux listes de mots pondérés. Le poids d'un mot dans une signature étant calculé dynamiquement à partir de son nombre d'occurrences grâce à la formule [2], l'agrégation peut être assimilée à une opération additive : si un mot m_i de l'UTL n'est pas présent dans la signature, il y est ajouté avec le nombre d'occurrences qu'il possède au sein de l'UTL, en l'occurrence $nbOcc(m_i,UTL)$; s'il y figure déjà, son nombre d'occurrences est uniquement augmenté de $nbOcc(m_i,UTL)$. Encore une fois, ce mécanisme est identique pour les mots des textes et les mots inférés et s'applique

indépendamment sur la liste des premiers et celle des seconds. En conformité avec les principes sous-tendant ANTHAPSI, il permet de faire émerger progressivement les mots les plus caractéristiques d'un thème en faisant l'hypothèse qu'ils correspondent aux mots apparaissant de façon récurrente au sein des segments de texte relatifs à un même thème.

1.3.5. Expérimentation sur un large corpus

Afin d'expérimenter la méthode présentée et la valider, nous avons appliqué sur un large ensemble de textes la chaîne d'opérations allant du pré-traitement des textes à la construction des signatures en passant par la segmentation thématique. Cet ensemble de textes se compose d'un mois de dépêches AFP (mois de mai 1994), soit 5949 textes d'une longueur moyenne de 250 mots. Le choix de ce corpus se justifie en premier lieu par le type des textes qu'il contient et ensuite par le mélange équilibré de similitude et de diversité des thèmes qui y sont traités. Pour que le test puisse être probant, il est en effet nécessaire que les textes abordent différents thèmes, mais en même temps que ceux-ci ne soient pas trop dispersés si l'on ne veut pas obtenir autant de signatures que de segments de texte. Les dépêches d'un même mois possèdent cet équilibre : les événements relatés sont assez divers mais conjointement, un même événement ou une configuration d'événements est assez fréquemment le sujet de plusieurs dépêches.

mots des textes			mots inférés		
mot	nbOcc	poids	mot	nbOcc	poids
attentat	52	0,435	voiture_piégée	48	0,551
bombe	26	0,244	attentat_à_la_bombe	38	0,441
police	30	0,226	forces_de_sécurité	39	0,416
explosion	25	0,222	grenade	38	0,407
revendiquer	24	0,209	couvre-feu	34	0,364
tuer	26	0,197	terroriste	37	0,339
blessé	21	0,180	commando	35	0,336
exploser	18	0,176	dégât_matériel	29	0,336
ulster	14	0,152	autobus	32	0,332
blesser	14	0,124	fusillade	31	0,328
source	14	0,113	blesser	37	0,319
lundi	14	0,108	endommager	30	0,314
mort	14	0,095	extrémiste	32	0,303
irlandais	10	0,095	blessé	35	0,303
personne	15	0,094	sentier_lumineux	27	0,300
matin	12	0,094	séparatiste	30	0,299
attribuer	11	0,094	explosif	31	0,293
séparatiste	9	0,090	lima	28	0,292
dimanche	12	0,090	armée_républicaine	22	0,253
policier	11	0,088	embuscade	23	0,252

Fig. 10.1 - Mots les plus représentatifs d'une signature thématique relative au terrorisme et résultant de 61 agrégations[7]

Pour cette expérimentation, nous avons utilisé la segmentation de SEGCOHLEX en nous plaçant ainsi dans l'optique de l'apprentissage des premières signatures d'un domaine, cas qui nous intéresse particulièrement dans la problématique d'ANTHAPSI. En l'absence d'un mécanisme d'adaptation aux caractéristiques de chaque texte traité, les valeurs adoptées pour les paramètres de la segmentation ont été choisies en fonction de la taille moyenne des textes : fenêtre de calcul de la cohésion de 19 mots, fenêtre de lissage de 9 mots et nombre de mots du texte nécessaires pour la sélection d'un mot du réseau de collocations égal à 2[2]. Dans ces conditions, la segmentation a produit 8601 UTLs à partir des 5949 textes initiaux.

La mémorisation des UTLs a donné lieu quant à elle à la construction de 3240 signatures thématiques. 691 (≈ 21%) d'entre elles sont le résultat d'au moins 2 agrégations. On peut donc considérer qu'au moins 2549 (≈ 79%) de ces signatures, c'est-à-dire une très nette majorité d'entre elles, ne sont pas significatives puisque résultant d'une UTL n'ayant jamais été trouvée similaire à une autre UTL. Il faut néanmoins préciser que ce jugement est limité au corpus considéré. De nouveaux textes pourraient tout à fait donner lieu à des UTLs s'agrégeant avec ces signatures. Ces signatures non significatives représentent une forte proportion de l'ensemble des signatures mais elles ne regroupent qu'une minorité des UTLs : environ 70% d'entre elles, 6052 pour être exact, sont en effet mémorisées dans des signatures issues d'au moins de 2 agrégations.

La répartition des signatures en fonction de leur nombre d'agrégations suit le type de distribution observé pour les collocations (cf. figure 9.4) : beaucoup de signatures rassemblant un petit nombre d'occurrences (dans le cas présent, les occurrences d'une signature correspondent aux UTLs qu'elle regroupe) et une diminution très brutale (décroissance de type exponentielle) de leur nombre à mesure de l'augmentation du nombre d'occurrences regroupées. La signature la plus fortement agrégée rassemble 413 UTLs mais on ne compte que 13 signatures au delà de 100 agrégations. La figure 10.1 donne à titre d'exemple la partie la plus représentative du contenu d'une des signatures construites dans le cadre de cette expérimentation. Cette signature réunit 61 UTLs faisant toutes référence au domaine du terrorisme. Cette unité thématique se retrouve au travers de la plupart de ses mots de plus fort poids. Les mots "attentat",

---

[1] On ne s'étonnera pas de trouver un mot comme "blesser" à la fois dans les mots des textes et dans les mots inférés. Il se peut en effet qu'il apparaisse explicitement dans certains textes et pas dans d'autres mais que dans ces derniers, il soit tout de même sélectionné à partir du réseau de collocations.

[2] Utiliser un seuil de 3 mots donnerait sans doute de meilleurs résultats du point de vue de la segmentation mais il n'est pas sûr que cela soit bénéfique pour la construction des signatures dans la mesure où cela limiterait également le nombre de mots inférés. Il faudrait en fait évaluer quelle proportion de ces mot perdus sont liés au thème représenté.

“voiture_piégée”, “bombe”, “explosion” ou “forces_de_sécurité” en sont à cet égard tout à fait représentatifs.

À côté de ces termes assez génériques, on trouve aussi trace des événements spécifiques à partir desquels ces signatures ont été construites. La présence des mots “sentier_lumineux”, “lima”, “autobus” (référence aux attentats anti-israéliens perpétués dans des bus) ou “ulster” en est une illustration assez nette pour la signature de la figure 10.1[1]. Si les termes généraux et représentatifs d’un domaine réussissent à émerger, on peut néanmoins penser que l’influence des mots attachés à des événements spécifiques devrait décroître à mesure qu’un ensemble de textes à la fois plus grand et plus diversifié aura été traité à propos du même sujet.

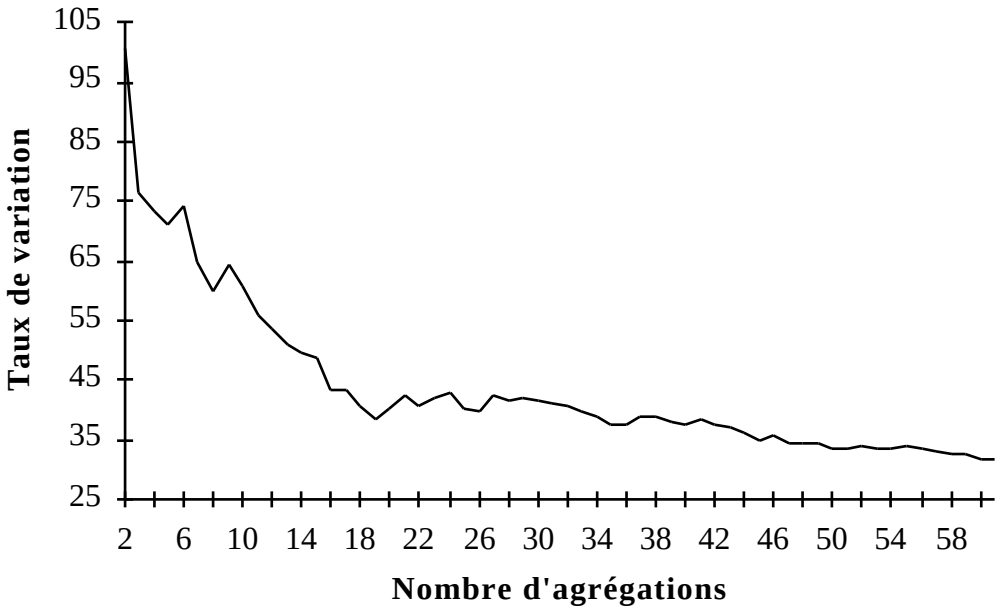


Fig. 10.2 - Taux de variation de la tête de la signature thématique de la figure 10.1

Globalement, on constate que les signatures formées apparaissent cohérentes. En particulier, leurs mots de plus fort poids présentent une homogénéité thématique souvent bien marquée. Il ne s’agit là bien entendu que d’une évaluation subjective, qui demanderait à être confirmée en recoupant le jugement de plusieurs sujets, ainsi que cela a été réalisé par Litman et Passonneau pour la segmentation thématique. Néanmoins, les modalités exactes d’une telle expérimentation restent à définir puisqu’à notre connaissance, aucune évaluation de ce type n’a encore été menée. Le caractère

[1] On a la confirmation à cette occasion que le TreeTagger étiquette beaucoup de noms propres comme des mots communs, ce qui explique que l’on retrouve en final un nombre non négligeable de noms propres dans les signatures.

non-supervisé de l'apprentissage interdit en particulier l'utilisation d'une référence constituée a priori comme cela se pratique généralement. Il semble donc que seul le jugement des sujets sur les signatures construites puisse servir de base à l'évaluation. Les psychologues cogniticiens, par leur expérience concernant des tâches proches (constitution de listes d'associations par exemple), pourraient sans doute fournir quelques pistes sur ce point.

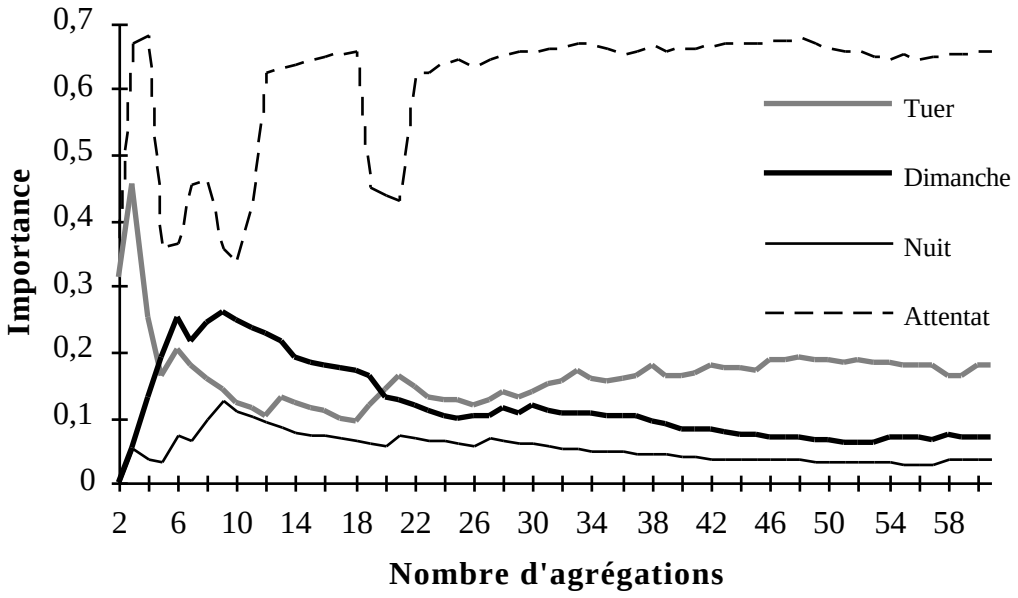


Fig. 10.3 - Évolution de l'importance de quatre mots de la signature thématique de la figure 10.1

Le degré de cohérence d'une signature dépend bien évidemment de son stade d'évolution. Il en est d'ailleurs de même du champ thématique qu'elle couvre. Plus la stabilisation d'une signature est avancée et plus sa cohérence devient grande tandis que dans le même temps, son champ d'application a tendance à s'élargir. Le fait de rassembler un ensemble d'événements spécifiques conduit en effet à faire émerger les éléments qui leur sont communs, c'est-à-dire les mots caractéristiques du domaine dans lequel ces événements s'inscrivent. Dans le cadre de l'expérimentation que nous avons menée, nous avons constaté que les signatures thématiques se stabilisent à la suite d'une vingtaine d'agrégations. On compte très exactement 48 signatures dans ce cas. Cette stabilisation est mesurée par un taux de variation, calculé par la formule suivante :

$$tauxVariation(ST(t)) = \sum_{i}^{n} poids(m_i, ST(t))\ \ rang(m_i, ST(t)) - \sum_{i}^{n} poids(m_i, ST(t-1))\ \ rang(m_i, ST(t-1)) \quad [3]$$

avec $ST(t)$: signature thématique ST après t agrégations;

rang(m_i,ST(t)) : rang du mot m_i dans la signature *ST*. Il est donné à partir du classement des mots de *ST* par ordre décroissant de leur poids;

poids(m_i,ST(t)) : poids du mot m_i, dans la signature *ST* (cf. [2]).

Ce taux de variation prend en considération les trois facteurs possibles de variation d'une agrégation à l'autre : les mots présents dans la tête de la signature (les *n* premiers mots de plus fort poids), leur poids ainsi que leur rang. Dans la formule [3], la référence est constituée par les *n* premiers mots de la signature *ST* au temps *t.* En conséquence, si l'un de ces mots ne faisait pas partie des *n* premiers mots de la signature *ST* au temps *t-1*, on considère que son rang et son poids sont égaux à 0 au temps *t-1*.

La figure 10.2 montre l'évolution de ce taux de variation pour les 30 premiers mots de la signature de la figure 10.1 (donc, pour *i* allant de 1 à 30). On constate de manière effective que ce taux de variation décroît assez rapidement jusqu'à une vingtaine d'agrégations pour devenir ensuite quasiment stable (très légère décroissance qui n'est pas réellement significative). Il ne devient en pratique jamais nul car la variabilité dans la présence des mots d'un texte à un autre entraîne toujours une variation résiduelle de faible ampleur. Par ailleurs, si le nombre de mots retenu pour définir la tête de la signature est trop important, on risque d'observer une stabilisation moins nette puisque les mots de faible poids ont par définition une présence très erratique dans les textes[1].

Le phénomène de stabilisation existant à l'échelle d'une signature toute entière se manifeste également à l'échelle plus microscopique des mots constituant cette signature. On peut le constater au niveau de la figure 10.3 pour quatre mots de la signature de la figure 10.1 : leur importance, évaluée par le rapport *poids du mot / rang du mot*, passe d'abord par des fluctuations plus ou moins erratiques couvrant un peu plus des vingt premières agrégations; mais cette période laisse ensuite la place à une phase de stabilisation progressive permettant de mettre en évidence le statut véritable de ces mots vis-à-vis de la signature. Des mots liés au domaine représenté comme "tuer" ou "attentat" ont ainsi une pente de stabilisation ascendante vers des valeurs d'importance élevées tandis que des mots tels que "nuit" ou "dimanche", sans attache particulière avec le domaine considéré, se stabilisent suivant une pente descendante vers de faibles valeurs d'importance.

[1] On pourrait d'ailleurs imaginer d'utiliser le critère de stabilisation d'une signature pour définir précisément l'ensemble de ses mots les plus significatifs. Il serait nécessaire pour ce faire d'enregistrer sur une période suffisamment longue l'évolution du taux de variation de la tête de la signature pour différentes tailles de cette tête. La frontière entre mots significatifs et mots non significatifs serait alors déterminée par la tête de plus grande taille pour laquelle on obtient une stabilisation sur la durée d'observation.

Comme on pouvait s'y attendre compte tenu de leur mode de formation, on observe également que les signatures comportent beaucoup de bruit, c'est-à-dire de mots qui ne sont pas spécifiquement attachés aux thèmes représentés par ces signatures. C'est le cas par exemple de mots comme "lundi" ou "dimanche" dans la signature de la figure 10.1. L'importance du bruit apparaît plus nettement encore sur le plan quantitatif : cette signature contient un total de 705 mots venant des textes et de 401 mots inférés[1], mais seulement 12 mots venant des textes et 66 mots inférés ont un poids suffisant pour intervenir dans la sélection des signatures ou dans la mesure de similarité. Au passage, on constate que filtrer le bruit en fixant un simple seuil de poids est un critère assez grossier puisque des mots tels que "mort" ou "séparatiste", pourtant liés au domaine du terrorisme, se retrouvent laissés de côté. C'est néanmoins la moins mauvaise solution que l'on puisse adopter dans un contexte aussi frustre, en espérant que les mots de plus fort poids soient suffisants sur le long terme pour attirer des UTLs qui contribueront à renforcer les autres mots propres au domaine représenté.

1.3.6. Évaluation et discussion

Ainsi que nous l'avons indiqué au paragraphe précédent, notre méthode de construction de représentation de thèmes n'a pas fait l'objet d'une évaluation directe de ses résultats par recoupement du jugement de plusieurs sujets. Elle n'a pas non plus été évaluée de façon indirecte au travers de l'utilisation des représentations construites dans une tâche tel que la tâche Détection de TDT. Si l'on se heurte dans le premier cas à la lourdeur de mise en œuvre d'une expérimentation de type psychologique, on est confronté dans le second à l'ampleur de la tâche consistant à réunir un corpus homogène et surtout, à annoter thématiquement les différents textes qui le composent. Or, les seuls corpus de ce genre (en particulier le corpus TDT, mais également un corpus de dépêches Reuters utilisé en catégorisation de textes) rassemblent des textes en anglais américain que nous ne pouvons donc pas utiliser pour le moment[2].

Concernant plus spécifiquement le cas de TDT, notons simplement que les valeurs de certains indicateurs données dans [Yang et alii 1997] pour le travail le plus proche du nôtre, en l'occurrence celui de CMU, sont comparables à nos valeurs pour les mêmes indicateurs : 15863 segments donnent lieu à la formation de 5907 regroupements dans le cas de CMU, soit un rapport moyen de 2,67 segments par regroupements, ce qui est équivalent à notre rapport de 2,65 UTLs par signature thématique; par ailleurs, pour une

[1] En moyenne, les mots inférés sont aussi nombreux que les mots venant des textes mais ce point est caractérisé par d'assez larges variations. Par ailleurs, il dépend bien entendu du nombre de mots imposé pour la sélection d'un mot du réseau de collocations.

[2] Tous les outils externes que nous utilisons (*Mtseg*, *TreeTagger* et *INTEX*) sont disponibles pour l'anglais mais nous sommes limité par la nécessité de constituer un réseau de collocations de large taille dans cette même langue si nous voulons véritablement y transposer notre travail.

mesure de similarité proche dans son principe de la nôtre, la valeur seuil de rattachement à un regroupement est égale à 0,23 pour CMU et à 0,25 dans notre cas.

À défaut d'une évaluation objective de la qualité et de la pertinence des représentations de thèmes construites, nous avons tout de même mené quelques expérimentations destinées au moins à tester certaines propriétés de la méthode ainsi qu'à confirmer, ou au contraire à infirmer l'intérêt de certains choix réalisés. Trois points ont ainsi été abordés : l'influence de l'ordre de traitement des textes, compte tenu du caractère incrémental de la méthode; l'intérêt de la présence des mots inférés pour améliorer la qualité des signatures et, dans le même esprit, l'intérêt de la segmentation des textes. En l'absence d'une procédure d'évaluation permettant de juger de la supériorité d'une condition par rapport à une autre, nous nous sommes appuyé uniquement sur des critères de jugement formels. Plus précisément, nous avons considéré qu'un état final *EM* de la mémoire est plus intéressant qu'un autre état *EM´* si *EM* se caractérise par un taux moyen d'agrégation des signatures plus important et un nombre de signatures plus faible. La justification de ce critère est simple : la forme des représentations au sein de SEGAPSITH favorise spontanément plutôt l'absence de similarité des UTLs que leur similarité. Tout mouvement dans le sens d'une meilleure détection de la similarité entre UTLs est donc considéré comme un progrès.

Influence de l'ordre de traitement des textes

À l'occasion de la présentation de la mémoire épisodique au chapitre 6, nous avions discuté des implications de l'incrémentalité de l'apprentissage au sein de MLK et plus précisément de l'influence de l'ordre de traitement des textes sur le résultat de cet apprentissage. La petite taille du jeu d'essai de MLK nous avait conduit surtout à examiner cette influence sur la forme d'une UT particulière. SEGAPSITH, qui est confronté au même problème étant donné sa communauté de principe avec MLK, offre l'opportunité de juger de l'impact de ce facteur sur une plus grande échelle. Pour ce faire, nous avons extrait du corpus originel le sous-ensemble de textes[1] impliqués dans la construction d'un certain nombre de signatures obtenues lors de l'expérimentation initiale et nous avons mémorisé leurs UTLs en modifiant à chaque nouvelle expérience leur ordre de présentation de façon aléatoire.

Bien entendu, l'état final atteint par la mémoire n'est pas le même d'une expérience à une autre. Au niveau global cependant, les différences observées sont la plupart du temps

[1] Nous n'avons pas travaillé sur l'ensemble du corpus dans la mesure où nous souhaitions pouvoir examiner assez précisément les différences entre les signatures construites dans une condition ou une autre, ce qui n'aurait pas été possible avec l'ensemble des textes.

relativement minimes. Plus exactement, les signatures les plus représentatives, c'est-à-dire celles ayant atteint un certain niveau de stabilité, se retrouvent sous des formes similaires d'une expérience à une autre. La variabilité est en revanche suffisamment forte en ce qui concerne les signatures faiblement agrégées pour empêcher leur mise en correspondance systématique entre deux expériences. Il faut à cette occasion souligner une difficulté que pose l'apprentissage non supervisé du point de vue de l'évaluation directe des représentations construites. Celui-ci implique une absence de référence commune quant au résultat à obtenir. Les comparaisons d'une expérience à une autre s'en trouvent de ce fait particulièrement compliquées car rien n'indique qu'une signature dans l'une est identique à une signature dans l'autre, en dehors de l'interprétation que peut en faire un sujet humain.

En dépit de cette difficulté et du caractère partiel des expérimentations menées ici, on peut néanmoins avancer que sur le long terme, l'ordre de traitement des textes ne semble pas influer de façon déterminante sur la nature des signatures thématiques construites dès lors que celles-ci sont en final suffisamment stables.

Bénéfice des mots inférés

Tous les travaux présentés au §1.2 concernant la construction de représentations de thèmes se fondent uniquement sur les mots apparaissant de façon explicite dans les textes puisqu'aucun d'entre eux n'a recours à une source de connaissances externe aux textes. En cela, l'utilisation des mots inférés constitue l'une des originalités de notre méthode. Rappelons que leur présence a pour objectif de favoriser la détection de la similarité entre des UTLs traitant du même thème mais ne partageant que peu de mots en raison de manières différentes de l'évoquer (variabilité de l'expression, niveaux de généralité différents ou degrés d'explicitation plus ou moins forts).

Afin de tester le bénéfice réel de ces mots inférés, nous avons relancé le processus de construction des signatures sur la totalité de notre corpus de dépêches AFP en ne faisant pas intervenir les mots inférés. La différence entre l'état final de la mémoire dans ce cas et cet état dans le cas initial est conforme à nos attentes. Le nombre de signatures est en effet plus important en l'absence des mots inférés et leur nombre d'agrégations est en moyenne plus faible. Ce phénomène touche y compris les signatures considérées dans l'expérimentation initiale comme stables et qui se trouvent donc éclatées en plusieurs signatures moins significatives. La signature de la figure 10.1 est une illustration exemplaire de ce phénomène. Alors qu'avec les mots inférés, elle rassemble 61 UTLs, ce chiffre tombe à 34 lorsque seuls les mots des textes sont pris en considération. Par ailleurs, les 27 UTLs ne s'agrégeant plus avec cette signature sont mémorisées pour la

plupart sous la forme de nouvelles signatures, contribuant ainsi l'augmentation du nombre global de signatures non significatives.

À la lumière des critères que nous avons adoptés, nous estimons donc que les mots inférés apportent un véritable bénéfice dans le sens d'une qualité et d'une pertinence accrues des signatures thématiques. Nous verrons même au §1.4 que ces mots sont plus intéressants que les mots venant des textes pour la représentation des thèmes.

Bénéfice de la segmentation thématique des textes

Le dernier point que nous avons testé est l'intérêt, pour construire les signatures, d'utiliser des segments de texte produits par un mécanisme de segmentation thématique, même imparfait, plutôt que des textes entiers. Ce point est d'ailleurs joint dans notre cas à l'intérêt de laisser de côté des segments évalués comme thématiquement peu cohérents. La segmentation thématique de SEGCOHLEX et celle de SEGAPSITH ont en effet plus vocation à mettre en évidence des portions de texte particulièrement cohérentes sur le plan thématique qu'à réaliser un strict découpage thématique des textes, tâche le plus souvent impossible dans les passages où les thèmes s'entremêlent étroitement.

De prime abord, l'intérêt de cette segmentation paraît évident : plus les unités textuelles agrégées sont homogènes et plus la cohérence des représentations résultantes devrait être grande. Au contraire, leur hétérogénéité accroît leur spécificité et conduit en final à la création d'une multitude de signatures ne rassemblant qu'une seule UTL. Force est néanmoins de constater que cette démarche n'est généralement pas adoptée dans les travaux que nous avons mentionnés précédemment, à l'exception du cas un peu particulier de ceux s'inscrivant dans TDT, confrontés non à des textes mais à un véritable flux de mots. Même la structuration explicite des textes mise en place par leurs auteurs (division en sections, sous-sections, etc.) n'est souvent pas exploitée.

Pour confirmer le bien-fondé de la segmentation des textes, nous avons à nouveau construit un ensemble de signatures à partir de la totalité du corpus de l'AFP mais en ne produisant cette fois-ci qu'une seule UTL pour chaque dépêche. Le résultat atteint est similaire dans sa tendance au résultat de l'expérience de suppression des mots inférés : le nombre de signatures augmente significativement par rapport à l'expérimentationinitiale et donc, chacune d'elles regroupe en moyenne moins d'UTLs. La tendance est encore plus marquée dans le cas présent, ainsi que l'illustre la signature de la figure 10.1 : les 61 UTLs initialement regroupées par cette signature se réduisent à seulement 21 UTLs lorsque les dépêches ne sont plus segmentées. On passe donc d'une signature que l'on peut considérer comme stable au vu des figures 10.2 et 10.3 à une signature dont le nombre d'agrégations est à peine suffisant pour établir une éventuelle stabilité.

Il ne fait donc guère de doute que la segmentation des textes possède une influence importante dans la production de signatures à la fois significatives et les moins bruitées possible. Il reste néanmoins à établir si le processus de segmentation que nous mettons en œuvre ici est plus performant de ce point de vue qu'une segmentation de référence établie sur des critères très simples. Ceci renvoie à la notion d'évaluation indirecte de la segmentation thématique évoquée à la fin de la section 2 du chapitre précédent. La segmentation de référence pourrait correspondre en l'occurrence à une succession de segments de même taille (voisine de la taille moyenne des segments délimités ici) ou bien reprendre une segmentation extraite de la forme des textes, comme celle portée par les paragraphes. Pour que la comparaison des performances respectives des deux segmentations soit complète, il sera nécessaire de calculer une mesure de cohésion pour chaque segment de la segmentation de référence, ce qui peut être fait de la même façon que dans la segmentation de SEGCOHLEX utilisée pour la présente expérimentation (utilisation d'une fenêtre glissante et sélection des mots du réseau de collocations).

1.4. Construction de signatures thématiques et structuration d'un réseau de collocations

Dans cette section, notre objectif n'est pas de proposer une nouvelle méthode de construction des signatures thématiques à côté de celle exposée ci-dessus mais plutôt d'en présenter une évolution possible, tirée d'une analyse des résultats obtenus lors de l'expérimentation sur le corpus des dépêches AFP.

1.4.1. Analyse des résultats de la construction automatique des signatures thématiques

Lorsqu'on examine en détail le contenu des signatures thématiques construites lors de l'expérimentation réalisée avec les dépêches AFP, comme celle de la figure 10.1, on est frappé de constater que les mots inférés sont généralement plus intéressants que les mots venant des textes. Ils sont plus homogènes sur le plan thématique et la proportion d'entre eux ayant un poids élevé est plus importante. À titre indicatif, parmi les signatures de cette expérimentation rassemblant au moins 20 UTLs, c'est-à-dire considérées comme à peu près stables, on observe que la proportion de mots significatifs (poids ≥ 0,1) est en moyenne égale à 16,9% pour les mots inférés tandis qu'elle n'est que de 2,66% pour les mots des textes.

Ce phénomène apparaît plus globalement au niveau de la figure 10.4. Celle-ci met en balance la distribution des nombres d'occurrences des mots inférés (ce qui équivalent à leur poids pour ce qui nous intéresse ici) avec celle des nombres d'occurrences des mots

des textes. Cette comparaison laisse d'abord apparaître que les mots inférés fortement récurrents, donc significatifs, sont globalement plus nombreux que leurs homologues venant des textes (courbe des mots inférés allant plus loin que celle des mots des textes vers les fortes valeurs de nombre d'occurrences). Elle montre ensuite que le nombre des mots inférés est plus important que celui des mots des textes lorsque leur nombre d'occurrences se situe au delà d'une valeur environ égale à 20 occurrences (courbe des mots inférés située au-dessus de celle des mots des textes à partir de cette valeur). Cette même valeur, transposée en nombre d'agrégations, définit comme on l'a vu le moment où la stabilité des signatures commence à se dessiner. On en déduit que les signatures stables contiennent en général davantage de mots significatifs inférés que de mots significatifs venant des textes.

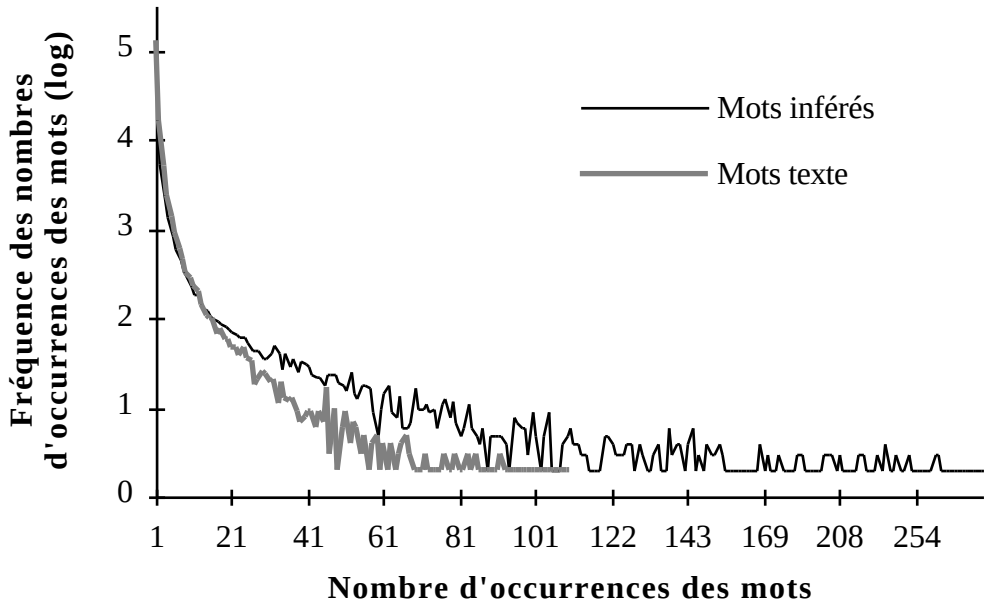


Fig. 10.4 - Distribution du nombre d'occurrences des mots des signatures

Enfin, la figure 10.5 donne un exemple plus direct de la différence entre les mots des textes et les mots inférés pour une signature relative au domaine de la justice. On a fait apparaître ici ses dix mots de plus fort poids après avoir supprimé les mots apparus à la fois comme mots du texte et comme mots inférés (ces derniers sont généralement cohérents sur le plan thématique). On voit sans conteste que les mots inférés résiduels restent majoritairement en liaison avec le thème abordé lorsqu'ils ont un poids suffisant, ce qui est le cas ici, alors que cette liaison est beaucoup plus erratique en ce qui concerne les mots des textes (seul le mot "appel" est lié au domaine concerné). Même s'il n'est pas toujours aussi marqué, ce phénomène est néanmoins général, ainsi que l'atteste la signature de la figure 10.1. En accord avec les données quantitatives exposées ci-dessus,

on constate également que le poids des mots des textes a tendance à baisser beaucoup plus rapidement que celui des mots inférés. En définitive, tous ces éléments laissent à penser que dans une signature, les mots inférés rassemblent la plupart des mots venant des textes lorsqu'ils sont significatifs.

mots des textes		mots inférés	
mot	poids	mot	poids
an	0,990	avocat_général	0,884
dollar	0,326	correctionnel	0,884
avril	0,323	réclusion_criminelle	0,840
ancien	0,314	juré	0,840
dernier	0,298	cour_d'assises	0,840
chef	0,268	condamner_à_mort	0,754
communiste	0,253	peine_de_mort	0,733
mois	0,242	homicide	0,647
albanais	0,224	chambre_correctionnelle	0,647
appel	0,172	extradition	0,625

Fig. 10.5 - Mots d'une signature relative au domaine de la justice après suppression de ses mots présents à la fois comme mots inférés et mots des textes

Qu'elles soient qualitatives ou quantitatives, les observations présentées ci-dessus vont toutes dans le même sens : les mots des textes apportent proportionnellement beaucoup plus de bruit que les mots inférés. Cette constatation n'est d'ailleurs pas surprenante sur le fond. Les mots inférés font l'objet d'une sélection visant à ne retenir que les mots en rapport avec le thème développé. Le mécanisme de sélection est certes assez imprécis étant donné la faible structuration de la source de connaissances utilisée mais il possède néanmoins une certaine efficacité ainsi que l'atteste le contenu des signatures construites. Les mots des textes sont pour leur part des données intangibles et ne contiennent en général que peu de mots fortement liés au domaine abordé. Ceci explique leur état de source importante de bruit du point de vue de la construction des signatures.

L'analyse développée à propos de la figure 10.5 nous a montré par ailleurs que ce bruit ne s'accompagne que d'un nombre faible de mots significatifs autres que ceux présents comme mots inférés. Dans la mesure où ils contribuent fortement à faire diminuer le rapport signal sur bruit, l'utilité des mots des textes dans la constitution des signatures peut donc apparaître comme discutable.

1.4.2. Évolution de la méthode de construction des signatures thématiques

Compte tenu de l'analyse exposée au paragraphe précédent, la solution la plus évidente pour améliorer la qualité des signatures thématiques consiste simplement à ne plus tenir compte des mots des textes dans le processus de leur construction. Les signatures

obtenues ne contiennent alors plus que des mots inférés. En dehors de ce changement, leur méthode de construction demeure toutefois strictement inchangée.

Cette évolution peut apparaître un peu étrange au regard des principes initialement développés : le but était en effet de créer des représentations de thème en agrégeant des segments de texte similaires, les mots inférés n'étant conçus que comme un moyen de faciliter la détection de la similarité entre les segments. Selon la nouvelle perspective, les segments n'ont plus pour rôle de fournir le matériau de construction des signatures mais de guider la sélection d'un ensemble de mots du réseau de collocations liés au thème abordé par le segment. La constitution des signatures est alors vue comme un processus de structuration de ce réseau sur le plan thématique. Considéré sous cet angle, elle crée un niveau parallèle à celui du réseau de collocation abritant un ensemble de nœuds sans relation les uns avec les autres mais projetant vers les mots du réseau un ensemble de liens pondérés. Chacun de ces nœuds s'identifie à une signature dont les constituants sont définis par les liens projetés vers les mots du réseau. Le poids attaché à chacun d'entre détermine l'importance du mot par rapport à la signature. Précisons à nouveau qu'il ne s'agit là que d'un changement de perspective et non d'une modification des représentations des UTLs et des signatures thématiques.

Afin de vérifier que la seule prise en compte des mots inférés se traduit effectivement par une meilleure qualité des signatures, nous avons relancé une dernière fois le processus de construction des signatures sur la totalité du corpus de l'AFP en faisant abstraction des mots des textes. Pour mener cette expérience, nous n'avons pu conserver que 7823 des 8601 UTLs initiales. 778 de ces UTLs ne possèdent en effet aucun mot inféré. Cette perte n'a cependant pas de conséquences néfastes puisque les segments pour lesquels aucun mot inféré n'a été sélectionné sont probablement assez peu cohérents sur le plan thématique.

Ces 7823 UTLs ont donné lieu à la construction de 783 signatures thématiques. Parmi elles, 570 résultent d'au moins 2 agrégations. 27% des signatures contiennent ainsi une seule UTL, à comparer aux 79% de l'expérimentation avec les mots des textes. Le nombre de signatures considérées comme stables (nombre d'agrégations ≥ 20) passe dans le même temps de 48 à 69. Les caractéristiques des signatures les plus agrégées demeurent en revanche globalement les mêmes : la signature la plus agrégée contient 351 UTLs, ce qui est à cette échelle assez proche des 413 UTLs de la première expérience, et le nombre de signatures ayant un nombre d'agrégations supérieur à 100 est dans ce cas aussi égal à 13. Cette proximité des résultats pour les hautes valeurs du nombre d'agrégations est illustratrice d'une insuffisance de la méthode de construction des signatures, insuffisance qui conduit à former régulièrement une minorité de regroupements assez massifs présentant le double inconvénient d'être trop hétérogènes et de jouer un rôle d'attracteur trop puissant.

En dehors de ce point précis, l'amélioration des signatures construites constatée au vu des indicateurs globaux nous servant de critères quantitatifs se retrouve sur le plan qualitatif lorsque l'on observe en détail le contenu des signatures. La proportion des mots d'une signature en relation étroite avec le thème représenté par celle-ci est ainsi nettement plus élevée ici que lors de l'expérience avec les mots des textes. On remarque par ailleurs que les mots restent pertinents vis-à-vis du thème représenté bien en dessous de la limite de poids de 0,1 fixée précédemment. Bien entendu, il ne s'agit là encore que d'une opinion subjective demandant à être confirmée par une évaluation adéquate mais nous estimons que les éléments quantitatifs avancés ci-dessus sont à eux seuls suffisamment significatifs pour admettre que la suppression des mots des textes dans les signatures permet d'améliorer la qualité de ces dernières.

2. La segmentation thématique de SEGAPSITH

2.1. Introduction

Le mécanisme de segmentation thématique que nous présentons dans cette section joue le même rôle au sein de SEGAPSITH que celui joué dans MLK par l'analyse thématique développée au chapitre 8. Il est à ce titre l'un des deux maillons de l'amorçage intra-niveau, l'autre étant incarné par l'extraction de signatures thématiques exposée ci-dessus. Il opère donc à partir des connaissances, en l'occurrences des signatures thématiques, produites par cette extraction. Son objectif est identique dans ses grandes lignes à celui de SEGCOHLEX, c'est-à-dire produire une représentation des textes sous la forme d'un ensemble d'Unités Thématiques Lexicales (UTLs), chacune d'entre elles caractérisant un segment de texte thématiquement homogène.

En dehors du mécanisme de segmentation proprement dit, les différences avec SEGCOHLEX portent essentiellement sur deux points. D'une part, la segmentation de SEGAPSITH peut être plus précise que celle de SEGCOHLEX dans la mesure où elle est fondée sur une source de connaissances rendant compte spécifiquement des relations thématiques entre mots. D'autre part, elle offre la possibilité de mettre en évidence des segments discontinus, c'est-à-dire d'associer en une seule UTL plusieurs manifestations d'un même thème situées à des endroits différents d'un texte. Pour ce faire, elle reprend une part importante des principes supportant la segmentation de MLK et offre par la même occasion à cette dernière une forme indirecte de validation.

2.2. L'utilisation des signatures thématiques pour la segmentation des textes

2.2.1. Principes

La segmentation de SEGAPSITH peut être vue comme une forme d'instanciation de la segmentation de MLK dans le cadre de ROSA. De ce fait, les deux segmentations sont sous-tendues par un même principe général. Une *unité d'analyse* des textes est définie. Il s'agit de la proposition dans le cas de MLK et du mot dans le cas de ROSA. Les différentes unités d'un texte sont traitées en séquence et lors de son traitement, chacune d'entre elles est considérée en englobant simultanément ses voisines les plus proches[1]. Il s'agit à la fois des unités d'analyse adjacentes ayant déjà été traitées et de celles qui seront traitées dans les cycles qui suivent. L'objectif est de considérer chaque unité d'analyse dans le contexte d'une portion de texte plus large, tout en conservant néanmoins un caractère central à cette unité d'analyse. Ce contexte permet de lever d'éventuelles ambiguïtés en même temps qu'il contribue à lisser les variations causées par le passage d'une unité à une autre.

L'analyse proprement dite consiste à déterminer si une unité d'analyse, abordée bien entendu à la suite de celles qui la précèdent dans le texte, se rattache à l'une des Unités Thématiques (UTs) déjà distinguées ou bien au contraire si elle doit donner lieu à la création d'une nouvelle UT. Le rattachement d'une unité d'analyse à une UT ne s'effectue pas par comparaison directe du contenu de l'unité et de celui de l'UT, ainsi que le pratiquent certaines méthodes de segmentation quantitatives (cf. TextTiling de Hearst par exemple). Il s'appuie sur la recherche d'un lien entre les connaissances sur les situations évoquées par l'unité d'analyse et celles évoquées par l'UT considérée. Ces connaissances prennent en l'occurrence la forme d'Unités Thématiques agrégées. Dans le cas de ROSA, ces UTs agrégées portent le nom plus spécifique de signature thématique.

Compte tenu de la nature incertaine, imprécise et incomplète de ces UTs agrégées, ce lien n'est pas établi comme dans le cas de Grau ou de Grosz et Sidner sur la base de relations explicites existant entre les UTs agrégées sélectionnées par l'unité d'analyse et celles sélectionnées par l'UT. Il est mis en évidence par une mesure de similarité entre ces deux ensembles d'UTs agrégées. On raisonne donc en termes de similitude des configurations d'UTs agrégées évoquées en essayant de compenser l'absence de précision par une vue d'ensemble. Chacune de ces configurations est appelée *contexte.* Le contexte d'une unité d'analyse est établi sur la base de la sélection directe des UTs

[1] La taille de ce voisinage fait partie des paramètres à fixer lorsque l'on met au point finement chaque méthode particulière.

agrégées les plus activées par le contenu de cette unité[1]. Celui d'une UT en construction est le résultat de la fusion de tous les contextes des unités d'analyse qui le composent.

Pour déterminer si une unité d'analyse se rattache à une UT en construction, on calcule la similarité entre leurs contextes. Si la valeur obtenue dépasse un seuil fixé a priori, le rattachement est accepté; sinon, il est rejeté. Globalement, la segmentation d'un texte consiste donc à essayer de rattacher chaque unité d'analyse à l'une des UTs en construction actives à ce moment-là du processus. Si ce rattachement échoue, l'unité en question donne lieu à la création d'une nouvelle UT en construction.

Ce principe général a subi quelques modifications afin de s'adapter aux caractéristiques de SEGAPSITH. Ces adaptations ne concernent que la stratégie de recherche d'une UT en construction pour le rattachement de l'unité d'analyse courante. Au niveau de SEGAPSITH, l'unité d'analyse est réduite au mot. Mais disperser une suite de mots indépendamment les uns des autres n'aurait pas grand sens. Par ailleurs, travailler à partir de la forme de surface des textes et non à partir d'une représentation sémantique ne permet pas d'être très précis. On ne peut donc pas, comme dans MLK, décider de l'affectation d'unités textuelles de petites tailles de façon assez libre. On cherche surtout à former des segments continus d'une certaine consistance. Une UTL peut être formée de plusieurs segments non contigus de ce type mais ceux-ci ne se réduisent pas à seulement quelques mots.

La stratégie générale de la segmentation de SEGAPSITH est en fait une forme de synthèse entre celle de MLK et celle de SEGCOHLEX. À l'instar de MLK, le processus maintient une liste des UTLs en construction avec pour chacune d'entre elles, un contexte composé de signatures thématiques. Néanmoins, comme dans SEGCOHLEX, le but que l'on cherche prioritairement à atteindre est de déterminer pour une position donnée du texte si l'on est en présence ou non d'un changement de thème. Cette décision est établie par rapport à un thème courant, supposé stable. Ce thème est bien entendu représenté par une des UTLs en construction gérées tout au long de la segmentation. Celle-ci est appelée *UTL active*. Il s'agit plus précisément de l'UTL à laquelle ont été rattachés les mots occupant les positions précédant la position courante.

Dans ce cadre, la mesure de similarité entre contextes est d'abord utilisée afin d'estimer le degré de proximité du contexte de l'UTL active et du contexte associé à la position courante du texte. Si cette proximité est jugée suffisante, le mot occupant la position courante est rattaché à cette UTL. Dans le cas contraire, on opte en faveur d'un

[1] L'activité antérieure des UTs agrégées intervient également dans le cas de la mémoire épisodique de MLK.

changement de thème, et donc également d'un changement de l'UTL active. La définition de cette dernière reprend les principes de la segmentation de MLK : on évalue la similarité du contexte associé à la position courante du texte avec les contextes de toutes les UTLs en construction, à l'exception de l'UTL active. Si cette similarité est assez forte pour l'une de ces UTLs, celle-ci devient la nouvelle UTL active. Il s'agit donc du retour d'un thème déjà abordé précédemment dans le texte. Une absence de similarité suffisante se traduit quant à elle par la création d'une nouvelle UTL en construction.

En revanche, nous ne reprenons pas au niveau de SEGAPSITH la notion d'UT inédite unique introduite dans MLK. Rappelons qu'une UT inédite a pour fonction de représenter une situation, elle-même désignée comme inédite, à propos de laquelle on ne dispose pas de connaissances en mémoire. Dans la segmentation de MLK, la reconnaissance explicite des situations de ce type est nécessaire dans la mesure où l'on risquerait, dans le cas contraire, de créer une nouvelle UT en construction pour chaque proposition représentant un événement propre à une situation inédite. La similarité entre le contexte d'une telle proposition et les contextes des UTs en construction est en effet généralement faible du fait de l'absence de cohérence interne et de représentativité du contexte de la proposition.

Ce problème ne se pose pas de façon aussi aiguë dans le cas de SEGAPSITH étant donné que l'analyse y est dirigée par la détection des transitions davantage que par l'affectation des unités d'analyse. L'évocation d'une situation inédite se traduit par un changement de thème et non pas par une succession de créations de nouvelles UTLs en construction. L'algorithme de segmentation peut rester dans un état de changement de thème jusqu'à ce qu'à ce qu'un nouveau thème stable ait été détecté, cette période pouvant être aussi longue que nécessaire pour couvrir l'évocation d'une situation inédite. Mais contrairement à ce qui se passe dans MLK, les différents passages de ce type ne sont pas réunis en une seule UT. Ils n'ont pas en effet de nécessité à être réunis par construction et il n'y a pas plus de raisons de penser qu'ils sont les différentes évocations d'une même situation inédite plutôt que les évocations de plusieurs situations inédites.

Pour terminer l'exposé des grands principes de la segmentation de SEGAPSITH, précisons qu'en dehors des signatures thématiques, celle-ci fait également appel au réseau de collocations de SEGCOHLEX. Cette utilisation n'a qu'un caractère optionnel, sans influence sur la forme de l'algorithme de segmentation, mais contribue en pratique à améliorer ses performances en rendant la sélection des signatures thématiques plus pertinente et plus sûre. Elle vient à ce titre pallier en partie l'inexistence d'une représentation explicite des concepts dans le cadre de SEGAPSITH. Outre une réduction des problèmes de polysémie et d'homonymie inhérents à l'utilisation des seuls mots, la présence de connaissances conceptuelles, même embryonnaires, permettrait idéalement

d'élargir les indices de rappel des signatures thématiques en ne se contentant pas des seuls mots présents dans les énoncés. À l'image de ce qui est fait dans MLK, des concepts sémantiquement proches de ceux évoqués par ces mots pourraient en effet être mobilisés et contribuer ainsi à focaliser davantage la sélection des signatures thématiques. À défaut de ces connaissances, le réseau de collocations constitue une solution de remplacement minimale. Il offre la possibilité de définir pour chaque unité d'analyse un ensemble de mots entretenant une forte cohésion lexicale avec les mots caractérisant cette unité. Les deux groupes de mots sont ensuite associés afin de sélectionner les signatures thématiques les plus en rapport avec le thème de l'unité.

2.2.2. Détail du mécanisme

L'algorithme définissant précisément la manière dont est réalisée la segmentation d'un texte dans SEGAPSITH est donné par les figures 10.6 et 10.7. Comme dans le cas de SEGCOHLEX, il prend comme entrée des textes pré-traités suivant la procédure décrite au paragraphe 2.3.1 du chapitre 9. Les textes ne contiennent donc plus que la forme canonique de leurs mots pleins. La définition de l'unité d'analyse et de son environnement immédiat s'effectue de façon strictement identique par rapport à SEGCOHLEX : on utilise une fenêtre de taille fixe que l'on fait glisser sur l'ensemble du texte. À chaque station de la fenêtre, l'unité d'analyse courante est définie par le mot situé au centre de la fenêtre. Le traitement du début et de la fin du texte fait intervenir également un mécanisme de recopie du premier ou du dernier mot.

La sélection des mots du réseau de collocations associés à une position de la fenêtre se déroule, quand elle a lieu, exactement suivant la méthode utilisée dans SEGCOHLEX lors du calcul de la cohésion au sein de la fenêtre glissante. Cette méthode est décrite au paragraphe 2.4.1 du chapitre 9 : on ne retient que les mots du réseau liés à un certain nombre, fixé a priori, de mots de la fenêtre. Les mots de la fenêtre et les mots du réseau de collocations, lorsqu'ils sont pris en compte, sont ensuite pondérés conformément à la procédure adoptée dans SEGCOHLEX : le poids initial de chacun de ces mots voit sa valeur croître plus ou moins fortement en fonction du nombre et de la force de cohésion des liens qu'il entretient, au travers du réseau de collocations avec les autres mots considérés (cf. également §2.4.1 du chapitre 9). Bien entendu, lorsque le réseau de collocations n'est pas utilisé, on se contente des poids initiaux.

Une fois pondérés, les mots composant la fenêtre glissante, éventuellement accompagnés des mots sélectionnés à partir du réseau de collocations, sont utilisés afin d'activer les signatures thématiques présentes en mémoire et définir ainsi le contexte associé à l'unité d'analyse courante (cf. fonction activationSignatures() de la figure 10.6). À

l'image du contexte de la proposition courante dans MLK, ce contexte est donc une liste de signatures thématiques pondérées en fonction du niveau d'activité qu'elles ont à la suite de cette phase d'activation. La procédure d'activation est identique en tout point à celle décrite au §1.3.3 de ce chapitre pour la mémorisation des UTLs. En particulier, la fonction d'activation des signatures thématiques est reprise telle quelle. La seule différence réside dans la sélection des signatures les plus activées : ici, on retient seulement les *N* signatures les plus activées, *N* correspondant à la taille choisie aussi bien pour les contextes des unités d'analyse que pour les contextes des UTLs en construction.

Le contexte d'une UTL en construction est défini quant à lui de la même façon que le contexte d'une UT en construction au niveau de la segmentation de MLK : il est le résultat de la fusion des contextes des différentes unités d'analyse qu'il a intégrées tout au long de l'analyse. Cette fusion est réalisée suivant la même procédure que celle décrite au §2.2 du chapitre 8 (cf. fonction majContexteUTLavecContexte() de la figure 10.6) : les deux listes de signatures sont fusionnées, avec un recalcul du poids des signatures communes, la liste résultante est triée par ordre décroissant des poids des signatures et finalement seules les *N* premières signatures sont conservées. La fonction *F* de recalcul du poids des signatures communes reprend la fonction linéaire donnée au chapitre 8 (cf. formule [3] du §2.2) :

$$poids(t+1, UTLC_j, ST_i) = poids(t, UTLC_j, ST_i) + \beta \;\; poids(UA_t, ST_i)$$

avec $poids(t,UTLC_j,ST_i)$: poids de la signature thématique ST_i dans le contexte de l'UTL en construction $UTLC_j$ au temps *t*;

$poids(UA_t,ST_i)$: poids de la signature ST_i dans le contexte de l'unité d'analyse considérée au temps *t*.

L'algorithme proprement dit de segmentation se définit à partir de trois paramètres, dont l'articulation est illustrée par l'automate de la figure 10.7 (celui-ci est représenté dans l'algorithme de la figure 10.6 au travers de la fonction GestionÉtatSegmentation()) :

- l'état du segmenteur thématique. Cet état peut prendre quatre valeurs : #changementDeThème, #développementThème, #détectionNouveauThème, #détectionChangementDeThème. La présence des deux premiers états est normale pour un segmenteur thématique. Lorsque les unités d'analyse qu'il considère se trouvent au sein d'un segment thématique, l'état du segmenteur est égal à #développementThème tandis que lorsque l'unité considérée marque la frontière entre deux segments, son état prend la valeur #changementDeThème.
  Néanmoins, dans le cadre de SEGAPSITH, l'unité d'analyse est très fine – c'est le mot – et malgré la présence d'un environnement important dans la fenêtre glissante, la méthode de segmentation ne peut pas être très sûre à un niveau aussi fin. Il nous a donc

paru nécessaire de considérer les transitions plutôt comme des intervalles que comme des points. On peut avoir de cette manière confirmation d'un changement d'état en observant que la tendance a effectivement changé sur toute l'étendue de l'intervalle. Ce principe s'applique aussi bien pour le passage d'un thème stable à un changement de thème que pour le passage d'un changement de thème à un nouveau thème stable. Dans le premier cas, l'état intermédiaire est l'état #détectionChangementDeThème tandis que dans le second, il s'agit de l'état #détectionNouveauThème. Le dispositif est plus particulièrement utile dans le premier cas. Il évite en effet qu'un passage de texte peu marqué thématiquement soit interprété nécessairement comme un changement de thème;

- la valeur de la similarité entre le contexte associé à la position courante de la fenêtre et le contexte associée à l'UTL active (variable *valSim* de la figure 10.6). Cette valeur est le résultat d'une mesure de similarité entre contextes strictement identique à celle présentée au chapitre 8 (cf. formule [2] du §2.2). La seule différence réside dans le fait de remplacer les UTs agrégées par des signatures thématiques. Cette mesure correspond à la fonction similaritéContextes() de la figure 10.6.
  Conformément aux principes présentés plus haut, l'UTL active correspond à l'UTL en construction associée au segment formé par la partie du texte s'étendant de la position du dernier changement d'état à la position courante de la fenêtre;
- le nombre de confirmations de l'état du segmenteur thématique. Cette variable, dénommée *nbConfirm* dans les figures 10.6 et 10.7, comptabilise pour les états #détectionNouveauThème et #détectionChangementDeThème le nombre de positions consécutives pendant lesquelles il y similarité entre le contexte de l'unité d'analyse courante et le contexte de l'UTL active. $S_{Confirm}$ représente la largeur maximale de l'intervalle correspondant à une phase de transition, c'est-à-dire la valeur de *nbConfirm* au delà de laquelle le segmenteur thématique passe à l'état #développementThème ou #changementDeThème.

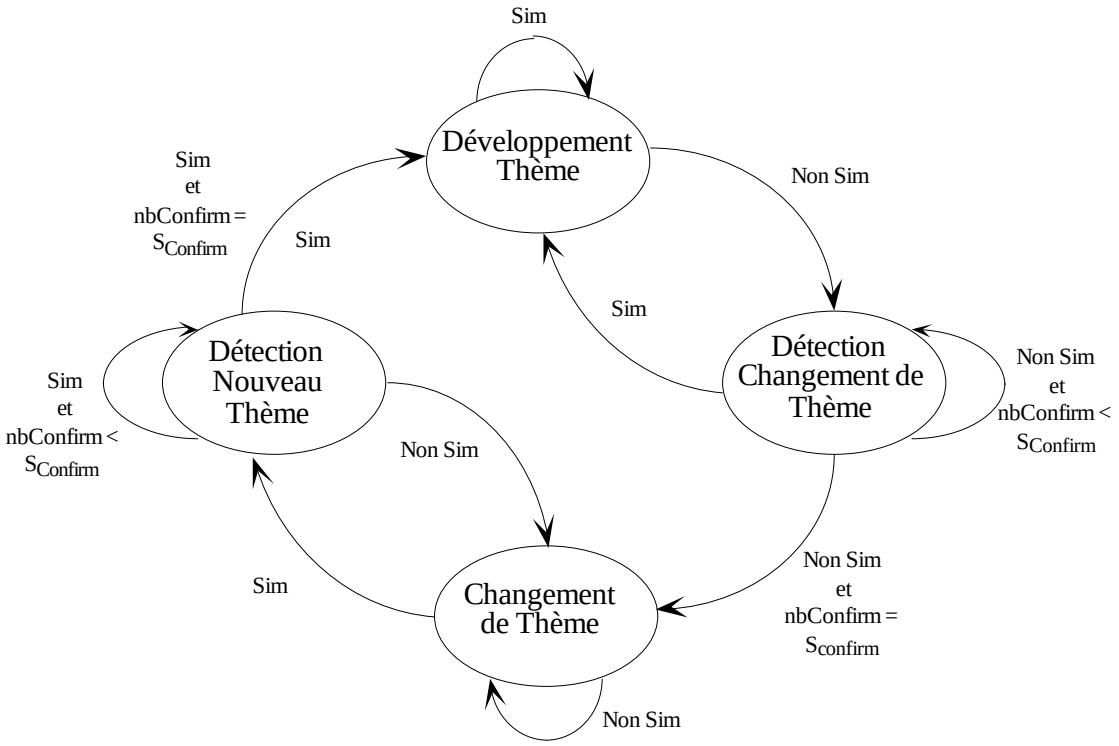
Après la construction du contexte de l'unité d'analyse courante, l'algorithme de segmentation se poursuit par l'évaluation de la similarité de ce contexte avec le contexte de l'UTL active, opération réalisée par la mesure de similarité évoquée plus haut. Le résultat de cette mesure est ensuite exploité conjointement avec l'état du segmenteur thématique ainsi que la valeur de la variable $nb_{Confirm}$ afin de déterminer la conduite à tenir du point de vue de la segmentation. Dans le cadre de cette détermination, le nouvel état du segmenteur est défini par l'automate de la figure 10.7.

```
positionCourante ← 2; nbConfirm ← 0; étatCourant ← #changementDeThème
UTLActive ← UTLVide; ListeUTLsEnConstruction ← ListeVide
contextePositionCourante ← activationSignatures(mémoireSEGAPSITH,fenêtre)
remplacementContexteUTLparContexte(UTLActive,contextePositionCourante)
ajoutMotaUTL(mot(1),UTLActive)
Tantque (positionCourante ≤ PositionFinTexte) faire
    contextePositionCourante ← activation(mémoireSEGAPSITH,fenêtre)
    Si (étatCourant = #détectionChangementThème) alors
        valSim ← similaritéContextes(contextePositionCourante,contexte(UTLenAttente))
    Sinon
        valSim ← similaritéContextes(contextePositionCourante,contexte(UTLActive))
    Fin_si
    Si (valSim > S_changementThème) alors
        Cas étatCourant
            #détectionChangementDeThème :
                nbConfirm ← 0
                majContenuUTLavecMots(UTLenAttente,mots(UTLActive))
                UTLActive ← UTLenAttente
            #changementDeThème :
                nbConfirm ← 1
                UTLenAttente ← UTLActive
                UTLActive ← UTLVide
            #détectionNouveauThème :
                Si (nbConfirm < S_Confirm) alors
                    nbConfirm ← nbConfirm + 1
                Sinon
                    Si (taille(UTLActive) ≥ T_minUTL) alors
                        ajoutUTLaListe(UTLenAttente,ListeUTLsEnConstruction)
                    Fin_si
                    UTLsim ← rechercheUTLsimilaire(UTLActive,ListeUTLsEnConstruction)
                    Si (UTLsim <> nil) alors
                        fusionUTLdansUTL(UTLActive,UTLsim)
                        UTLActive ← UTLsim
                    Sinon
                        ajoutUTLaListe(UTLActive,ListeUTLsEnConstruction)
                    Fin_si
                Fin_si
        Fin_cas
        majContexteUTLavecContexte(UTLActive,contextePositionCourante)
    Sinon
        Cas étatCourant
            #développementThème :
                nbConfirm ← 1
                UTLenAttente ← UTLActive
                UTLActive ← UTLVide
            #détectionNouveauThème :
                nbConfirm ← 0
                majContenuUTLavecMots(UTLenAttente,mots(UTLActive))
                UTLActive ← UTLenAttente
            #détectionChangementDeThème : nbConfirm ← nbConfirm + 1
        Fin_cas
        remplacementContexteUTLparContexte(UTLActive,contextePositionCourante)
    Fin_si
```

```
    ajoutMotaUTL(mot(positionCourante),UTLActive)
    étatCourant ← GestionÉtatSegmentation(étatCourant,indSim,nbConfirm)
    positionCourante ← positionCourante + 1
    déplacerFenêtreVersPosition(fenêtre,positionCourante)
Fin_tantque
```

Fig. 10.6. - Algorithme de segmentation de SEGAPSITH

Cet automate met en évidence la symétrie existant entre les deux états stables #changementDeThème et #développementThème ainsi qu'entre les deux états transitoires #détectionNouveauThème et #détectionChangementDeThème. Cette symétrie a pour origine le parti pris de considérer les états #changementDeThème comme de potentiels états de développement d'un thème inédit, c'est-à-dire d'un thème non représenté en mémoire par des signatures thématiques. Le déroulement habituel de la segmentation consiste à passer d'un thème à un autre, la phase de changement étant en elle-même assez ponctuelle. En général, le segmenteur thématique reste donc beaucoup plus longtemps dans l'état #développementThème que dans l'état #changementDeThème.



(Non)Sim : (non) similarité du contexte courant et du contexte sélectionné pour la position courante de la fenêtre

nbConfirm : nombre de positions de la fenêtre pour lesquelles la tendance actuelle (similarité ou non similarité des contextes) est vérifiée

$S_{confirm}$: nombre de confirmations de la tendance actuelle au-dessus duquel on change d'état

Fig. 10.7 - Automate de gestion des transitions entre états (définition de la fonction GestionÉtatSegmentation() de l'algorithme de la figure 10.6)

Il est néanmoins tout à fait possible que se retrouvant dans l'état #changement-DeThème initialement en raison de la fin du développement d'un thème, le segmenteur reste dans cet état pendant un nombre de positions important du fait de l'impossibilité de trouver une similarité suffisante entre le contexte de l'UTL active et le contexte associé aux différentes unités d'analyse occupant ces positions. Cette impossibilité traduit l'absence de cohérence entre elles de ces différentes unités du point de vue des signatures thématiques présentes en mémoire. Elle révèle également la présence probable d'un thème inédit.

Cette interprétation est liée à la façon dont est défini le contexte de l'UTL active lorsque le segmenteur est dans l'état #changementDeThème. Ce contexte n'a plus ici pour vocation de représenter un thème stable mais plutôt de constituer une mémoire à très court terme afin de détecter le plus rapidement possible un nouveau thème. C'est pourquoi il reprend le contexte de l'unité d'analyse occupant la position précédant la position courante (cf. fonction remplacementContexteUTLparContexte() de la figure 10.6). Le fait de ne pas trouver de similarité entre le contexte de l'UTL active et le contexte de l'unité d'analyse courante signifie donc que celle-ci n'est pas cohérente avec l'unité qui la précède.

La symétrie entre les états transitoires se traduit quant à elle par une manière commune de gérer le passage d'un état stable à un état stable différent du premier. L'idée guidant cette gestion est de séparer les structures associées à l'état stable initial, en l'occurrence une UTL et son contexte, de nouvelles structures spécifiquement dédiées à l'état transitoire considéré. De cette façon, en fonction du devenir de la transition, retour à l'état stable initial ou passage vers un autre état stable, on pourra fondre ces nouvelles structures dans les structures de l'état stable initial ou bien s'en servir comme point de départ pour bâtir celles du nouvel état stable.

En pratique, cette gestion donne lieu à la création d'une nouvelle UTL dès que le segmenteur thématique entre dans l'état #détectionChangementDeThème ou #détection-NouveauThème ainsi qu'au remplacement de l'UTL active par cette nouvelle UTL. L'ancienne UTL active est conservée dans une zone tampon (cf. variable *UTLenAttente* de la figure 10.6) jusqu'à ce que le résultat exact de la transition soit connu. On peut ainsi restaurer l'état exact du segmenteur en cas d'échec de la transition. Dans le même esprit, cette nouvelle UTL n'est initialement pas rattachée à la liste des UTLs en construction.

Durant la période où l'état transitoire considéré est actif, cette UTL est chargée de recueillir toutes les unités d'analyse traitées. Un contexte lui est bien entendu adjoint, comme pour toutes les UTLs en construction. Dans le cas d'un état #détection-NouveauThème, il assure pleinement les deux fonctions propres à un contexte d'UTL en construction : caractériser sur le long terme le thème de l'UTL et servir ainsi de point de référence pour décider si une nouvelle unité d'analyse doit lui être intégrée. Plus

précisément ici, la première fonction prépare le contexte de l'UTL qui pourrait être associée au prochain état #développementThème tandis que la seconde contribue à établir si cette transition aura effectivement lieu.

Dans le cas d'un état #détectionChangementDeThème, la présence d'un contexte est en revanche sans importance : d'une part, le contexte de l'état #changementDeThème qui est susceptible de suivre reprend directement celui des unités d'analyse au fur et à mesure de leur traitement mais ne les intègre pas; d'autre part, le contexte servant de point de référence pour le calcul de la similarité avec le contexte de l'unité d'analyse courante est le contexte de l'UTL située dans la zone tampon[1]. En effet, ce contexte constitue la représentation stable du thème vis-à-vis duquel les unités d'analyse les plus récentes semblent se démarquer. Or, tel ne serait justement pas le cas d'un contexte construit à partir des contextes de ces unités.

Le traitement de l'échec d'une transition vers un autre état stable est traité exactement de la même façon à partir d'un état #détectionNouveauThème ou d'un état #détection-ChangementThème. On restaure l'UTL active avant la transition à partir de la zone tampon et l'on ajoute au corps de cette UTL les unités d'analyse traitées dans l'espace de cette transition (cf. fonction majContenuUTLavecMots() de la figure 10.6). En revanche, on ne met pas à jour son contexte. L'opération serait inutile dans le cas d'un état #détectionNouveauThème puisque l'état restauré, #changementDeThème, ne nécessite pas de contexte cumulatif; elle serait même néfaste dans le cas d'un état #détectionChangementDeThème dans la mesure où l'on ne souhaite pas "contaminer" le contexte de l'UTL sortie de la zone tampon par des signatures thématiques non pertinentes, sélectionnées par un passage du texte apparemment non spécifique du thème courant.

Le traitement du basculement vers un autre état stable, lorsque la transition est menée à son terme, s'accompagne en revanche de quelques différences suivant que l'état destination est l'état #changementDeThème ou l'état #développementThème. Les attentes sont en effet différentes suivant que l'on se prépare à un développement de thème ou au contraire, à un changement de thème. Dans le premier cas, la perspective est de construire une nouvelle UTL ou bien éventuellement de compléter une UTL déjà existante s'il s'agit de la reprise d'un thème déjà abordé. Dans le second, les efforts sont dirigés vers la détection d'un nouveau thème. La création d'une nouvelle UTL, associée à un thème inédit, constitue un objectif secondaire, dont l'opportunité ne peut être reconnue qu'a posteriori, après avoir constaté que le changement de thème couvre un passage du texte suffisamment important.

[1] C'est l'objet du premier test de l'algorithme de la figure 10.6

De ce fait, le passage à l'état #changementDeThème ne se traduit par aucune opération particulière. On se contente de reprendre l'UTL active créée précédemment lors du passage à l'état #détectionChangementDeThème. En revanche, le passage à l'état #développementThème met en œuvre une procédure plus complexe. La première opération consiste à s'assurer qu'une UTL correspondant à un thème inédit n'a pas été produite lors du précédent état de type #changementDeThème. On examine pour ce faire comment la taille de l'UTL située dans la zone tampon, qui est en l'occurrence l'UTL associée au précédent changement de thème, se situe par rapport au seuil fixé a priori T_{minUTL}. Si elle le dépasse, on estime que le segment est suffisamment large pour correspondre à l'évocation d'un thème. L'UTL est alors mémorisée dans la liste des UTLs en construction, sans tenir compte de son contexte. Les UTLs en construction correspondant à des thèmes supposés inédits se caractérisent donc par leur absence de contexte. Si le seuil T_{minUTL} n'est pas dépassé, l'UTL est détruite. On perd par la même occasion les mots qu'elle contenait puisque l'on ne sait pas où les rattacher. Une autre solution pourrait être de les rattacher arbitrairement à l'UTL du segment précédent ou à la nouvelle UTL.

La deuxième opération liée au passage à l'état #développementThème est la recherche dans la liste des UTLs en construction d'une éventuelle UTL faisant référence au même thème que le nouveau segment. Cette recherche est réalisée en parcourant l'ensemble des UTLs en construction et en évaluant la similarité du contexte de chacun d'entre eux avec le contexte de l'UTL active. L'une de ces UTLs est sélectionnée à condition que cette valeur de similarité dépasse le seuil $S_{changementThème}$ de rattachement d'une unité d'analyse à une UTL en construction. Dans ce cas, l'UTL active, qui est l'UTL nouvellement créée lors du précédent état #détectionNouveauThème, est fusionnée dans l'UTL en construction trouvée (cf. fonction fusionUTLdansUTL() de la figure 10.6). Cette opération consiste à ajouter les mots de la première à la seconde et à fusionner les deux contextes suivant les principes appliqués pour la fusion du contexte d'une unité d'analyse et du contexte d'une UTL en construction. Si aucune UTL en construction n'a été sélectionnée au terme de la recherche, on continue à considérer comme UTL active l'UTL créée lors du précédent état #détectionNouveauThème.

Le dernier point que nous évoquerons concernant la segmentation thématique de SEGAPSITH concerne la forme exacte des UTLs produites et leur devenir après la segmentation. Bien que les UTLs produites par la segmentation de SEGAPSITH revêtent comme dans SEGCOHLEX la forme d'ensembles de mots pondérés, elles présentent la particularité d'associer à cette structure d'ensemble une structure en segments. Une UTL de SEGAPSITH peut être en effet formée à partir de plusieurs segments de texte non

adjacents. La structure ajoutée permet de conserver la trace de cette origine en identifiant chacun de ces segments par un groupe de mots spécifique. Pour le moment cependant, cette structure n'est pas exploitée par les opérations de mémorisation qui continuent, comme nous l'avons décrit dans la première section de ce chapitre, à considérer chaque UTL comme un seul ensemble de mots.

Ces mêmes opérations de mémorisation prennent en compte en revanche à la fois les mots dits inférés et le niveau de cohésion de chaque UTL. Compte tenu de l'utilisation que nous faisons du réseau de collocations, nous conservons pour l'essentiel les principes adoptés dans SEGCOHLEX concernant ces deux points. Les mots inférés sont donc toujours les mots sélectionnés à partir du réseau de collocations et présents dans au moins dans 75% des positions recouvertes par l'UTL. L'obtention du niveau de cohésion suppose quant à lui qu'une valeur de cohésion soit calculée pour chaque position de la fenêtre glissante, à l'instar de ce qui se fait dans SEGCOHLEX. Cette valeur est en l'occurrence identique à celle calculée dans SEGCOHLEX. Le niveau de cohésion de l'UTL est finalement donné par la moyenne des valeurs de cohésion associées à l'ensemble des positions recouvertes par cette UTL. Il est à noter qu'en raison de la méthode de segmentation, le niveau de cohésion d'une UTL est calculé ici après l'avoir formée alors que dans SEGCOHLEX, la valeur de ce niveau est un pré-requis à sa formation. Cette différence n'a toutefois pas d'influence sur la mémorisation : les UTLs n'ayant pas un niveau de cohésion suffisant, au sens défini par SEGCOHLEX, ne sont pas mémorisées.

En toute généralité, l'utilisation du réseau de collocations n'est pas obligatoire dans la méthode de segmentation de SEGAPSITH. Elle n'est supposée que suppléer l'absence de représentation des concepts. Il serait donc nécessaire de disposer d'un moyen autre que le réseau de collocations pour mettre en évidence des mots inférés ainsi que pour calculer un niveau de cohésion. Il est pour cela logique de se retourner en direction des signatures thématiques. Nous proposons ainsi de retenir comme mots inférés les mots les plus représentatifs des signatures thématiques sélectionnées pour l'ensemble des positions couvertes par l'UTL considérée. Dans cette optique, le niveau de cohésion d'une UTL pourrait être évalué sur la base de la moyenne de l'activité de ces mêmes signatures thématiques, en supposant que l'activité des signatures pour un passage de texte est un bon reflet du degré de cohérence de ce passage.

Plus précisément en ce qui concerne les mots inférés, la méthode pressentie consiste à recueillir pour chaque position l'ensemble des mots présents dans les N signatures formant le contexte de l'unité d'analyse courante. Pour une signature thématique, seuls les mots possédant un poids supérieur à un seuil fixé a priori sont retenus. Les mots

inférés d'une UTL sont alors formés de l'ensemble des mots ainsi pris en considération et présents dans une certaine proportion, par exemple 75%, des positions couvertes par l'UTL considérée.

L'évaluation du niveau de cohésion nécessiterait quant à elle de calculer pour chacune de ces positions un niveau de cohésion local égal à la moyenne des activités des N signatures constituant le contexte courant. Le niveau global de cohésion de cette UTL serait finalement donné par la moyenne de ces niveaux de cohésion locaux.

Pour finir, il faut préciser que la mémorisation des UTLs produites par la segmentation de SEGAPSITH est guidée, comme dans MLK, par les signatures thématiques sélectionnées tout au long de l'analyse et que l'on retrouve dans la forme finale du contexte qui leur est associé[1]. Suivant le processus décrit dans la première section de ce chapitre, chaque UTL est comparée aux signatures formant son contexte grâce à une mesure de similarité, ces signatures étant considérées selon l'ordre décroissant de leur poids. L'UTL est agrégée à la première signature avec laquelle la similarité dépasse le seuil fixé a priori pour déclencher l'agrégation. Si toutes les signatures sont passées en revue sans succès, l'UTL est mémorisée en créant une nouvelle signature.

2.2.3. Résultats

En l'état actuel du travail, l'évaluation de la méthode de segmentation thématique de SEGAPSITH n'est pas complète, en particulier du point de vue de son objectivation. Nous avons principalement réalisé des tests sur une dizaine de textes, à nouveau des dépêches d'agence de presse et des textes extraits de journaux, afin de mettre au point la méthode et ajuster ses différents paramètres. En dépit de leur caractère assez partiel, ces tests permettent néanmoins de mettre en évidence certaines caractéristiques intéressantes de la méthode, en particulier vis-à-vis de la méthode de segmentation de SEGCOHLEX. Nous donnons ici une illustration de ces différences sur l'exemple précis du texte de la figure 10.8.

Une jeune indienne remporte le titre de Miss Univers 1994

<*ST*> Un mannequin indien de 18 ans, Sushmita Sen, a créé la surprise samedi à Manille en remportant le titre de Miss Univers 1994, devançant deux beautés sud-américaines, Miss Colombie, Carolina Gomez Correa, et surtout Miss Venezuela, Minorka Mercado, qui faisait figure de favorite du concours.
La jeune Indienne, une beauté brune aux yeux noisettes de 1,75 mètre, est la première candidate de son pays à remporter ce titre. Elle succède à Miss Porto Rico, Dayanara Torres, 22 ans, qui lui a remis sa couronne devant une audience

[1] Cela ne concerne évidemment pas les UTLs supposées représenter un thème inédit puisqu'elles n'ont pas de contexte. Ces UTLs particulières sont mémorisées sous la forme de nouvelles signatures thématiques.

télévisée estimée à 600 millions de personnes à travers le monde. Parmi les six finalistes, figuraient également Miss Etats-Unis, Frances Louis Parker, Miss Philippines, Charlene Gonzales, et Miss République Slovaque, Silvia Lakatosova. Elles avaient été choisies parmi un groupe de dix demi-finalistes qui comprenait également les **représentantes** de l'Italie, de la Grèce, de la Suède et de la Suisse. *</ST>*

<ST> Quelques heures avant la cérémonie, un homme avait été tué par l'explosion d'un engin qu'il transportait à un kilomètre environ du Centre des Congrès où s'est tenu le concours de beauté, face à la baie de Manille. La police n'a pas pu établir immédiatement si cet incident avait un rapport avec le concours.

Jeudi, une bombe artisanale de faible puissance avait explosé dans une poubelle du Centre des **Congrès**, sans faire de dégâts. *</ST>*

<ST> La nouvelle Miss Univers, qui a remporté plus de 150.000 dollars de prix divers, a déclaré qu'elle se destinait au théâtre, à la publicité ou à l'écriture. Mais son voeu le plus cher, a-t-elle assuré, était de rencontrer Mère Teresa, parce qu'elle est "un exemple parfait d'une personne totalement dévouée, désintéressée et entière". *</ST>*

<ST> Alors que se déroulait l'élection, une centaine de féministes ont manifesté pacifiquement devant le Centre des Congrès, pour dénoncer le concours, affirmant qu'il servait à promouvoir le tourisme sexuel aux Philippines. *</ST>*

Fig. 10.8 - Exemple de dépêche de l'AFP traitée (AFP - mai 1994)

Dans ce texte, nous avons fait figurer entre les bornes *<ST>* et *</ST>* les quatre segments que nous avons intuitivement mis en évidence sur un critère thématique. Ces segments recouvrent trois thèmes différents : l'élection de la nouvelle Miss Univers, qui constitue le thème principal du texte, les attentats à la bombe qui ont accompagnés cette élection et enfin, la manifestation de protestation à l'encontre du concours. Une grande partie du texte est dédiée au premier thème, qui recouvre les segments 1 et 3. Il faut cependant noter qu'il n'est pas abordé de façon très spécifique. La partie du texte relatif à ce thème comporte beaucoup de noms propres (noms de pays et de personnes) et donne des indications sur les caractéristiques de la gagnante qui ne sont pas thématiquement très spécifiques.

Le deuxième thème, présent dans le deuxième segment, est abordé quant à lui de façon plus succincte mais également plus significative. On y trouve ainsi des mots ou des expressions comme "tué", "explosion", "engin", "police", "bombe artisanale", "de faible puissance", "explosé", "dégâts", qui sont véritablement spécifiques des attentats à la bombe. Le troisième et dernier thème enfin, qui est associé au quatrième segment, n'est évoqué que de façon très furtive puisque ce segment ne représente qu'une seule phrase du texte. Par ailleurs, il ne transparaît qu'au travers des seuls les mots "manifesté" et "dénoncer", le second n'étant pas d'ailleurs sans ambiguïté.

La figure 10.9 montre le résultat de l'application de la méthode de segmentation thématique de SEGCOHLEX sur le texte ci-dessus. Les valeurs utilisées pour les différents paramètres reprennent celles adoptées pour l'essentiel des tests du chapitre 9 :

taille de la fenêtre glissante égale à 19 mots, seuil minimum de 15 en fréquence et de 0,15 en cohésion pour les collocations prises en considération, taille de la fenêtre de lissage égale à 9 mots et enfin, nécessité pour un mot du réseau de collocations d'être lié à au moins 3 mots de la fenêtre glissante pour être retenu.

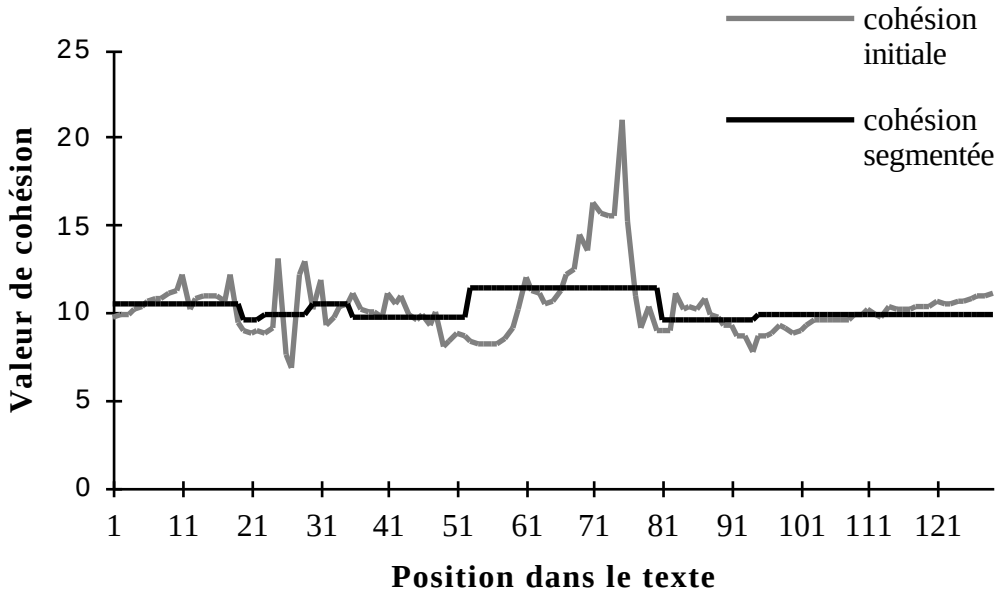


Fig. 10.9 - Résultat de l'application de la méthode de segmentation de SEGCOHLEX sur le texte de la figure 10.8

La figure met en évidence assez clairement deux points. Tout d'abord, la segmentation de SEGCOHLEX a une tendance naturelle à produire trop de segments, en particulier lorsque les thèmes n'apparaissent pas de façon très explicite comme c'est le cas ici pour le premier segment relatif au thème de l'élection de la nouvelle Miss Univers. On peut s'en apercevoir encore plus précisément à partir de la liste des segments (accompagnés chacun de leur valeur de cohésion) :

[1–12] : 10,45; [13–19] : 10,39; [20–22] : 9,58; [23–29] : 9,87; [30–35] : 10,39; [36–52] : 9,64; **[53–80] : 11,45;** [81–94] : 9,53; [95–129] : 9,87

On compte ainsi 6 segments sur un total de 9 pour la seule première moitié du texte, qui n'est pourtant sensée renvoyer qu'à un seul thème.

La seconde caractéristique notable de la segmentation de SEGCOHLEX est son manque de précision dans la délimitation des segments. Dans le cas du texte ci-dessus, le seul segment considéré comme significatif (mis en évidence en gras dans la liste des segments), c'est-à-dire sélectionné pour être ensuite mémorisé, correspond au deuxième thème abordé, celui des attentats à la bombe. Ce choix est assez compréhensible dans la

mesure où il s'agit du thème, parmi les trois identifiés dans le texte, qui est exprimé de la façon la plus spécifique.

En dépit de la pertinence de ce choix, la définition même du segment concerné est très approximative par rapport à la segmentation de référence présentée précédemment. Selon cette segmentation, ce segment devrait en effet s'étendre de la position 63 à la position 92 alors qu'il couvre en réalité les positions 53 à 80. Non seulement il débute nettement trop tôt mais il se finit trop rapidement. Ce dernier point pourrait être corrigé en lui adjoignant le segment suivant mais aucun élément ne permet de le décider puisque le niveau de cohésion de ce segment est même assez bas par rapport au niveau de cohésion des autres segments. L'UTL construite à partir du segment sélectionné ne peut donc qu'être assez bruitée et incomplète vis-à-vis du thème des attentats à la bombe.

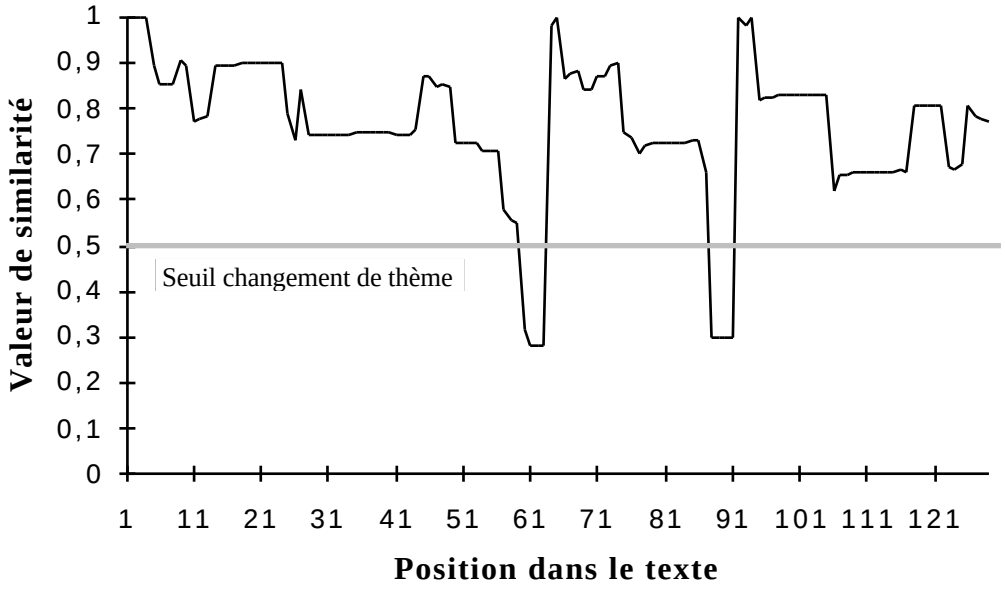


Fig. 10.10 - Résultat de l'application de la méthode de segmentation de SEGAPSITH sur le texte de la figure 10.8

La figure 10.10 illustre pour sa part le résultat de l'application de la segmentation thématique de SEGAPSITH, réalisée également sur le texte de la figure 10.8. La courbe dans ce cas ne rend pas compte d'une mesure de cohésion mais de la similarité entre le contexte de l'unité d'analyse située à la position précisée par l'axe des abscisses et du contexte de l'UTL active à ce moment donné de l'analyse. Les points de changement de thème correspondent donc là aussi aux minima les plus importants, comme ceux figurant aux positions 62 et 90 (ces positions sont marquées dans le texte de la figure 10.8 par une mise en gras des mots les occupant).

Les tests ont été réalisés avec les valeurs suivantes des paramètres propres à la méthode : taille des contextes (N) = 5; seuil de similarité pour la détection d'un changement de thème ($S_{changementThème}$) = 0,5; nombre de confirmations d'un état transitoire avant passage vers un autre état stable ($S_{Confirm}$) = 3; valeur du cœfficient d'atténuation β dans la fonction F de recalcul du poids des signatures communes lors de la fusion de deux contextes : 0,75; taille de la fenêtre glissante = 19 mots. Par ailleurs, en ce qui concerne les paramètres liés à l'utilisation du réseau de collocations (seuils minimaux de fréquence et de cohésion pour la prise en compte d'une collocation et nombre de liaisons nécessaires avec des mots de la fenêtre pour la sélection d'un mot du réseau de collocations), nous avons repris les mêmes valeurs que celles adoptées pour l'utilisation ci-dessus de SEGCOHLEX.

Précisons enfin que les signatures thématiques utilisées pour mener les tests sont celles élaborées à la suite de l'expérimentation relatée au paragraphe 1.4.2 de ce chapitre. Elles ne sont donc formées que de mots inférés mais présentent l'avantage d'une assez grande cohérence thématique. Afin de limiter le bruit inhérent aux signatures peu agrégées et aux mots de faible poids, nous n'avons pris en considération que les signatures résultant d'au moins 4 agrégations et les mots d'un poids supérieur ou égal à 0,1.

La figure 10.10 montre assez directement les avantages de la méthode de segmentation de SEGAPSITH par rapport à celle de SEGCOHLEX. On constate d'abord qu'elle engendre naturellement moins de segments que la précédente et que ceux-ci sont plus pertinents. Dans l'exemple ci-dessus, le nombre de segments n'est que de 3, à comparer aux 9 obtenus avec SEGCOHLEX. De plus, les deux premiers segments correspondent assez parfaitement aux deux premiers segments de la segmentation de référence. Seule la distinction entre les deux derniers segments échappe à la méthode. Il en est d'ailleurs de même pour la première méthode. Il faut souligner néanmoins que la petite taille de ce dernier segment et le peu de mots spécifiques de son thème à comparer au nombre de mots liés au troisième segment rend cette distinction quasi-impossible pour une segmentation n'intervenant qu'au niveau lexical.

Le nombre de segments produit par l'analyse de SEGAPSITH dépend bien entendu de la valeur donnée au seuil $S_{changementThème}$ de détermination des changements de thèmes. Ce seuil présente l'avantage d'offrir un contrôle plus direct sur la finesse de la segmentation que celui offert par SEGCOHLEX au travers de la taille de la fenêtre glissante et de la taille de la fenêtre de lissage. On constate par ailleurs qu'avec une valeur médiane ne résultant pas d'un choix pointu, on obtient un nombre de segments tout à fait proche de la valeur souhaitable. Cette robustesse résulte sans aucun doute du fait que les principaux changements de thèmes sont en général très nettement marqués comme on peut le voir sur la figure 10.10 pour les positions 62 et 90. Les minimaux de la courbe de

similarité qui leur correspondent sont suffisamment profonds et étroits pour qu'ils puissent être repérés avec une gamme de valeurs de seuil assez large.

Le second avantage de la méthode d'analyse de SEGAPSITH que laisse apparaître la figure 10.10 est sa plus grande précision dans la délimitation des segments : elle place le début du deuxième segment à la position 62 alors qu'il commence à la position 63 et le début du troisième segment à la position 90 alors qu'il débute à la position 93. L'étroitesse des minimaux correspondant aux changements de thème contribue à cette plus grande précision mais on peut penser plus globalement que la sélection des domaines est plus sensible et réagit plus rapidement aux changements que celle des mots du réseau de collocations.

Afin d'illustrer cette sensibilité de la sélection des domaines, nous allons observer ce qui se passe à ce niveau à différents moments représentatifs de l'analyse. Le tableau de la figure 10.11 montre pour 7 positions dans le texte de la figure 10.8 à la fois le contexte de l'UTL active (les deux colonnes de gauche) et le contexte de l'unité d'analyse courante (les deux colonnes de droite). Pour chaque contexte, on donne sa liste de signatures thématiques accompagnées de leur poids. Chaque signature est représentée par un titre généré automatiquement en associant ses deux mots de plus fort poids (ce qui peut expliquer des similitudes de nom entre deux signatures différentes mais thématiquement proches (cf. signatures demi-finale/tournoi)). Pour chaque position, un en-tête précise le numéro de la position, la valeur de la similarité entre les deux contextes considérés ainsi que l'état dans lequel se trouve le segmenteur thématique à cette position.

Ce tableau permet d'abord de constater que les évolutions de la similarité entre contextes observées au niveau de la figure 10.10 sont bien sous-tendues par un changement du contenu des contextes. Ce changement se manifeste aussi bien lors du premier changement de thème, où l'on passe de signatures à dominante sportive à des signatures plutôt relatives à des événements guerriers, que lors du second changement de thème où le passage s'effectue dans l'autre sens : des signatures à dominante guerrière, on revient à des signatures à dominante sportive.

32	0,743		#développementThème	
demi-finale/tournoi		109,45	demi-finale/tournoi	4,57
championnatMonde/championMonde		65,94	majoritéAbsoudre/majoritéAbsolu	3,00
demi-finale/tournoi		63,56	championnatMonde/championMonde	2,71
charge_social/préretraite		31,20	demi-finale/tournoi	2,64
rectorat/pédagogique		4,33	skipper/mille	1,84
60	0,317		#détectionChangementDeThème	
demi-finale/tournoi		191,13	fusillade/blessé	2,41
demi-finale/tournoi		113,01	peuplement/couvre-feu	2,37
championnatMonde/championMonde		104,62	demi-finale/tournoi	2,00
charge_social/préretraite		40,06	guérillero/caserne	1,78
rectorat/pédagogique		4,33	bombarder/artillerie	1,53
64	0,98		#détectionNouveauThème	
peuplement/couvre-feu		2,58	peuplement/couvre-feu	2,58
fusillade/blessé		2,53	fusillade/blessé	2,52
guérillero/caserne		2,06	guérillero/caserne	2,06
demi-finale/tournoi		2,06	demi-finale/tournoi	2,03
arme_lourde/belligérant		1,84	arme_lourde/belligérant	1,84
72	0,87		#développementThème	
fusillade/blessé		19,61	fusillade/blessé	3,66
peuplement/couvre-feu		19,26	peuplement/couvre-feu	3,32
demi-finale/tournoi		14,96	guérillero/caserne	2,42
guérillero/caserne		14,74	demi-finale/tournoi	2,12
arme_lourde/belligérant		4,60	bombarder/artillerie	1,98
88	0,30		#détectionChangementDeThème	
fusillade/blessé		50,56	demi-finale/tournoi	3,49
peuplement/couvre-feu		45,86	argent_frais/provisionné	2,99
demi-finale/tournoi		38,53	championnatMonde/championMonde	2,07
guérillero/caserne		22,04	demi-finale/tournoi	2,05
arme_lourde/belligérant		4,60	taux_d'escompte/billet_vert	1,60
93	0,98		#détectionNouveauThème	
demi-finale/tournoi		6,11	demi-finale/tournoi	3,49
argent_frais/provisionné		5,28	argent_frais/provisionné	2,99
championnatMonde/championMonde		3,64	championnatMonde/championMonde	2,08
taux_d'escompte/billet_vert		3,56	demi-finale/tournoi	2,03
demi-finale/tournoi		3,56	taux_d'escompte/billet_vert	1,97
100	0,83		#développementThème	
demi-finale/tournoi		167,92	demi-finale/tournoi	3,58
demi-finale/tournoi		99,13	argent_frais/provisionné	2,98
championnatMonde/championMonde		93,05	demi-finale/tournoi	2,12
charge_social/préretraite		40,06	championnatMonde/championMonde	2,10
argent_frais/provisionné		20,94	taux_d'escompte/billet_vert	1,66

Fig. 10.11 - Évolution du contexte de l'UTL active et du contexte de l'unité d'analyse courante lors de la segmentation du texte de la figure 10.8

Ce retour correspond en l'occurrence au segment 3, qui est une reprise du thème initial développé dans le premier segment. On peut d'ailleurs juger facilement de la similitude existant entre le contexte de l'UTL active à la position 32, ou même 60,

(segment 1) et ce même contexte aux positions 93 et 100 (segment 3). C'est cette similitude qui permet d'ailleurs de réunir les deux segments dans une même UTL, possibilité que n'offre pas la segmentation de SEGCOHLEX (cf. phase de recherche d'une UTL en construction similaire lors du passage à l'état #développementThème).

L'autre constatation importante que permet de faire la figure 10.11 est que la présence de signatures thématiques très spécifiques par rapport aux thèmes abordés par les textes n'est pas systématiquement indispensable. Ceci est un point très important du point de vue de la capacité de généralisation de la méthode, c'est-à-dire de sa capacité à traiter des textes situés en dehors du corpus d'apprentissage des signatures thématiques. Dans le cas concret du texte de la figure 10.8, le thème principal est l'élection d'une reine de beauté, événement bien trop singulier dans le mois de mai 1994 des dépêches de l'AFP pour avoir donné lieu à la création d'une signature thématique[1], ou tout du moins d'une signature significative (même avec le seuil placé très bas ici d'au moins 4 agrégations).

On voit cependant au niveau des contextes qu'une représentation stable de ce thème parvient à se construire avec les signatures thématiques disponibles. En l'occurrence, le fait que cette élection soit une compétition a conduit à sélectionner des signatures relatives à des compétitions sportives pour l'incarner. À défaut de pouvoir être représenté globalement, ce thème a donc été représenté par l'intermédiaire d'une de ses dimensions. Bien entendu, cette stratégie de remplacement ne peut pas toujours être un succès. Si le deuxième thème important du texte considéré avait été lié à une compétition sportive au lieu d'être relatif aux attentats à la bombe, qui est de plus un thème assez éloigné de l'élection d'une reine de beauté, la séparation des deux thèmes n'aurait pu être réalisée. L'utilisation de signatures seulement liées à un thème de façon partielle pour le représenter constitue malgré tout une possibilité intéressante offerte par cette méthode pour traiter des textes dont le thème dépasse a priori le champ d'action défini par ses connaissances.

Pour achever la présentation de ces premiers résultats, nous évoquerons rapidement deux points que nous n'avons qu'effleurés pour le moment : l'influence des différents paramètres de la méthode et l'évaluation de la méthode sur une tâche de redécouverte de frontières de texte. L'étude du premier point se heurte, comme usuellement dans un tel cas, au problème de l'interdépendance des différents paramètres. En toute généralité, ce problème demanderait l'élaboration d'un plan d'expérimentation rigoureux permettant de contrôler soigneusement les paramètres les uns par rapport aux autres.

---

[1] On a vu par ailleurs avec la figure 10.9 que les segments correspondant au thème de l'élection de Miss Univers sont considérées comme non significatifs par la segmentation de SEGCOHLEX, qui est en l'occurrence celle ayant permis de construire les signatures utilisées ici.

En l'absence d'une expérimentation de cette nature, nous nous limiterons à souligner que parmi les paramètres nous semblant les plus influents, la taille des contextes arrive assez nettement en tête. Cette taille ne doit pas être trop petite afin que l'assise de la représentation d'un thème soit suffisante et évite ainsi que de petits changements dans le classement des signatures n'entraîne des variations trop importantes du contenu des contextes sur un petit intervalle de positions. Elle ne doit pas non plus être trop grande si l'on ne veut pas que la similarité soit altérée par le changement d'une position à une autre de signatures non significatives. Ce point dépend bien entendu de la définition précise donnée à la mesure de similarité.

En revanche, pour une taille de contexte donnée, la nature de la mesure de similarité ne semble pas affecter de façon très sensible les résultats. En dehors de la mesure proposée ici, nous avons également appliqué trois autres mesures de similarité (une fondée uniquement sur les poids des signatures communes, une autre fondée seulement sur leur différence de rang et une dernière similaire à celle utilisée entre les UTLs et les signatures au §1 de ce chapitre) en obtenant des résultats comparables. La fonction *F* intervenant dans la fusion des contextes n'est pas non plus un paramètre très sensible : les tests menés avec $\beta = 1$ ou prenant la moyenne des deux poids n'ont pas conduit à des résultats significativement différents.

On constate en revanche que la présence des mots apportés par le réseau de collocations lors de l'activation des signatures thématiques contribue assez nettement à rendre les minima de similarité plus marqués, donc plus précis et plus faciles à détecter.

En dehors de l'évaluation qualitative dont nous avons fait état ci-dessus, nous sommes également intéressé, comme on a pu le voir au chapitre 9, par une évaluation plus quantitative des méthodes de segmentation développées. Dans un premier temps, l'objectif consiste à reprendre la tâche de redécouverte des frontières de texte présentée au chapitre 9, en attendant de constituer des données de référence fiables (recoupement de jugements humains), par ailleurs plus adaptées au type de segmentation que nous mettons en œuvre ici.

Nous avons donc testé la segmentation de SEGAPSITH sur le même ensemble de textes que SEGCOHLEX en utilisant exactement le même protocole[1]. Nous avons constaté à cette occasion que la politique de pondération des minima utilisée par Hearst pour mettre en évidence les frontières de texte, politique qui posait problème pour SEGCOHLEX, est tout à fait inadaptée dans le cas présent. On ne peut pas s'appuyer en pratique sur la profondeur des minima de similarité entre contextes pour juger de

[1] Les valeurs utilisées pour les paramètres de la méthode sont les mêmes que précédemment, à l'exception de $S_{changementThème}$, porté à 0,6 afin de faire apparaître un peu plus de changements de thèmes.

l'importance des changements de thème. En conséquence, seul les résultats obtenus pour l'ensemble des minima peuvent être prise en compte : pour les 47 changements de thème détectés (sur 38 frontières recherchées)[1], on obtient une précision de 48,9% et un rappel de 60,5%. Ces chiffres sont identiques à ce que nous obtenions avec SEGCOHLEX pour le même nombre de minima.

Cette relative contre-performance a selon nous deux origines. Tout d'abord, les tests de mise au point de la segmentation de SEGAPSITH n'ont pas pu être poussés aussi loin que ceux de SEGCOHLEX. L'évaluation n'a donc été réalisée qu'à titre indicatif et provisoire. Par ailleurs, le corpus utilisé pour cette évaluation est constitué d'une suite entrelacée de textes relatifs au domaine de la justice et au domaine des accidents ferroviaires. Or, si certaines des signatures construites à partir des dépêches de l'AFP sont bien en liaison avec le domaine de la justice, aucune ne fait référence au domaine ferroviaire. Cette insuffisance se trouve renforcée par le fait que certains des textes traitant d'un accident de train font usage d'un vocabulaire également utilisé dans le domaine de la justice (investigation de la police, ouverture d'une enquête, ...). L'absence de connaissances spécifiques sur le domaine ferroviaire, conjuguée à ces recouvrements, contribue donc assez largement à expliquer les performances obtenues.

3. Les mécanismes d'amorçage inter-niveau

3.1. Aspects généraux

Bien qu'en l'état actuel du travail, l'amorçage dans ANTHAPSI en soit davantage au stade des principes que de la réalisation concrète, il est déjà possible de mettre en avant des spécifications assez précises de la façon dont il pourrait être réalisé au sein de ROSA, entre SEGCOHLEX et SEGAPSITH, ainsi qu'entre ROSA et MLK[2]. Au niveau le plus général, cet amorçage inter-niveau est sous-tendu par deux grands principes. Le premier d'entre eux, déjà implicitement présent au travers de la figure 3.2 du chapitre 3 (cf. flèches doubles), stipule que l'amorçage n'intervient que sous la forme d'échanges directs entre processus. Le processus amorcé n'exploite donc pas les représentations construites par le processus amorceur mais plutôt des informations que celui-ci lui transmet explicitement. Ces informations ont une signification fonctionnelle en rapport avec la tâche à réaliser. Il découle de ce principe que l'amorçage au sein d'ANTHAPSI ne peut intervenir qu'entre les processus de segmentation thématique des différents niveaux.

[1] Le nombre total de minima est encore une fois moins important que dans le cas de SEGCOHLEX (64) bien que le seuil pour la détection des changement de thème soit plus haut que précédemment.

[2] Lorsque nous faisons allusion à l'amorçage de MLK par ROSA nous faisons plus spécifiquement référence à l'amorçage de MLK par SEGAPSITH.

Les processus d'apprentissage sont en effet directement dépendant de la forme des représentations de textes manipulées au niveau où ils opèrent.

Dans le cas concret de l'amorçage de MLK par SEGAPSITH par exemple, le segmenteur thématique de SEGAPSITH communique ainsi avec celui de MLK non pas en lui fournissant les représentations de texte qu'il produit sous la forme d'ensembles d'UTLs mais de façon plus générique en lui transmettant seulement les positions du texte à l'endroit desquelles il a détecté un changement de thème. La communication entre processus amorceur et amorcé repose donc sur l'utilisation exclusive des notions propres à la tâche, comme c'est le cas ici des changements de thème. Cette façon de faire permet de faire abstraction des différences de niveau de représentation pouvant exister entre le processus amorceur et le processus amorcé.

La capacité ainsi offerte est particulièrement indispensable dans un amorçage tel que celui intervenant entre ROSA et MLK dans la mesure où la différence de niveau entre les représentations manipulées par l'un et par l'autre est très importante : le premier ne travaille qu'à partir de regroupements non structurés de mots tandis que le second opère en utilisant des ensembles hautement structurés de concepts. Une communication par l'intermédiaire des représentations construites demanderait la mise en place d'un mécanisme spécifique de mise en correspondance, démarche à la fois plus lourde et moins générique que la communication directe des processus, lorsque celle-ci est possible.

Le second principe sur lequel se fonde l'amorçage dans ANTHAPSI concerne la façon dont ce mécanisme prend place dans le fonctionnement du processus qui se trouve amorcé. Dans ce que nous avons exposé jusqu'à présent, nous avons toujours abordé les processus de ce type, que ce soit la segmentation thématique de MLK ou celle de SEGAPSITH, de manière identique par rapport aux autres processus du même niveau, c'est-à-dire en les considérant comme autonome vis-à-vis des composantes d'ANTHAPSI situées en dehors de son propre niveau. C'est ainsi que la segmentation thématique de MLK peut fonctionner indépendamment de celle de SEGAPSITH et que parallèlement, cette dernière peut fonctionner indépendamment de la segmentation de SEGCOHLEX.

En revanche, chacun des processus susceptibles d'être amorcés possède les moyens de reconnaître les cas où il se trouve mis en défaut en raison d'une insuffisance de ses connaissances. La notion d'UT inédite dans la segmentation de MLK et celle de changement de thème prolongé dans celle de SEGAPSITH en sont la manifestation. Cette forme de réflexivité permet donc aux processus concernés de savoir quand faire appel à des compétences externes, même s'ils ont été conçus afin de conserver une capacité de

fonctionnement minimale dans ce type de situation. Ces compétences externes prennent ici la forme du processus de même nature existant au niveau immédiatement inférieur.

La conjugaison d'un mécanisme d'auto-diagnostic de chaque segmentation thématique et du recours à la segmentation du niveau inférieur en cas de détection d'une insuffisance au niveau considéré forment le cœur de l'amorçage inter-niveau que l'on souhaite mettre en œuvre.

À côté de ces deux principes généraux, se pose le problème concret du mode de fonctionnement de deux segmentations en interaction potentielle. Sur un plan général, la segmentation thématique n'est pas en effet un processus que l'on peut solliciter sur commande pour savoir ponctuellement s'il existe un changement de thème à une position donnée d'un texte. Elle nécessite de prendre en compte un large contexte textuel et doit donc être réalisée sur l'ensemble d'un texte, ou tout du moins une grande partie de celui-ci, pour être capable de prodiguer un avis ponctuel.

Du point de vue de l'amorçage, cette contrainte est assez gênante dans la mesure où elle impose en pratique de réaliser deux analyses complètes, à deux niveaux différents, au lieu d'une seule. Il semble en effet difficile de prévoir à l'avance si la segmentation du niveau N sera capable ou non de traiter l'intégralité du texte considéré et donc, si elle aura ou non besoin de la segmentation du niveau N-1. Dans le doute, on accomplit les deux analyses, ce qui est assez coûteux.

La principale solution envisageable pour faire face à ce surcoût consiste à réserver principalement l'amorçage aux situations dans lesquelles les connaissances nécessaires à la caractérisation de la plus grande partie des thèmes du texte sont absentes. Il est en effet sans doute plus facile d'estimer avec un faible coût que les thèmes principaux d'un texte ne sont globalement pas représentés dans les connaissances disponibles (l'analyse du tout début du texte suffit souvent à s'en rendre compte) que de faire la même estimation pour un thème précis, de plus non central. Dans ce dernier cas, on devra donc se reposer sur les capacités de gestion des thèmes inédits proposées par chacune des méthodes de segmentation, en espérant que le texte ne comporte pas plusieurs thèmes de ce genre. Ceux-ci pourraient en effet ne pas être distingués les uns des autres.

La solution proposée ci-dessus présente la caractéristique d'être très proche de la solution consistant, dans le cas où les thèmes principaux d'un texte ne sont pas représentés au niveau N, à faire uniquement appel à la segmentation du niveau N-1. L'amorçage se résume alors au choix exclusif d'une méthode pour chaque texte en fonction des connaissances disponibles mais ne cherche pas à établir une interaction entre deux méthodes à un niveau plus fin, comme celui d'un passage de texte. Cette définition

de l'amorçage nous paraît cependant un peu restrictive et en dépit de son efficacité, nous examinerons dans ce qui suit comment une collaboration plus étroite peut être établie.

3.2. Amorçage de SEGAPSITH par SEGCOHLEX

Bien que SEGCOHLEX et SEGAPSITH soient assez proches l'un de l'autre quant au niveau des connaissances et des représentations qu'ils manipulent, le rapport entre leurs deux mécanismes de segmentation mérite une attention toute particulière étant donné que ceux-ci ont des modes de fonctionnement assez différents. La segmentation de SEGAPSITH est capable de détecter les changements de thème au fur et à mesure de l'analyse tandis que celle de SEGCOHLEX opère de façon plus globale et nécessite donc de traiter tout le texte avant de savoir où sont placés les changements de thème. Dans le cas où l'on fait appel à l'amorçage entre les deux, il faut donc segmenter d'abord le texte dans son intégralité avec SEGCOHLEX avant de lancer l'analyse de SEGAPSITH.

Sur le plan pratique, le recours aux services de SEGCOHLEX n'engendre pas de surcoût notable pour la segmentation de SEGAPSITH lorsque celle-ci fait appel au réseau de collocations. Les opérations de sélection des mots du réseau et de calcul d'une valeur de cohésion pour chaque position (uniquement pour la définition d'un niveau de cohésion des segments dans SEGAPSITH) sont en effet communes et auront donc déjà été réalisées par l'analyse de SEGCOHLEX lorsque celle de SEGAPSITH sera lancée. Le seul problème pour les textes un peu longs réside dans la taille représentée par toutes ces données.

L'amorçage en lui-même intervient à partir du moment où le segmenteur thématique de SEGAPSITH passe dans l'état #changementDeThème. Plus précisément, il ne se déclenche qu'à partir du moment où le segmenteur est resté dans cet état pendant un nombre de positions supérieur à T_{minUTL}, le seuil à partir duquel on considère que le segment ainsi délimité possède suffisamment de mots pour donner lieu à la création d'une UTL. Pour chaque position au delà de ce seuil, on examine donc si la segmentation de SEGCOHLEX fait apparaître un changement de thème. Si tel est le cas, l'UTL active est mémorisée dans la liste des UTLs en construction et une nouvelle UTL vide est prise comme UTL active. Le segmenteur conserve quant à lui le même état et le processus de déclenchement de l'amorçage est réinitialisé.

Bien que nous ne l'ayons pas implanté pour le moment, ce mécanisme ne présente pas de difficulté particulière. Il nécessite seulement une modification mineure de l'algorithme de segmentation de SEGAPSITH au niveau du traitement des états #changementDe-Thème.

3.3. Amorçage de MLK par SEGAPSITH

Contrairement à l'amorçage de SEGAPSITH par SEGCOHLEX, celui de MLK par SEGAPSITH n'est pas confronté à une différence de mode de fonctionnement des segmentations thématiques. Dans les deux cas, les changements de thème sont détectés au fur et à mesure de l'analyse et les deux processus sont donc menés en parallèle. En revanche, il est caractérisé par une différence très importante du niveau des représentations manipulées. Bien que nous ne sommes concerné ici que par la dimension thématique, cette différence possède tout de même une influence dans la mise en correspondance des analyses réalisées par l'une et l'autre des méthodes de segmentation dans la mesure où les unités élémentaires d'analyse ne sont pas les mêmes : des mots dans le cas de SEGAPSITH et des graphes conceptuels représentant des propositions dans celui de MLK.

Le mécanisme de l'amorçage en lui-même reprend pour l'essentiel celui décrit au paragraphe 3.2. Il est mis en éveil dès qu'une proposition est affectée à l'UT inédite courante. Il n'est toutefois mis en action que si une affectation similaire se répète pour un certain nombre de propositions venant à sa suite (le nombre de ces propositions est un paramètre à fixer). Dès lors que ce seuil est franchi, le traitement de chaque nouvelle proposition, si elle possède un contexte non significatif, déclenche la consultation de l'analyse de SEGAPSITH afin de déterminer si elle détecte la présence d'un changement de thème dans l'espace correspondant à la proposition. Dans l'affirmative, une nouvelle UT en construction est créée, ajoutée à la liste des UTs en construction et la proposition considérée y est rattachée. Il en sera de même des propositions qui suivront si elles ont également un contexte non significatif et ce, jusqu'au prochain changement de thème détecté par SEGAPSITH ou jusqu'à atteindre une proposition dotée d'un contexte significatif.

Le problème de la différence des niveaux de représentation intervient plus spécifiquement lors de la recherche d'un éventuel changement de thème dans l'espace du texte correspondant à la proposition traitée, ou plus à exactement sa représentation sémantique. Cette mise en relation suppose en fait qu'un lien de filiation ait été conservé entre la forme de surface du texte, sa représentation syntaxique et sa représentation sémantique. Nous n'avons pas cependant abordé ce point plus en détail étant donné qu'il sort un peu du champ de notre travail.

Sur un plan plus général, l'amorçage de MLK par SEGAPSITH pose davantage de problèmes de réalisation que celui de SEGAPSITH par SEGCOHLEX. Néanmoins, en supposant que la segmentation de MLK soit implémentée et que les étapes permettant de

passer de la forme de surface des textes à leur représentation sémantique soient précisées, il semble tout à fait possible de mettre en œuvre cet amorçage de façon effective dès à présent. Bien entendu, il ne permettra pas de créer des UTs très structurées mais celles-ci devraient tout de même suffire à alimenter le processus de segmentation de MLK.

4. Discussion

Le premier point à souligner est que SEGAPSITH, aussi bien au travers de l'extraction des signatures thématiques que de sa segmentation thématique, a fait l'objet d'une implémentation complète. Les seuls points n'ayant pas été concrétisés concernent la communication de SEGAPSITH avec les autres composantes d'ANTHAPSI, c'est-à-dire son amorçage de MLK et son amorçage par SEGCOHLEX. Cette implémentation a été réalisée comme précédemment en Smalltalk. Elle s'appuie en partie sur des outils développés pour SEGCOHLEX. C'est le cas bien évidemment du gestionnaire de réseau de collocations mais également de l'outil de test de la segmentation thématique de SEGCOHLEX (cf. annexe H), que nous avons pu reprendre pour la segmentation de SEGAPSITH en y ajoutant des possibilités de suivi des contextes. Un outil spécifique a par ailleurs été développé pour le test de la construction des signatures thématiques. Il permet notamment d'inspecter facilement le contenu de la mémoire des signatures, son activité, les signatures en elles-mêmes et de déclencher les calculs de similarité ainsi que les agrégations voulus. On pourra en avoir un aperçu à l'annexe J.

Le deuxième point que l'on peut mettre en avant est que le travail réalisé dans le cadre de SEGAPSITH, qui est le reflet à un niveau plus directement opérationnel de MLK, a permis de valider certaines hypothèses de ce dernier. C'est le cas en particulier de la possibilité de faire émerger des représentations stables de situations prototypiques par agrégation de segments de textes similaires évoquant ces situations. Cette validation ne concerne à strictement parler que SEGAPSITH et sa transposition au niveau de MLK reste assurément spéculative. Elle constitue néanmoins une indication intéressante en même temps qu'un terrain d'expérimentation tout à fait nécessaire à l'affinement des idées.

Cette validation devra d'ailleurs être poussée plus loin en direction de l'amorçage. L'amorçage intra-niveau n'a ainsi pas véritablement été validé en tant que démarche reposant sur l'interdépendance entre l'analyse des textes et l'apprentissage de connaissances dans la mesure où les UTLs utilisées dans les expériences de construction de signatures thématiques ont été entièrement produites par SEGCOHLEX et non par la segmentation de SEGAPSITH. L'enrichissement de signatures déjà existantes ou la

création de nouvelles signatures réalisés à partir d'UTLs produites par la segmentation de SEGAPSITH n'ont donc pas été illustrés.

On peut considérer en revanche que ces expériences rendent compte d'une phase initiale caractérisée par l'absence de toute connaissance pragmatique sur le domaine abordé. À ce titre, elles offrent un premier aperçu de l'amorçage inter-niveau. Celui-ci reste toutefois très partiel, notamment parce que le mécanisme présenté au paragraphe 3 de ce chapitre n'a pas été implémenté. Dans les expérimentations réalisées, la phase de transition entre un fonctionnement fondé uniquement sur les capacités d'analyse du niveau précédent et un fonctionnement s'appuyant sur le processus d'analyse du niveau considéré se limite donc au passage brusque d'un mode de fonctionnement à l'autre. Cette commutation brutale, qui n'est déjà pas une démarche très naturelle à l'échelle d'un domaine spécifique, ne peut être appliquée globalement. Elle irait en effet à l'encontre de bon nombre des principes développés ici, qui mettent plutôt en avant l'incrémentalité de l'apprentissage. L'amorçage inter-niveau mis en œuvre ici devra donc être profondément étendu.

Le troisième et dernier objet de discussion à propos de SEGAPSITH concerne ses extensions possibles. Les propositions dans ce sens ont d'abord trait à l'extraction des signatures thématiques. L'essentiel des extensions relatives à la structure de la mémoire épisodique de MLK et au mécanisme d'apprentissage qui lui est associé peuvent en effet être reprises et appliquées aux signatures thématiques de SEGAPSITH. Nous citerons à cet égard : l'auto-adaptation du seuil contrôlant la détection de la similarité entre une UTL et une signature en fonction des caractéristiques globales de la mémoire des signatures; l'adaptation de ce même seuil, mais au niveau local de chaque signature, selon le niveau de stabilité de cette dernière; la hiérarchisation des signatures thématiques; le découpage éventuel des signatures trop hétérogènes, et enfin, la détection de la stabilité d'une signature. Pour obtenir davantage de précision sur chacune de ces extensions envisagées, on pourra se reporter au paragraphe 5.3 du chapitre 6. Les solutions préconisées dans le cadre de MLK sont en effet transposables dans le cadre de SEGAPSITH.

Les extensions de la segmentation thématique de SEGAPSITH dépendent d'abord des extensions liées à l'extraction des signatures thématiques. Par exemple, si les signatures sont organisées de manière hiérarchique, il sera sans aucun doute dans l'intérêt de la segmentation d'en tenir compte. La mise en évidence d'un niveau de stabilité pour chaque signature est également un facteur potentiellement exploitable par la segmentation thématique. Nous n'avons pas cependant réfléchi de façon plus approfondie sur ces points car ils dépendent de la forme exacte que prendront les extensions touchant les signatures.

En ce qui concerne la segmentation à proprement parler, il nous semble que la principale amélioration doit intervenir au niveau de la sélection des signatures thématiques. Nous utilisons à l'heure actuelle le réseau de collocations afin d'augmenter le nombre de mots liés au thème courant mais cette utilisation reste assez rudimentaire. En particulier, elle a tendance à faire apparaître trop de bruit, ce qui, à côté de l'impact sur la pertinence des signatures activées, présente surtout l'inconvénient de pénaliser lourdement les performances. Il est donc nécessaire de limiter le nombre des mots ainsi sélectionnés en éliminant le plus grand nombre de ceux que l'on peut assimiler à du bruit. Le nombre de liens qu'un mot du réseau de collocations doit entretenir avec des mots de la fenêtre glissante pour être sélectionné est un premier moyen d'action. Son utilisation est cependant délicate : il permet d'éliminer effectivement une grande partie du bruit mais il conduit aussi à supprimer une part importante des mots intéressants.

Une autre possibilité consiste à s'appuyer sur la fréquence de réapparition des mots, comme c'est le cas pour la sélection des mots inférés des UTLs. La seule différence est qu'ici, cette sélection devra être réalisée au fur et à mesure de l'analyse, et non a posteriori. Dans cette optique, ne seraient utilisés pour activer les signatures que les mots du réseau de collocations apparus suffisamment fréquemment depuis le dernier changement de thème. Cela suppose bien évidemment de prévoir des périodes de mise en place après les changements de thème étant donné que les tendances obtenues sur une très courte période ne sont pas significatives.

Récapitulatif

Ce chapitre nous a permis de présenter SEGAPSITH, la partie d'ANTHAPSI qui est l'équivalent de MLK à un niveau où l'unité de représentation est le mot et non le concept. Cette équivalence a à la fois pour vocation de valider à un niveau pleinement opérationnel les principes de MLK et d'offrir simultanément les moyens d'amorcer ce dernier sur le plan thématique. À l'image de MLK, SEGAPSITH se divise en deux grandes composantes : la mémoire des signatures thématiques et le segmenteur thématique de textes.

La première met en œuvre l'apprentissage réalisé par SEGAPSITH et a pour objectif, à ce titre, de faire émerger des signatures thématiques. Celles-ci sont des représentations de thèmes prenant la forme d'ensembles de mots pondérés. Leur mode de formation est en tout point identique à celui des UTs agrégées. Il s'appuie à la base sur des représentations de texte composées d'un ensemble d'Unités Thématiques Lexicales (UTLs). Chacune de ces UTLs constitue la représentation d'un thème au sein du texte considéré. Elle est formée elle-même de deux ensembles de mots pondérés : un ensemble de mots provenant

du texte, appelés mots du texte, et un second ensemble de mots du réseau de collocations sélectionnés à l'occasion de la segmentation du texte, appelés mots inférés. Chaque signature est le résultat de l'agrégation d'un ensemble d'UTLs appartenant à des textes différents et ayant été jugées similaires. Les signatures conservent en outre la structure des UTLs en deux ensembles de mots de provenances différentes.

L'agrégation d'une UTL à une signature est équivalente à la simple fusion de deux listes de mots pondérés. La procédure globale de mémorisation d'une UTL est la même que celle définie pour les UTs dans MLK : en supposant qu'un ensemble de signatures supposées proches de l'UTL considérée ait été constitué (plus ou moins directement en faisant appel à l'activation de la mémoire), on évalue la similarité entre l'UTL à mémoriser et chacune des signatures composant cet ensemble. Si la valeur obtenue est supérieure à un seuil fixé a priori, l'UTL est agrégée à la signature concernée. Si au contraire, ce seuil n'est franchi pour aucune des signatures en présence, l'UTL est mémorisée en tant que nouvelle signature.

La méthode de construction des signatures thématiques a été testée sur un vaste ensemble de textes, en l'occurrences 5949 dépêches de l'AFP. Les résultats obtenus montrent l'intérêt de la segmentation des textes et de la prise en compte des mots inférés pour la construction de signatures à la fois plus stables (absence de changement de la partie significative de leur contenu), plus représentatives (signatures de fort poids) et plus homogènes sur le plan thématique. Ils rendent compte également d'une certaine robustesse de la méthode utilisée vis-à-vis des effets de séquence résultant de l'ordre de traitement des textes. Enfin, nous avons pu montrer que considérer la formation des signatures thématiques en laissant de côté les mots des textes, c'est-à-dire sous l'angle de la structuration d'un réseau de collocations, permet d'obtenir de meilleures signatures au sens des critères énoncés ci-dessus.

La seconde composante d'ANTHAPSI, la segmentation thématique, est chargée de produire les représentations de texte que nous avons décrites précédemment comme des ensembles d'UTLs. Elle s'appuie pour ce faire sur les signatures thématiques présentes dans la mémoire de SEGAPSITH. Elle reprend elle aussi une part importante des principes développés pour la composante équivalente de MLK. Ceux-ci consistent pour l'essentiel à décider du rattachement d'une unité d'analyse (une proposition dans le cas de MLK ou un mot dans celui de SEGAPSITH) à une Unité Thématique en construction sur la base de la similitude de leurs contextes respectifs. Le contexte d'une unité d'analyse ou d'une UT en construction est un ensemble d'UTs agrégées (au sens générique) de la mémoire, sélectionnées comme étant représentatives du thème évoqué par cette UT ou cette unité d'analyse.

En pratique, la segmentation suit le processus suivant : l'unité d'analyse courante ainsi que ses unités voisines (à la fois en amont et en aval) servent de point de départ à l'activation de la mémoire des signatures. Un contexte composé des signatures les plus activées est ainsi défini pour cette unité d'analyse. Par ailleurs, le thème en cours de développement est figuré par l'une des UTLs en construction gérées par le segmenteur thématique. Cette UTL est appelée UTL active. L'opération de base de la segmentation consiste à évaluer si l'unité d'analyse courante se rattache ou non à l'UTL active. Ce rattachement repose sur le calcul d'une mesure de similarité entre le contexte de l'UTL active et celui de l'unité d'analyse courante. Si la similarité dépasse un seuil fixé a priori, le rattachement est décidé et le segmenteur reste dans un état de développement de thème. Ce rattachement se traduit notamment par la fusion du contexte de l'UTL active avec celui de l'unité d'analyse courante.

Si au contraire, la similarité n'est pas suffisante, le segmenteur entre dans un état qui, s'il est confirmé par les unités d'analyse suivantes, conduira à un état de changement de thème et donc, à un changement de l'UTL active. Celle-ci prendra la forme d'une autre UT en construction déjà existante si le contexte des unités d'analyse venant à la suite est trouvé similaire au contexte de l'une de ces UTs. On note au passage que les UTLs formées peuvent donc regrouper des segments de texte non connexes. En revanche, l'absence de retour d'un thème déjà abordé entraînera la création d'une nouvelle UT en construction qui reprendra alors le rôle d'UTL active.

En l'état actuel du travail, l'évaluation de cette méthode de segmentation n'a été réalisée que de façon très partielle. Au vu des expérimentations effectuées, on peut dire cependant que sur le plan qualitatif, l'utilisation des signatures comme source de connaissances semble apporter, par rapport à la segmentation de SEGCOHLEX, une précision plus importante ainsi qu'une sûreté accrue des changements de thème mis en évidence. Sur un plan plus quantitatif, en l'occurrence l'expérience de redécouverte des frontières de texte présentée au chapitre 9, les premiers résultats sont comparables à ceux produits par la segmentation de SEGCOHLEX. De meilleurs résultats devraient toutefois pouvoir être obtenus avec des signatures plus proches des thèmes abordés par le corpus de référence.

Du fait de sa position centrale au sein d'ANTHAPSI, SEGAPSITH est au cœur des processus d'amorçage inter-niveau. Il est en effet amorcé par SEGCOHLEX et doit lui-même amorcer MLK. La réalisation actuelle de SEGAPSITH ne tient pas encore compte de cette dimension mais les principes de cet amorçage sont néanmoins fixés : celui-ci intervient directement entre les processus de segmentation thématique et prend la forme d'une transmission d'informations sur les positions des changements de thème. Ces informations sont exploitées par le segmenteur amorcé lorsqu'il détecte la présence d'un thème ou d'une situation inédite. Elles lui permettent alors de déterminer si le

segment de texte concerné par cette détection recouvre éventuellement plusieurs thèmes ou plusieurs situations inédites successives. Ce principe est applicable aussi bien lorsque le segmenteur ne dispose d'aucune connaissance sur les thèmes principaux d'un texte que lorsque ces manques ne touchent qu'un thème secondaire.

Le seul test que nous ayons réalisé pour le moment relatif à l'amorçage s'inscrit dans une perspective beaucoup plus rudimentaire que les principes exposés. La segmentation de SEGAPSITH a en effet été testée entièrement à l'aide de signatures construites en faisant appel à la segmentation de SEGCOHLEX. Cette expérience est équivalente en fait à un amorçage dans lequel on solliciterait un processus de segmentation ou à un autre en fonction des connaissances disponibles au sein des différents niveaux, sans que les processus en question n'interagissent le moins du monde. La réalisation d'un amorçage plus élaboré ne devrait cependant pas se heurter à des difficultés trop importantes, au moins entre SEGCOHLEX et SEGAPSITH, dans la mesure où tous les éléments nécessaires sont déjà réunis.

Conclusion et perspectives

1. Conclusion

1.1. Synthèse

Au travers du travail que nous avons présenté, nous avons cherché à apporter une solution au problème de l'apprentissage automatique de connaissances pragmatiques à partir de textes, plus précisément de connaissances sur les situations prototypiques du monde. Ainsi que nous l'avons mis en évidence dans le chapitre 1, cet objectif se heurte à la très forte interdépendance présente entre la compréhension de textes et l'apprentissage de ce type de connaissances : mettre en évidence les connaissances sur les situations véhiculées par les textes suppose l'existence de capacités minimales de compréhension ; or ces dernières se fondent en règle générale sur les connaissances pragmatiques que nous souhaitons mettre à jour. Cette interdépendance explique que les travaux réalisés jusqu'à présent dans ce domaine (cf. chapitre 2) se soient majoritairement orientés vers la spécialisation de connaissances générales déjà possédées par les systèmes considérés.

Afin de nous affranchir de cette interdépendance et nous orienter ainsi vers un apprentissage moins tributaire de connaissances fournies a priori, nous avons proposé de mettre en œuvre un amorçage en deux phases. Pour un domaine de connaissances donné, par exemple le domaine ferroviaire, celui de l'édition ou encore celui de la guerre, la première phase intervient lorsque les connaissances relatives à ce domaine sont inexistantes ou trop peu nombreuses et trop incertaines. Dans une telle situation, les mécanismes d'analyse des textes exploitant ces connaissances sont bien entendu dans l'impossibilité d'opérer.

Pour surmonter cette difficulté inhérente au début du développement de la représentation d'un domaine de connaissances, la première phase de l'amorçage exploite le fait qu'un type d'analyse de texte peut être mené à différents niveaux. Pour un niveau N de structuration et de précision des connaissances, elle fait plus précisément appel aux capacités d'analyse opérant à un niveau de connaissances de plus faible degré de structuration et de précision afin de produire une représentation des textes permettant d'alimenter le processus d'apprentissage du niveau N. Ce processus peut alors donner naissance à des connaissances spécifiques de ce niveau N, connaissances qui pourront ensuite être utilisées par le mécanisme d'analyse également propre au niveau N.

La première phase d'amorçage décrite ci-dessus, appelée phase d'amorçage inter-niveau, est complétée par une seconde, dite phase d'amorçage intra-niveau. Du fait de leur mode de formation, les premières connaissances relatives à un domaine sont nécessairement incomplètes et imprécises. Il est donc indispensable de poursuivre leur développement afin de pallier ces faiblesses. C'est l'objet de l'amorçage intra-niveau. Celui-ci repose sur un principe stipulant que le processus d'analyse des textes à un niveau donné repose sur les connaissances produites par le processus d'apprentissage de ce même niveau. La liaison entre ces deux composantes est réalisée par une mémoire permettant de stocker ces connaissances tout en y assurant un accès compatible avec les exigences des processus qui les utilisent. Compte tenu du principe posé ci-dessus, le processus d'analyse des textes doit donc être capable d'opérer à partir de connaissances incomplètes et imprécises. Pour sa part, le processus d'apprentissage doit être de nature incrémentale et non supervisée.

En vue de concrétiser ces principes, de les expérimenter et de les valider, nous avons conçu le système ANTHAPSI (ANalyse THématique et APprentissage de SItuations). Celui-ci est plus précisément le résultat de l'application des principes présentés précédemment au problème de l'apprentissage automatique de connaissances pragmatiques à partir de textes. L'amorçage inter-niveau y est illustré en faisant intervenir trois niveaux de connaissances pragmatiques : dans SEGCOHLEX, celles-ci sont représentées de façon implicite au travers d'un réseau de cooccurrences lexicales ; dans SEGAPSITH, une représentation explicite des thèmes existe mais elle est peu structurée, prenant la forme d'un regroupement de mots pondérés en fonction de leur importance ; dans MLK enfin, des situations sont représentées à la fois de façon précise, les briques élémentaires de leur représentation étant des concepts, et de façon structurée, ces concepts faisant partie de graphes conceptuels, eux-mêmes organisés au sein de la représentation d'une situation en fonction de leur rôle vis-à-vis de celle-ci et de leurs relations avec les autres graphes qui la composent.

L'amorçage inter-niveau intervient donc entre SEGCOHLEX et SEGAPSITH ainsi qu'entre SEGAPSITH et MLK. Il s'effectue plus précisément entre les processus d'analyse de texte propres à ces trois composantes. Compte tenu du type de connaissances que nous avons pour objectif d'apprendre ici, ces processus relèvent de l'analyse thématique des textes. Suivant la composante considérée, celle-ci est réalisée avec un niveau de précision différent.

Ne pouvant travailler qu'à partir des mots, SEGCOHLEX et SEGAPSITH sont tributaires de la présence explicite dans les textes de mots représentatifs des thèmes abordés. La sensibilité de SEGAPSITH est cependant plus grande que celle de SEGCOHLEX du fait de la représentation explicite des thèmes qu'il possède. Celle-ci

permet en effet de reconnaître plus facilement un thème à partir d'une petite configuration de mots. En opérant à partir d'une représentation sémantique des propositions des textes, MLK s'affranchit des ambiguïtés de sens présentes au niveau lexical et s'appuie sur une représentation structurée : textes découpés en propositions avec détermination du rôle de chacun des concepts qui les constituent. Ces deux facteurs permettent une mise en correspondance plus aisée avec les connaissances qu'il détient sur les situations prototypiques, connaissances elles-mêmes dotées de ce même niveau de précision et de structuration.

L'analyse thématique de SEGAPSITH ainsi que celle de MLK ont en outre la caractéristique commune de savoir détecter qu'un passage de texte fait référence à un thème sur lequel ils ne possèdent pas de connaissances. L'amorçage inter-niveau intervient dans de telles situations. Lorsque l'analyse thématique de MLK décèle un thème inédit du point de vue de ses connaissances, elle se repose sur celle de SEGAPSITH. Dans une situation similaire, l'analyse thématique de SEGAPSITH fait appel pour sa part à celle de SEGCOHLEX. Lors du traitement de textes relatifs à un nouveau domaine, l'analyse d'un texte à un niveau *N* peut ainsi reposer entièrement sur les moyens offerts par le niveau *N-1*. À mesure que les connaissances du niveau *N* se développent concernant ce domaine, l'analyse du même niveau prend de plus en plus d'importance par rapport à celle du niveau *N-1*, jusqu'à devenir autonome lorsque ce développement a été suffisamment important.

La poursuite de ce même développement est du ressort de l'amorçage intra-niveau. Celui-ci est mis en œuvre de façon identique au niveau de SEGAPSITH et de MLK, l'objectif étant de montrer que des principes similaires peuvent s'appliquer à des niveaux différents de précision et de structuration des connaissances. SEGCOHLEX est de ce point de vue un peu à part : il est supposé constituer l'amorce initiale permettant d'initier l'amorçage inter-niveau. Son fonctionnement est de ce fait différent de celui des autres composantes d'ANTHAPSI : sa tâche est de réaliser la segmentation thématique des textes en se fondant sur une source de connaissances constituée une fois pour toutes. Il s'agit donc d'un module autonome que l'on pourrait utiliser hors d'ANTHAPSI. C'est pourquoi il a été réuni avec SEGAPSITH au sein d'un même ensemble, ROSA, que l'on peut considérer comme auto-amorcé.

L'amorçage intra-niveau s'organise quant à lui autour d'une mémoire abritant les connaissances pragmatiques en formation. Compte tenu de son contexte d'usage, cette mémoire ne possède pas de structure a priori. Le rappel des connaissances qu'elle contient s'y effectue donc de manière associative. On utilise pour ce faire un mécanisme

d'activation de la mémoire, plus ou moins complexe suivant que l'on se trouve dans MLK ou dans SEGAPSITH.

Le mécanisme d'apprentissage présidant au développement proprement dit des connaissances se définit comme suit. L'analyse thématique produit pour chaque texte un ensemble d'Unités Thématiques (UTs) représentant chacune la façon dont une situation est évoquée dans ce texte. À chacune de ces UTs est associée un ensemble d'UTs agrégées de la mémoire sélectionnées lors de l'analyse comme étant les plus proches de l'UT considérée. Une UT agrégée est un agrégat d'UTs similaires possédant une structure mettant en évidence les éléments communs de ces UTs et pondérant plus généralement tous leurs constituants en fonction de la fréquence de leur présence au sein des UTs regroupées. Pour chaque UT issue de l'analyse thématique, on évalue la similarité entre cette UT et chacune des UTs agrégées issues de la mémoire. Cette évaluation s'effectue suivant l'ordre décroissant du niveau d'activité des UTs agrégées et s'arrête lorsqu'une similarité suffisante a été trouvée ou bien lorsque la liste des UTs agrégées sélectionnées a été épuisée. Dans le premier cas, la nouvelle UT est agrégée à l'UT agrégée similaire. Dans le second, elle conduit à la création en mémoire d'une nouvelle UT agrégée.

L'apprentissage ainsi réalisé est de fait non supervisé et incrémental. Le contenu des UTs agrégées évolue au fur et à mesure de l'agrégation de nouvelles UTs en faisant apparaître de façon progressive leurs traits les plus caractéristiques, en l'occurrence évalués comme étant leurs traits les plus récurrents. Lorsque les UTs agrégées sont considérées comme stables, c'est-à-dire dotées d'un noyau de constituants possédant un poids suffisamment élevé et n'évoluant plus de façon significative, elles peuvent être abstraites suivant la procédure décrite au chapitre 7.

La troisième composante de l'amorçage intra-niveau, l'analyse thématique, qui prend principalement ici la forme d'une segmentation thématique des textes, repose aussi sur des principes similaires dans SEGAPSITH et dans MLK, même si quelques différences existent entre les deux. Dans les deux cas, l'idée de base consiste à utiliser les UTs agrégées présentes en mémoire afin de caractériser à la fois l'unité d'analyse courante, que ce soit la représentation sémantique d'une proposition dans MLK ou une suite de mots dans SEGAPSITH, et les Unités Thématiques en cours de construction, représentant les situations identifiées dans la partie du texte déjà traité. Cet ensemble d'UTs agrégées est appelé *contexte*, aussi bien pour l'unité courante d'analyse que pour les UTs en construction. Dans le premier cas, ces UTs agrégées sont sélectionnées par l'intermédiaire du mécanisme de rappel associé à la mémoire. Dans le second, le contexte est constitué à partir de la fusion des contextes adjoints aux différentes unités d'analyse ayant formé l'UT en construction considérée.

La compatibilité d'une unité d'analyse et d'une UT en construction, donc la possibilité de rattacher la première à la seconde, est définie par une mesure de similarité entre leurs deux contextes. La détection des changements de thème s'effectue donc en comparant le contexte de l'unité d'analyse courante avec le contexte de l'UT en construction à laquelle ont été rattachées les unités d'analyse précédentes.

On obtient ainsi un mécanisme de segmentation thématique des textes capable d'utiliser des connaissances, les UTs agrégées, n'obéissant pas aux contraintes habituelles de précision et de complétude, contraintes que satisfont par essence difficilement des connaissances en construction progressive comme le sont les UTs agrégées. Cette capacité est au cœur de l'amorçage intra-niveau dans la mesure où elle conduit à une amélioration progressive de l'analyse des textes, laquelle induit à son tour une amélioration progressive des connaissances qui sont construites par le processus d'apprentissage, et ainsi de suite.

1.2. Bilan

Lorsque nous avons indiqué au début de la synthèse ci-dessus que notre objectif est d'apprendre de manière automatique des connaissances pragmatiques à partir de textes, nous avons omis de spécifier le niveau de précision et de structuration de ces connaissances. Implicitement, il est bien entendu souhaitable d'obtenir les connaissances les plus précises et les plus structurées possible. À cet égard, le niveau définit par MLK, en particulier pour ce qui est des schémas abstraits, constitue une référence supérieure. ROSA représente au contraire un niveau de référence minimum, plus exactement un point de départ. L'amorçage inter-niveau se fixe comme objectif de passer progressivement de l'un à l'autre. Mais il est évident que ce passage ne peut être direct. ANTHAPSI fait apparaître les extrêmes de ce processus mais il reste à définir un ensemble de niveaux intermédiaires caractérisés par un accroissement progressif de la précision et de la structuration des connaissances manipulées ainsi que des représentations de texte construites.

Il est également évident que combler le fossé séparant ROSA de MLK est un travail de grande ampleur dépassant largement le cadre d'une thèse. Le fait de nous limiter à ces deux composantes nous a conduit à les investir de deux missions de nature pratique, s'ajoutant à leur rôle au sein d'ANTHAPSI vis-à-vis de l'amorçage inter-niveau. En raison du niveau élevé des connaissances qu'il utilise et des processus qu'il met en œuvre, MLK nous a ainsi permis de spécifier précisément les mécanismes régissant l'amorçage intra-niveau et par là même, la façon dont est réalisé l'apprentissage au sein

d'ANTHAPSI. Nous avons vu cependant que le degré d'élaboration des représentations manipulées à ce niveau interdit pour le moment une expérimentation sur une large échelle, l'intervention humaine restant nécessaire à de nombreuses occasions.

Au contraire, le caractère pleinement opérationnel de ROSA nous a offert la possibilité d'expérimenter ces principes assez largement et dans une certaine mesure, de les valider. Cette validation reste très partielle puisque les représentations manipulées à ce niveau sont peu élaborées et de ce fait, assez éloignées de celles présentes dans MLK. Les résultats obtenus dans le cadre de ROSA ne présument donc pas de ce que l'on pourrait obtenir dans MLK. Au mieux donnent-ils une indication intéressante sur ce qu'ils pourraient y être.

L'ensemble des points abordés ci-dessus définissent un cadre permettant de dresser un bilan circonstancié du travail réalisé. Nous commencerons celui-ci par l'amorçage intra-niveau. Au niveau de MLK, notre objectif n'était pas, comme nous l'avons dit plus haut, de valider la démarche mais de la concrétiser suffisamment afin de l'illustrer concrètement sur quelques exemples. Ceci a été réalisé en ce qui concerne les structures de la mémoire épisodique et le mécanisme de mémorisation, donc d'apprentissage des UTs agrégées. En revanche, le bilan est plus contrasté pour le mécanisme de rappel et l'analyse thématique. Dans les deux cas, des spécifications très précises existent et ont donné lieu à une conception détaillée. L'implémentation reste cependant à accomplir. Précisons, à propos du mécanisme de rappel, que des expérimentations dans un environnement de simulation ont tout de même permis de valider les choix réalisés pour la phase de sélection, ce qui constituait le point le plus délicat et le plus incertain du rappel. Plus généralement à propos de MLK, il serait sans doute nécessaire de développer un plus grand nombre d'exemples, non pas tant pour tendre vers une validation quantitative mais plutôt afin d'explorer qualitativement différents cas de figure possibles.

Au niveau de ROSA, et plus précisément de SEGAPSITH, l'objectif était à la base plus ambitieux et allait davantage dans le sens d'une évaluation objective des résultats. Au stade actuel, SEGAPSITH a été réalisé dans son intégralité mais son évaluation n'est encore que partielle. La raison en est double. Tout d'abord, nous ne disposons pas d'un cadre d'évaluation pré-établi (méthodologie, corpus de référence, outils de mesure, etc.), cadre qui est très lourd à mettre en place. Comme nous l'avons vu aux chapitres 8 et 10 néanmoins, nous pourrions nous rattacher à une action d'évaluation institutionnelle proche, telle que Topic Detection and Tracking (TDT)[1]. Cela nous obligerait à travailler sur la langue anglaise alors que nous n'avons travaillé jusqu'à présent que sur la langue

[1] Sur un plan purement matériel, ce rapprochement est plutôt à classer parmi les perspectives à la fois en raison de l'ampleur du travail qu'il représente et du fait que cette action d'évaluation n'a été ouverte que très récemment (1998 a vu le lancement de la première campagne publique).

française. Cette transposition est toutefois envisageable puisque les outils que nous utilisons sont également applicables à la langue anglaise.

La seconde raison du caractère partiel de l'évaluation actuelle tient à ce qu'il est intrinsèquement difficile d'évaluer SEGAPSITH indépendamment de SEGCOHLEX du fait de l'amorçage du premier par le second. L'analyse thématique de SEGAPSITH ne peut fonctionner sans celle de SEGCOHLEX dans le cas où elle ne dispose pas des signatures thématiques relatives au domaine abordé. Or, cet amorçage a été spécifié mais pas encore implémenté. Les tests que nous avons menés n'ont donc pas concerné SEGAPSITH dans son ensemble mais ont été réalisés en découpant le processus initial en plusieurs étapes.

La première de ces étapes a consisté à construire des signatures thématiques à partir des résultats de la segmentation thématique de SEGCOHLEX. On a considéré en première approximation que SEGAPSITH ne possédait pas les signatures nécessaires au traitement des différents domaines abordés et que son analyse thématique reposait dès lors entièrement sur celle de SEGCOHLEX. En réalité, compte tenu de la taille du corpus de test (5949 dépêches de l'AFP) et de sa relative homogénéité thématique (les dépêches ne couvrent qu'un seul mois), on observe que des signatures atteignent un niveau de stabilité élevé au cours de ce processus de construction et pourraient être utilisées dès ce moment-là par l'analyse thématique de SEGAPSITH.

La deuxième étape a permis pour sa part de tester l'analyse thématique de SEGAPSITH en utilisant les signatures construites à la suite de l'étape précédente. Le prolongement logique de cette démarche voudrait qu'une troisième étape conduise à tester la coopération directe entre apprentissage et analyse thématique en veillant à ce que les résultats de l'analyse thématique de SEGAPSITH soient mémorisés afin de compléter et d'étendre les signatures déjà présentes. Cette troisième étape n'a pas été réalisée pour le moment, notamment du fait de la nécessité d'optimiser les performances de la segmentation de SEGAPSITH[1]. Compte tenu de la quantité de travail relativement modérée que représente la mise en œuvre de l'amorçage entre SEGCOHLEX et SEGAPSITH, il sera plus intéressant à l'avenir de tester l'ensemble de ROSA plutôt de s'intéresser à cette troisième étape.

[1] Nous avons pu appliquer l'algorithme de segmentation de SEGAPSITH sur une cinquantaine de dépêches AFP concaténées mais les temps de traitement (de l'ordre de 2 à 3 heures) sont trop importants pour que l'on puisse traiter des corpus aussi importants que dans SEGCOHLEX. Il est possible de réduire ce temps de traitement de façon importante en limitant le nombre de mots provenant du réseau de cooccurrences lexicales. Les performances obtenues sont cependant moins bonnes. Précisons d'autre part que cette analyse a été implémentée à la manière d'un prototype et que de nombreuses optimisations sont encore possibles à ce niveau pour atteindre des temps de traitement plus raisonnables.

Le second point principal du bilan est constitué par l'amorçage inter-niveau. Compte tenu de l'absence d'implémentation de l'analyse thématique de MLK, il est évident que cet amorçage n'a pas été mis en œuvre entre ROSA et MLK, même s'il a été spécifié assez précisément. Quoi qu'il en soit, même si cet amorçage est envisageable en l'état[1], il ne se justifie pas sur le long terme. Si l'on suppose en effet l'existence d'un ensemble de niveaux intermédiaires entre ROSA et MLK, l'amorçage interviendra non pas directement de ROSA vers MLK mais de ROSA vers le premier de ces niveaux intermédiaires et du dernier de ceux-ci vers MLK.

La situation est différente en ce qui concerne l'amorçage de SEGAPSITH par SEGCOHLEX puisque dans ANTHAPSI, la seconde composante est spécifiquement dédiée à l'amorçage de la première. Comme nous l'avons indiqué ci-dessus, celui-ci n'a pas encore été implémenté bien qu'ayant été spécifié suffisamment finement pour l'être assez directement. L'expérimentation menée à propos de la construction de signatures thématiques à partir des dépêches de l'AFP a néanmoins contribué à valider la démarche. Elle montre en effet que des segments de texte comparables à ce que produit l'analyse thématique de SEGAPSITH, ces derniers étant même généralement plus précis, peuvent être agrégés afin de construire une représentation homogène d'un ensemble de thèmes. Par ailleurs, nous avons également montré que ces représentations sont utilisables afin de mettre en œuvre une segmentation thématique plus précise que celle de SEGCOHLEX.

2. Perspectives

La première des perspectives n'en ait pas vraiment une puisqu'elle consiste à achever l'implémentation et le test d'ANTHAPSI dans ses spécifications actuelles. De ce cadre, le point sans doute le plus délicat est constitué par l'analyse thématique de MLK. Bien que celle-ci ait été spécifiée de façon détaillée, elle demandera en effet un travail de mise au point important, travail rendu particulièrement difficile par la lourdeur de l'environnement de test à mettre en place : nécessité de modéliser les connaissances conceptuelles utilisées et de construire manuellement un nombre suffisant de représentations de textes pour former des UTs agrégées exploitables par cette analyse thématique.

Le deuxième volet des perspectives s'inscrit dans un terme un peu plus long. Il vise à étudier et à concrétiser les différentes extensions proposées lors de l'exposé des composantes d'ANTHAPSI. Nous ne reprendrons pas ici l'ensemble de ces

[1] La segmentation thématique de SEGAPSITH peut tout à fait fournir des indications utiles à celle de MLK sur la position des changements de thème en dépit de la différence de niveau des connaissances manipulées et des représentations construites.

propositions. Pour plus de détails, nous renvoyons le lecteur à la dernière partie de chacun des chapitres concernés. Nous nous contenterons de souligner l'intérêt et l'importance de la hiérarchisation des Unités Thématiques agrégées. Cette extension présente en effet la particularité d'être commune à la fois à MLK et à SEGAPSITH et d'être l'évolution ayant le plus grand impact général puisque touchant à l'organisation et à la forme des connaissances au sein des mémoires abritant les différentes formes d'UTs agrégées.

Le dernier volet des perspectives envisageables se place quant à lui dans un terme encore plus lointain dans la mesure où il s'intéresse à l'extension de la démarche proposée pour l'analyse thématique des textes et l'apprentissage de connaissances pragmatiques à l'ensemble de la compréhension de textes et de l'apprentissage de connaissances à partir de textes. L'analyse thématique n'est en effet qu'une des nombreuses composantes d'un processus de compréhension de textes. L'essentiel des processus impliqués dans l'analyse des textes narratifs, auxquels nous nous intéressons plus spécifiquement ici, relèvent plus précisément des cinq composantes suivantes :

- la composante conceptuelle. Cette composante définit les briques de base des représentations construites, ces briques étant des concepts[1] à leur stade le plus avancé. Elle recouvre également un ensemble de connaissances rendant compte des propriétés générales des objets et des actions du monde représentées par ces briques de base. Sur le plan des processus, la composante conceptuelle se concrétise par l'analyse sémantique, permettant de passer d'une représentation morpho-syntaxique d'un texte à une représentation sémantique où les mots sont remplacés par des concepts ou par leurs précurseurs, suivant le degré d'évolution atteint;

- la composante référentielle. Cette composante s'intéresse de façon générale à la reconnaissance des objets du monde[2] désignés par les énoncés. Le problème le plus typique relevant de cette composante est l'identification du référent d'un pronom mais le champ couvert par cette composante est beaucoup plus large puisqu'il peut aller jusqu'à l'identification des noms propres (cf. tâche appelée "Named Entities" des évaluations MUC). Dans le cadre de MLK, la composante conceptuelle et la composante référentielle sont supposées associées afin de produire la représentation pré-thématique des textes constituant l'entrée de cette partie d'ANTHAPSI;

---

[1] Rappelons qu'à l'instar de Vygotsky, on distingue dans MoHA différentes étapes dans la constitution des concepts, allant du regroupement d'expériences en tas sur la base de la simple similitude d'un trait, en passant par la notion de complexe, où ce regroupement est fondé sur un plus grand réseau de liens, pour finir avec les concepts abstraits, dotés d'une définition en intension. Dans le cas présent, nous reprenons l'essentiel de ces grandes lignes de développement de la composante conceptuelle.

[2] Il est à noter que les actions, et même plus généralement les événements, peuvent faire l'objet d'une telle désignation.

- la composante causale. Du point de vue de MLK, cette composante met en évidence les relations causales existant entre les propositions d'une Unité Thématique. Plus généralement, elle recouvre les connaissances sur les déterminismes causaux élémentaires unissant les objets et les actions du monde ainsi que les mécanismes permettant de les reconnaître dans un texte;

- la composante thématique. Nous ne nous attarderons pas davantage sur cette composante puisque c'est celle que nous avons étudiée tout au long du présent travail;

- la composante spatio-temporelle. Cette composante est responsable d'une part de la recherche des relations temporelles entre les actions explicitées dans les textes (le fait qu'une action en précède une autre par exemple), et d'autre part de la mise en évidence des relations spatiales entre les objets désignés par les énoncés ainsi que de l'évolution de ces relations en fonction des actions relatées par ces mêmes énoncés (par exemple, se déplacer vers un endroit conduit à se trouver à cet endroit à la fin du déplacement). Dans MLK, cette composante est plus spécifiquement sollicitée afin de découvrir les relations temporelles présentes entre les propositions d'une UT.

Bien qu'ANTHAPSI ne soit pas un système de compréhension de texte généraliste, il constitue une première cible pour l'extension évoquée ci-dessus. Nous avons souligné la présence d'une différence importante de niveau entre ses deux constituants, ROSA et MLK, différence relative à la nature des connaissances manipulées et des représentations construites. Le saut résultant du passage de l'un à l'autre ne concerne pas tant la composante thématique que les autres composantes mentionnées ci-dessus[1]. Les représentations de texte dans MLK sont fondées sur des graphes conceptuels, dans lesquels on suppose résolus les problèmes de référence. Des relations temporelles et causales sont d'autre part présentes entre les graphes d'une même UT. Les représentations de texte de MLK font donc intervenir, outre la composante thématique, les composantes conceptuelle, référentielle, spatio-temporelle et causale. En comparaison, les représentations de texte de ROSA, constituées de regroupements de mots relatifs à un même thème, ne font guère intervenir que la seule composante thématique.

L'intervention de toutes ces composantes au niveau de MLK se fait par ailleurs à un niveau de précision supposé élevé, comparable à celui de la composante thématique. Le problème posé est donc semblable à celui rencontré pour cette dernière et la solution que nous proposons est similaire. Il s'agit à nouveau d'avoir recours à l'amorçage,

[1] C'est d'ailleurs pour cette raison que l'amorçage de MLK par SEGAPSITH n'est pas irréaliste sur le seul plan thématique.

c’est-à-dire de distinguer différents niveaux de précision et de structuration des connaissances pour chaque composante, de définir pour chacun de ces niveaux des moyens d’analyse permettant d’exploiter les connaissances du niveau en question et enfin, de spécifier la façon dont s’effectue le passage d’un niveau à un autre.

La particularité, dans ce cas, réside dans la mise en œuvre simultanée de plusieurs composantes, laquelle complexifie à la fois le processus d’analyse des textes, qui est alors le produit de la coopération de ces différentes composantes, et le processus d’amorçage, qui n’intervient plus seulement à l’intérieur d’une seule composante mais prend place également entre des composantes différentes. La première dimension de cette complexification renvoie plus généralement aux problèmes d’architecture des systèmes de traitement du langage naturel et à la nécessité d’y introduire une certaine flexibilité afin de pallier les problèmes d’ambiguïté et d’incomplétude des connaissances qui sont inhérents à l’analyse de textes en monde ouvert. Un système tel que CARAMEL [Sabah & Briffault 1993] est un exemple d’un pas fait dans cette direction.

Le seconde axe de la complexification considérée entraîne pour sa part une multiplication des schémas d’évolution possibles. Chaque niveau ne correspond plus, comme dans l’architecture actuelle d’ANTHAPSI, à un degré de structuration et de précision s’appliquant globalement à toutes les connaissances manipulées mais représente une évolution d’un certain type de connaissances attaché à une composante spécifique. Bien entendu, cette évolution peut également toucher plusieurs composantes simultanément. Cependant, l’idée centrale est que l’évolution d’un système tel qu’ANTHAPSI doit se faire de façon progressive, presque continue, par avancées successives de ses différentes composantes.

À tout moment de cette évolution, on dispose de cette manière d’un système d’analyse des textes couvrant les différentes composantes visées, même si toutes ne se trouvent pas nécessairement au même stade de leur développement. Selon ce mode de fonctionnement, l’activité d’une partie des composantes peut ainsi contribuer à l’amélioration d’une ou de plusieurs autres composantes, situation qui pourra éventuellement s’inverser par la suite. L’amorçage ne se fait donc plus seulement entre deux versions successives d’une même composante mais également de façon transversale, entre composantes de natures différentes.

Un processus de résolution des anaphores pronominales, même assez simple, peut contribuer par exemple à améliorer un système de segmentation thématique opérant au niveau lexical. À son tour, les bornes de segment qu’il fournira pourront être utilisées comme indices complémentaires pour perfectionner la résolution des anaphores. Cette

aide réciproque peut intervenir lors du traitement d'un texte particulier. C'est l'objet de la première dimension dégagée précédemment.

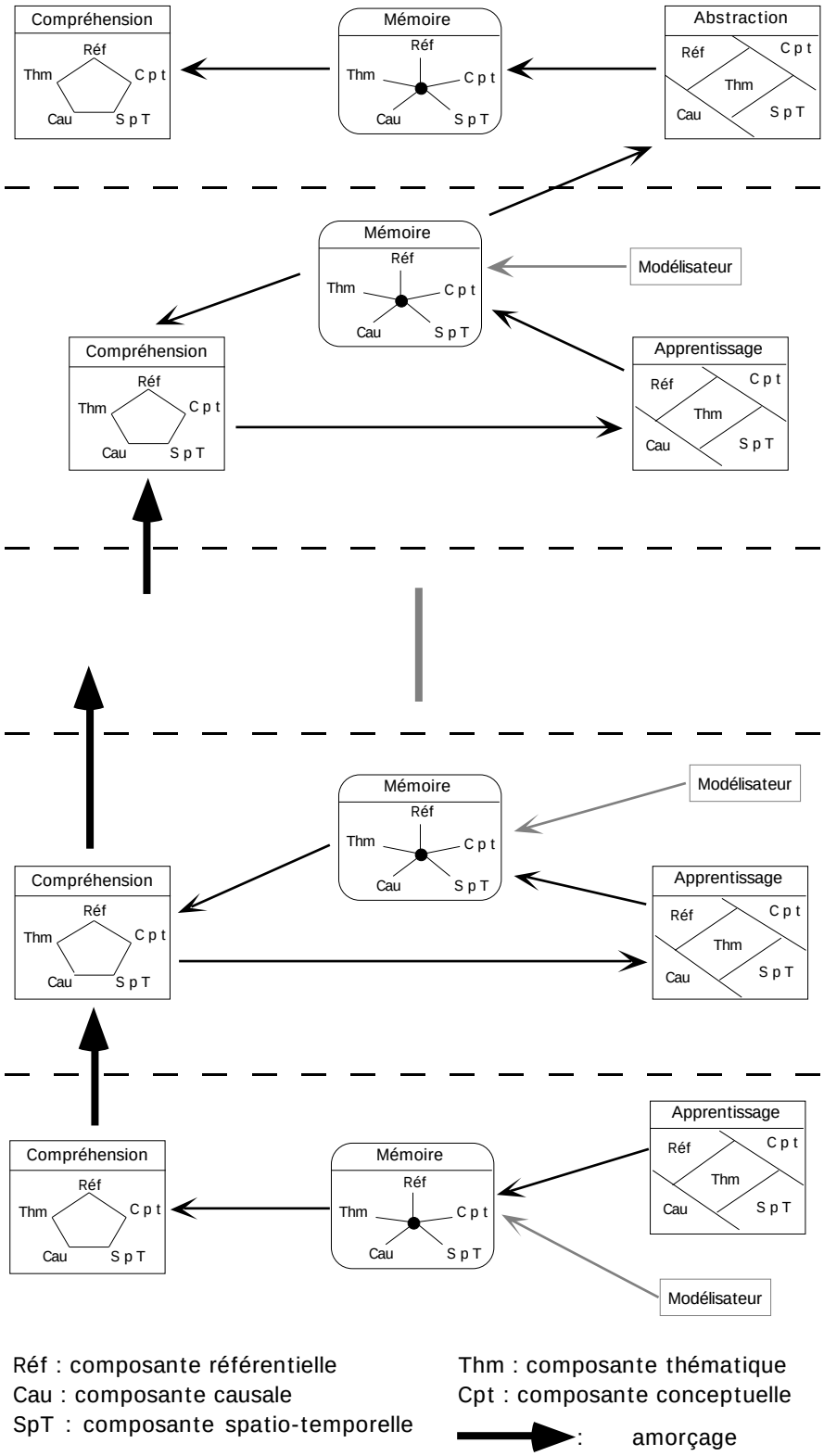


Fig. 1 - Architecture générale d'un système de compréhension de textes et d'apprentissage de connaissances fondé sur l'amorçage

Mais elle peut aussi intervenir sur un plus long terme lorsque ces apports respectifs sont capitalisés au travers de l'apprentissage de nouvelles connaissances. L'association de la segmentation thématique de SEGCOHLEX et d'un module de résolution des anaphores pronominales pourrait ainsi contribuer plus efficacement à l'amorçage de SEGAPSITH que la seule segmentation en mettant en évidence certains mots oubliés dans les signatures et en modifiant le poids d'autres mots. L'apprentissage permet en outre d'opérer sur le long terme un recoupement entre les résultats obtenus et donc, de faire la distinction entre les produits de cette association jugés intéressants, puisqu'apparaissant de façon suffisamment fréquente, et ceux supposés peu pertinents, parce que très rares.

Au delà d'ANTHAPSI, la démarche décrite ci-dessus peut s'appliquer plus globalement à un système complet de compréhension de texte, système qui ne sera pas centré autour de la composante thématique, comme c'est le cas d'ANTHAPSI, mais qui verra une implication beaucoup plus équilibrée des différentes composantes évoquées. L'extension de l'amorçage à un tel système est illustrée par la figure 1. On y retrouve les caractéristiques principales de l'amorçage dans ANTHAPSI : l'inter-dépendance entre compréhension et apprentissage ; la présence d'une mémoire permettant de faire le lien entre ces deux dimensions ; la nécessité d'un niveau initial fondé sur des connaissances établies une fois pour toutes ; la présence d'un niveau final abordé non pas via un amorçage entre processus de compréhension mais par abstraction des connaissances les plus précises et les plus structurées.

En dehors de la diversification de l'amorçage sur laquelle nous avons mis l'accent ci-dessus, ce cadre plus général se caractérise par la présence possible d'un modélisateur humain, élément que nous n'avions pas fait apparaître dans ANTHAPSI. Le terme de modélisateur est à prendre d'ailleurs selon un sens très générique puisque dans la plupart des cas, on n'attend pas d'un opérateur humain éventuel qu'il fournisse des connaissances mais plutôt qu'il valide ou qu'il réfute les connaissances apprises par le système. L'intervention humaine dans un tel contexte est à la fois nécessaire et très difficile. La nécessité vient de ce qu'il semble utopique en pratique qu'une chaîne d'amorçage puisse se dérouler de façon parfaitement autonome et sans aucun retour externe pour passer d'un niveau tel que ROSA jusqu'à un niveau comparable à MLK. La difficulté vient quant à elle de l'absence de restriction à un domaine précis. Pour certains types de connaissances, par exemple une partie des connaissances conceptuelles (cf. Wordnet), il est possible que cette absence de restriction soit humainement gérable. Pour d'autres, comme dans le cas des connaissances pragmatiques, elle ne l'est clairement pas. Dans toute sa généralité, le problème d'un système de compréhension de textes généraliste reste donc ouvert pour l'essentiel, même si nous espérons que les mécanismes proposés constituent au moins l'amorce d'une solution.

Pour achever cette discussion, nous nous ressituerons par rapport à MoHA, le modèle qui a initialement inspiré notre travail. En comparaison avec le modèle illustré par la figure 1, MoHA est à la fois plus spécifique et plus général. Il est plus spécifique dans la mesure où dans sa forme actuelle, il s'intéresse principalement aux connaissances conceptuelles et pragmatiques. Cette limitation n'est cependant pas théorique : le traitement de la référence, de la causalité et des aspects spatio-temporels s'y inscrivent assez naturellement, même s'ils n'ont pas fait pour le moment l'objet de développements très avancés. En revanche, MoHA est plus général par sa volonté de ne pas se limiter à la seule modalité langagière.

Sur le plan théorique, l'apport principal de notre travail par rapport à MoHA nous semble résider, outre la concrétisation de l'apprentissage de connaissances pragmatiques, dans la proposition d'un mécanisme de gestion de la transition d'un niveau de connaissances à un autre, mécanisme a priori largement applicable au sein de MoHA. Sur le plan pratique, nos travaux concernant ROSA s'articulent assez directement avec ceux de Jean-Pierre Gruselle sur l'émergence de concepts [Gruselle 1997]. Il nous semble en particulier qu'à court terme, il est possible d'utiliser les mécanismes de segmentation thématique de SEGCOHLEX afin de réaliser le découpage des textes en situations requis par le travail de Jean-Pierre Gruselle et à l'inverse, d'exploiter les résultats de ce même travail pour remplacer progressivement les mots des signatures thématiques par des concepts résultant d'un processus d'apprentissage.

Bibliographie personnelle relative au travail de thèse

Ferret Olivier 1998. *How to thematically segment texts by using lexical cohesion?,* Actes ACL-COLING'98 (Student Session), Montréal, Canada, pp. 1481-1483.

Ferret Olivier 1998. *Une segmentation thématique fondée sur la cohésion lexicale,* Actes TALN'98, Paris, pp. 32-41.

Ferret Olivier 1996. *Un système qui s'appuie sur son expérience pour segmenter des textes,* Actes Récital'96, Courcelles, France, pp. 129-136.

Ferret Olivier & Grau Brigitte 1998. Construire une mémoire épisodique à partir de textes : pourquoi et comment?, *Revue d'Intelligence Artificielle,* volume 12, n°3, pp. 377-409.

Ferret Olivier & Grau Brigitte 1998. *Structuration d'un Réseau de Cooccurrences Lexicales en Domaines Sémantiques par Analyse de Textes,* Actes Natural Language Processing and Industrial Applications (NLP+IA 98), Moncton, Canada, pp. 220-226.

Ferret Olivier & Grau Brigitte 1998. *A Thematic Segmentation Procedure for Extracting Semantic Domains from Texts,* Actes ECAI'98, Brighton, pp. 155-159.

Ferret Olivier & Grau Brigitte 1997. *An Aggregation Procedure for Building Episodic Memory,* Actes 15th International Joint Conference on Artificial Intelligence (IJCAI), Nagoya, Japan, pp. 280-285.

Ferret Olivier & Grau Brigitte 1997. An episodic memory for understanding and learning, dans *Recent Advances in Natural Languages Processing : Selected Papers from RANLP'95*, édité par R. Mitkov and N. Nicolov, Current Issues in Linguistic Theory (CILT), John Benjamins, Amsterdam/Philadelphia, pp. 173-184.

Ferret Olivier & Grau Brigitte 1997. *Une Analyse Thématique s'Appuyant sur une Mémoire Épisodique,* Actes 1ères Journées Scientifiques et Techniques FRANCIL, Avignon, France, pp. 161-168.

Ferret Olivier & Grau Brigitte 1996. *Construire une mémoire épisodique à partir de textes : pourquoi et comment ?,* Actes RFIA'96, Rennes, pp. 1115-1124.

Ferret Olivier & Grau Brigitte 1995. *An episodic memory for understanding and learning,* Actes Recent Advances in Natural Language Processing, Tzigov Chark, Bulgarie, pp. 221-229.

Ferret Olivier, Grau Brigitte & Masson Nicolas 1998. *Thematicsegmentation of texts: two methods for two kinds of texts,* Actes ACL-COLING'98, Montréal, Canada, pp. 392-396.

Ferret Olivier, Grau Brigitte & Masson Nicolas 1998. Utilisation d'un réseau de cooccurrences lexicales pour améliorer une analyse thématique fondée sur la distribution des mots, dans *Organisation des connaissances en vue de leur intégration dans les systèmes de représentation et de recherche d'information*, Presses Universitaires de Lille, Lille, France.

Bibliographie

Aamodt Agnar & Plaza Enric 1994. Case-Based Reasoning: fondational issues, methodological variations and system approaches, *AICommunication*, volume 7, n°1, pp. 39-59.

Adam Jean-Michel 1984. *Le Récit*, Que sais-je?, Presses Universitaires de France, Paris.

Adda Gilles, Calmès Martine de, Lamel Lori, Pérennou Guy, Rajman Martin, Rosset Sophie & Zeilinger Jérôme 1997. *Ressources pour l'apprentissage, le développement et l'évaluation des systèmes de dictée vocale en français : corpus de texte, de parole et lexical*, Actes 1ères Journées Scientifiques et Techniques FRANCIL, Avignon, France, pp. 305-309.

Anderson John 1983. *The Architecture of Cognition*, Havard University Press.

Apté Chidanand, Damerau Fred & Weiss Sholom 1994. Automated learning of decision rules for text categorization, *ACM Transactions on Informations Systems*, volume 12, n°3, pp. 233-251.

ARPA & Agency Advanced Research Projects (éd.) 1996. *Sixth Message Understanding Conference (MUC-6)*, Morgan Kaufmann, San Mateo, CA.

Asher Nicholas 1993. *Reference to Abstract Objects in Discourse*, Kluwer, Norwell, MA.

Bardou Bruce 1995. *Etude et réalisation d'un mécanisme de propagation d'activation pour la sélection de connaissances*, Rapport de DEA (Sciences Cognitives), Université d'Orsay.

Bartlett F. C. 1932. *Remembering : A study in experimental and social psychology*, Cambridge University Press, Londres.

Beeferman D., Berger A. & Lafferty J. 1997. *Text segmentation using exponential models*, Actes Second Conference on Empirical Methods in Natural Language Processing, Providence.

Béguin Annette, Jouis Christophe & Mustafa Widad 1997. *Évaluation d'outils d'aide à la construction de terminologie et de relations sémantiques entre termes à partir de corpus*, Actes 1ères Journées Scientifiques et Techniques FRANCIL, Avignon, France, pp. 419-425.

Bérard-Dugourd A., Fargues J. & Landau M.-C. 1988. *Natural language analysis using conceptual graphs*, Actes International Computer Science Conference' 88, Hong-Kong, pp. 265-272.

Biber Douglas 1993. Using Register-Diversified Corpora for General Language Studies, *Computational Linguistics*, volume 19, n°2, pp. 219-241.

Bichindaritz Isabelle 1994. *Apprentissage de concepts dans une mémoire dynamique : raisonnement à partir de cas adaptable à la tâche cognitive*, Thèse de doctorat, Université René Descartes - Paris V.

Bobrow D. G. & Norman D. A. 1975. Some principles of memory schemata, dans *Representation and understanding : studies in Cognitive Science*, édité par D. G. Bobrow and A. M. Collins, Academic Press, New York.

Bonnard Henri 1981. *Code du Français Courant,* Magnard.

Bordeaux François 1993. *Association entre Perceptions Visuelles et LangageNaturel pour l'Apprentissage de Connaissances Sémantiques : le Modèle Hybride MoHA,* Actes Atelier sur la Formation des Symboles dans les Modèles de la Cognition, Grenoble.

Bouaud Jacques, Bachimond Bruno & Zweigenbaum Pierre 1996. *Processing metonymy: a domain-model heuristic graph traversal approach,* Actes 16th COLING, Copenhagen, Denmark, pp. 137-142.

Brachman Ronald J. 1979. On the epistemological status of semantic networks, dans *Associative networks : representation and use of knowledge by computers,* édité par N. V. Findler, Academic Press, New York.

Briffault Xavier, Chibout Karim, Sabah Gérard & Vapillon Jérome 1997. *An object-oriented linguistic engineering environment using LFG (Lexical-Functional Grammar) and CG (Conceptual Graphs),* Actes ACL'97 Workshop on Computational Environments for Grammar Development and Linguistic Engineering, Madrid.

Brown Gillian & Yule George 1983. *Discourse Analysis,* Textbooks in Linguistics Series, Cambridge University Press.

Carletta Jean 1996. Assessing agreement on classification tasks: The kappa statistic, *Computational Linguistics,* volume 22, n°2, pp. 249-254.

Chalendar Gaël de 1997. *Abstraction de Schémas à partir de Situations Agrégées,* Rapport de DEA (Sciences Cognitives), Université d'Orsay.

Charolles Michel 1993. Les plans d'organisation du discours et leurs interactions, dans *Parcours linguistiques de discours spécialisés,* édité par S. Moirand and al., Peter Lang, Berne, pp. 301-315.

Chibout Karim 1993. *Traitement automatique des tropes,* Rapport de DEA (Sciences Cognitives), Université d'Orsay.

Chibout Karim & Vilnat Anne 1997. *Primitives Sémantiques, Classification des Verbes et Polysémie,* Actes 1ères Journées du Chapitre Français de l'ISKO, Lille, France.

Church Kenneth Ward & Hanks Patrick 1990. Word Association Norms, Mutual Information, And Lexicography, *Computational Linguistics,* volume 16, n°1, pp. 22-29.

Church Kenneth W. & Mercer Robert L. 1993. Introduction to the Special Issue on Computational Linguistics Using Large Corpora, *Computational Linguistics,* volume 19, n°1, pp. 1-24.

Collins A. M. & Quillian M. R. 1969. Retrieval time from semantic memory, *Journal of Verbal Learning and Verbal Behavior,* volume 8, pp. 240-248.

Cornuejols Antoine 1989. *De l'Apprentissage Incrémental par Adaptation Dynamique : le système INFLUENCE,* Thèse de doctorat, Université de Paris-Sud Orsay.

Cullingford R. E. 1978. *Script Application: Computer Understanding of Newpaper Stories,* Doctoral Dissertation, Department of Computer Science, Yale University.

Cutting Douglass R., Karger David R., Pedersen Jan O. & Tukey John W. 1992. *Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections,* Actes 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'92).

Dahlgren Kathleen 1993. Discourse Coherence and Segmentation, dans *Computational and Conversational Discourse. Burning Issues - An Interdisciplinary Account*, édité par E. H. Hovy and D. R. Scott, N. A. Series, Series F: Computer and System Sciences, Vol. 151, Springer Verlag, pp. 111-138.

Danyluk Andrea Pohoreckyj 1987. *The use of explanations for similarity-based learning*, Actes Tenth International Conference on Artificial Intelligence, Milan, pp. 274-276.

DeJong Gerald 1983. *Acquiring schemata through understanding and Generalizing plans,* Actes Eighth International Joint Conference on Artificial Intelligence (IJCAI), Karlsruhe, pp. 462-464.

DeJong Gerald 1982. *Automatic schema acquisition in a natural language environment,* Actes Nat'l. Conference on Artificial Intelligence, Pittsburgh, pp. 410-413.

DeJong Gerald 1981. *Generalizations Based on Explanations,* Actes Seventh International Joint Conference on Artificial Intelligence (IJCAI), Vancouver, pp. 67-69.

DeJong Gerald & Mooney Raymond 1986. Explanation-Based Learning: An Alternative View, *Machine Learning*, volume 1, pp. 145-176.

Dijk T. van & Kinstch W. 1983. *Strategies of Discourse Comprehension*, Academic Press, New York.

Domeshek E. 1991. *What Abby cares about,* Actes Workshop on case-based reasoning (DARPA), Washington, D.C., Morgan Kaufmann.

Dormont Agnès & Gruselle Jean-Pierre 1993. *A Constructivist Approach to Tense and Aspect: from text to Polytyped Strings,* Actes IEEE Tools for Artificial Intelligence, Boston, pp. 484-485.

Dowty D. 1982. Tenses, Time Adverbs and Compositional Semantic Theory, *Linguistics and Philosophy*, volume 5, pp. 23-33.

Ellis Gerard 1992. Compiled Hierarchical Retrieval, dans *Conceptual Structures - current research and practice*, édité par T. E. Nagle, J. A. Nagle, L. L. Gerholz and P. W. Eklund, Ellis Horwood, New York, pp. 271-294.

Esch John 1992. Linear Forms for Conceptual Structures, dans *Conceptual Structures: current research and practice*, édité par T. E. Nagle, J. A. Nagle, L. L. Gerholz and P. W. Eklund, Ellis Horwood Workshops, England, pp. 595-604.

Esch John, Pagnucco Maurice, Wermelinger Michel & Pfeiffer Heather 1994. *LINEAR — Linear Notation Interface*, Actes Third International Workshop on PIERCE: A Conceptual Graphs Workbench, University of Maryland, College Park.

Fano R. 1961. *Transmission of Information: A Statistical Theory of Communivations,* MIT Press, Cambridge, MA.

Ferrari Stéphane 1993. *Étude sur le traitement automatique des métaphores,* Rapport de DEA (Informatique), Université d'Orsay.

Fillmore C. J. 1968. The case for case, dans *Universals of linguistic theory*, édité par E. Bach and R. T. Harms, Holt, Rinehart & Winston, New York, pp. 1-90.

Fodor Jerry A. 1981. *Representations: Philosophical Essays on the Foundations of Cognitive Science*, MIT Press, Cambridge, Massachusetts.

Fodor Jerry A. 1975. *The Language of Thought*, Thomas Y. Crowell Co., New York.

Forbus Kenneth D. 1988. Quantitative physics: Past, present and future, dans *Exploring artificial intelligence*, édité par H. Shrobe, Morgan Kaufmann, California.

Forest Françoise 1997. *Comment représenter l'expérience individuelle qui donne leur sens aux mots, approche informatique,* Actes Vèmes journées LTT (Lexcologie, Lexicologie, Traduction), Tunis.

Forest Françoise 1991. Se donner les moyens d'une approche constructiviste de la représentation du sens - le traitement massivement parallèle des données LIMSI, 91-21, Décembre 1991.

Forest Françoise & Grau Brigitte 1992. *HLM, an Hybrid Learning Model. A model based on the perception of the environment by an individual,* Actes IPMU'92 (on Information Processing and Management of Uncertainty in knowledge-based systems), Mallorca, pp. 607-610.

FraCaS Consortium 1996. Chapter 3: A Semantic Test Suite, dans *Public Deliverables of the FRACAS Project, Deliverable D16*, CE, DG XIII.

Francis Anthony G. 1995. *Memory-Based Opportunistic Reasoning*, Thesis Proposal, Georgia Institute of Technology.

Francis Anthony G. 1994. Psychological Aspect of MOORE: The Memory Organization and Optimized Retrieval Engine .

Gale William A., Church Kenneth W. & Yarowsky David 1992. *Estimating upper and lower bounds on the performance of word-sense disambiguation programs,* Actes 30th Annual Meeting of the Association for Computational Linguistics, pp. 249-256.

Gennari John H., Langley Pat & Fisher Doug 1989. Models of Incremental Concept Formation, *Artificial Intelligence,* volume 40, n°1-3 Special Volume on Machine Learning, pp. 11-61.

Grau Brigitte 1984. *Stalking Coherence in the Topical Jungle*, Actes FGCS, Fifth Generation Computer System, Tokyo.

Grau Brigitte 1983. *Analyse et représentation d'un texte d'après le thème du discours,* Thèse de troisième cycle, Université Pierre et Marie Curie - Paris VI.

Grau Brigitte & Sabah Gérard 1985. *Acquisition automatique des connaissances pragmatiques,* Actes Cognitiva, Paris, pp. 687-693.

Grau Brigitte & Vilnat Anne 1997. *Cooperation in Dialogue and Discourse Structure,* Actes Workshop of IJCAI'97 on Collaboration, Cooperation and Conflicts in Dialogue Systems, Nagoya, Japan.

Greimas A. J. 1970. *Du Sens - Essais sémiotiques,* Seuil.

Greimas A. J. 1966. *Sémantique structurale*, Larousse, Paris.

Grosz Barbara, Joshi Aravind & Weinstein Scott 1983. *Providing an unified account of definite noun phrases in discourse,* Actes 21th Annual Meeting of the Association of Computational Linguistics, pp. 44-50.

Grosz Barbara J. & Sidner Candace L. 1986. Attention, Intentions and the Structure of Discourse, *Computational Linguistics,* volume 12, pp. 175-204.

Grumbach Alain 1994. *Cognition artificielle - Du réflexe à la réflexion,* Addison-Wesley France, Paris.

Gruselle Jean-Pierre 1997. *Le rôle du mot dans la formation des concepts : un modèle informatique et son implantation*, Thèse de doctorat, Université d'Orsay.

Guha Amal 1995. *Compréhension de textes : une exploration du modèle de W. Kintsch*, Rapport de DEA (Sciences Cognitives), Université Paris-Sud.

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

Harnad Stevan 1987. Category induction and representation, dans *Categorical perception - The groundwork of cognition*, édité par S. Harnad, Cambridge University Press, Cambridge.

Hearst Marti 1993. TextTiling: A quantitative approach to discourse segmentation, Technical Report Sequoia Computer Science Division, University of California, Berkeley, 93/24.

Hearst Marti A. 1997. TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages, *Computational Linguistics*, volume 23, n°1, pp. 33-64.

Hearst Marti A. 1994. *Multi-paragraph segmentation of expository text,* Actes 32th Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, USA, pp. 9-16.

Hillis W. D. 1985. *The Connection Machine*, MIT Press.

Hirschberg Julia & Grosz Barbara J. 1992. *Intonational Features of Local and Global Discourse Structure,* Actes Darpa Workshop on Speech and Natural Language.

Hirschberg Julia & Litman Diane J. 1993. Empirical Studies on the Disambiguation of cue phrases, *Computational Linguistics*, volume 19, n°3, pp. 501-530.

Hirst Graeme & St-Onge David 1995. Lexical chains as representations of context for the detection and correction of malapropisms, dans *WordNet*, édité par C. Fellbaum, MIT Press, Cambridge, MA.

Ho H.G. 1990. *Représentation des valeurs sémantiques du passé composé en français en vue du traitement informatique,* Thèse d'état, Université Paris-Sorbonne.

Hobbs Jerry R. 1979. Coherence and Co-reference, *Cognitive Science,* volume 3, n°1, pp. 67-82.

Hovy Eduard & Lin Chin Yew 1997. *Automated Text Summarization in SUMMARIST,* Actes ACL 97 Workshop on Intelligent Scalable Text Summarization, Madrid, Espagne, pp. 18-24.

Irandoust Hengameh 1997. *Vers un Modèle Cognitif de la Structure Temporelle des Textes Narratifs*, Thèse de doctorat, Université d'Orsay.

Jackiewicz Agata 1996. *Filtrage d'informations textuelles par une approche contextuelle de la causalité,* Actes Récital'96, Courcelles, France, pp. 114-120.

Jodouin Jean-François 1993. *Réseaux de Neurones et Traitement du Langage Naturel - Étude des réseaux de neurones récurrents et de leurs représentations,* Thèse de doctorat, Université Paris XI Orsay.

Kamp H. 1981. A Theory of Truth and Semantic Representation, dans *Formal Methods in the Study of Language*, édité par J. Groenendijk, T. Janssen and M. Stockhof, Mathematisch Centrum, Amsterdam, pp. 277-322.

Kaplan R.M. & Bresnan J. 1982. Lexical-Functional Grammar: A formal system for grammatical representation, dans *The Mental Representation of Grammatical Relations*, édité par J. Bresnan, MIT Press, Cambridge, Mass., pp. 173-281.

Kass Alex, Leake David & Owens Christopher 1986. SWALE, A Program that Explains, dans *Explanations Patterns: Understanding Mechanically and Creatively*, édité par R. C. Schank, Lawrence Erlbaum, Hillsdale, NJ, pp. 232-254.

Kayser Daniel 1987. Une sémantique qui n'a pas de sens, *Langages,* volume 87, pp. 33-45.

Kekenbosh C. & Denhière G. 1988. L'Activation et la Diffusion de l'Activation, *L'Année Psychologique,* volume 88, pp. 237-255.

Kenny A. 1963. *Action, Emotion and Will,* Routledge, London.

Kintsch W. & Dijk T.A. van 1978. Toward a Model of Text Comprehension and Production, *Psychological Review,* volume 85, n°5, pp. 363-394.

Kodratoff Yves & Michalski Ryszard S. 1990. Research in Machine Learning: Recent Progress, Classification of Methods, and Future Directions, dans *Machine Learning: An Artificial Intelligence Approach - Volume III,* édité par Y. Kodratoff and R. S. Michalski, Morgan Kaufmann, pp. 3-30.

Kolodner Janet 1993. *Case-Based Reasoning*, Morgan Kaufmann Publishers.

Kolodner Janet L. 1983. Maintaining Organization in a Dynamic Long-Term Memory, *Cognitive Science,* volume 7, n°4, pp. 243-280.

Kolodner Janet L. 1983. Reconstructive Memory: A Computer Model, *Cognitive Science,* volume 7, n°4, pp. 281-328.

Kolodner Janet L. & Simpson R. L. 1989. The MEDIATOR: Analysis of an early case-based problem solver, *Cognitive Science,* volume 13, n°4, pp. 507-549.

Kozima Hideki 1993. *Computing Lexical Cohesion as a Tool for Text Analysis*, Doctoral Thesis, University of Electro-Communications.

Kozima Hideki 1993. *Text Segmentation Based on Similarity between Words*, Actes 31th Annual Meeting of the Association for Computational Linguistics (Student Session), Colombus, Ohio, USA, pp. 286-288.

Lange Trent E. & Dyer Michael G. 1989. High-level Inferencing in a Connectionist Network, *Connection Science,* volume 1, n°2, pp. 181-217.

Lange Trent E. & Wharton Charles M. 1993. *Dynamic Memories: Analysis of an Integrated Comprehension and Episodic Memory Retrieval Model,* Actes Thirteenth International Conference on Artificial Intelligence, Chambery, pp. 208-213.

Lappin Shalom & Leass Herbert J. 1994. An Algorithm for Pronominal Anaphora Resolution, *Computational Linguistics,* volume 20, n°4, pp. 535-561.

LDOCE 1987. *Longman Dictionary of Contemporary English,* Longman, Harlow, Essex.

Lebowitz Michael 1990. The Utility of Similarity-Based Learning in a World Needing Explanation, dans *Machine Learning: An Artificial Intelligence Approach - Volume III*, édité par Y. Kodratoff and R. S. Michalski, Morgan Kaufmann, pp. 399-422.

Lebowitz Michael 1988. *Deferred Commitment in UNIMEM: Waiting to learn,* Actes Fifth Machine Learning Conference, Ann Arbor, Michigan, pp. 80-86.

Lebowitz Michael 1986. Concept Learning in a Rich Input Domain: Generalization-Based Memory, dans *Machine Learning: An Artificial Intelligence Approach - Volume II*, édité par R. S. Michalski, J. G. Carbonell and T. M. Mitchell, Morgan Kaufmann, Los Altos, California, pp. 193-214.

Lebowitz Michael 1983. Generalization From Natural Language Text, *Cognitive Science*, volume 7, pp. 1-40.

Lenat Douglas B., Guha R., Pittman K., Pratt D. & Sheperd M. 1990. Cyc: towards programs with common sense, *Communications of the ACM*, volume 33, n°8, pp. 30-49.

Lewis David D. & Ringuette M. 1994. *A comparison of two learning algorithms for text categorization,* Actes Third Annual Symposium on Document Analysis and Information Retrieval, pp. 81-93.

Lin Chin-Yew 1997. *Robust Automated Topic Identification,* Doctoral Dissertation, University of Southern California.

Litman Diane J. & Passonneau Rebecca J. 1995. *Combining Multiple Knowledge Sources for Discourse Segmentation,* Actes 33th Annual Meeting of the Association for Computational Linguistics.

Maire-Reppert D. 1990. *L'imparfait de l'indicatif en vue d'un traitement automatique du français*, Thèse de doctorat, Université Paris IV.

Mann W.C. & Thompson S.A. 1987. Rhetorical Structure Theory: A Theory of Text Organization ISI, ISI-RS-87-190.

Martin J. R. 1992. *English Text - System and Structure*, John Benjamins Publishing Company, Philadelphia/Amsterdam.

McClelland J. L. 1988. Connectionist Models and Psychological Evidence, *Journal of Memory and Language,* volume 27, pp. 107-123.

Michalski Ryszard S. 1977. *A system of programs for computer aided induction: A summary,* Actes Fifth International Joint Conference on Artificial Intelligence (IJCAI), Morgan Kaufmann.

Miller A. G., Fellbaum C. & Gross D. 1989. *WordNet: a Lexical Database Organised on Psycholinguistic Principles,* Actes First International Lexical Acquisition Workshop - IJCAI, Detroit.

Mineau Guy W. 1992. Induction on Conceptual Graphs: Finding Common Generalizations and Compatible Projections, dans *Conceptual Structures - current researchandpractice*, édité par T. E. Nagle, J. A. Nagle, L. L. Gerholz and P. W. Eklund, Ellis Horwood, New York, pp. 295-310.

Minsky Marvin 1975. A framework for representing knowledge, dans *The Psychology of Computer Vision*, édité par P. H. Winston, McGraw Hill, New York.

Mitchell Tom M. 1982. Generalization as search, *Artificial Intelligence,* volume 18, n°2, pp. 203-226.

Mitchell Tom M., Keller Richard M. & Kedar-Cabelli Smadar T. 1986. Explanation-Based Generalization: A Unifying View, *Machine Learning,* volume 1, pp. 47-80.

Mooney Raymond & DeJong Gerald 1985. *Learning schemata for natural language processing,* Actes Ninth International Joint Conference on Artificial Intelligence, Los Angeles, pp. 681-687.

Moore Johanna D. & Pollack Martha E. 1992. A Problem for RST: The Need for Multi-Level Discourse Analysis, *Computational Linguistics*, volume 18, pp. 537-544.

Morris Jane & Hirst Graeme 1991. Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text, *Computational Linguistics,* volume 17, n°1, pp. 21-48.

Moulinier Isabelle, Raskinis Gailius & Ganascia Jean-Gabriel 1996. *Text Categorization: a Symbolic Approach.,* Actes de SDAIR'96, Las Vegas.

Nazarenko Adeline 1994. *Compréhension du Langage Naturel : le problème de la causalité,* Thèse de doctorat, Université Paris XIII - Institut Galilée.

Nédellec Claire 1994. *APT, apprentissage interactif de règles de résolution de problèmes en présence de théorie du domaine,* Thèse de doctorat, Université Paris-Sud.

Nogier Jean-François 1991. *Génération automatique de langage et graphes conceptuels,* Langue, Raisonnement, Calcul, Hermès, Paris.

Nomoto Tadashi & Nitta Yoshihiko 1994. *A Grammatico-Statistical Approach To Discourse Partitioning,* Actes 15th International Conference on Computational Linguistics (COLING), Kyoto, Japan, pp. 1145-1150.

Okumura Manabu & Honda Takeo 1994. *Word Sense Disambiguation and Text Segmentation Based on Lexical Cohesion,* Actes 15th International Conference on Computational Linguistics (COLING), Kyoto, Japan, pp. 755-761.

Passonneau Rebecca J. 1993. Coding Scheme and Algorithm for Identification of Discourse Segment Boundaries on the Basis of the Distribution of Referential Noun Phrases, Technical Report Columbia University.

Passonneau Rebecca J. & Litman Diane J. 1996. Empirical Analysis of Three Dimensions of Spoken Discourse: Segmentation, Coherence, and Linguistic Devices, dans *Computational and Conversational Discourse. Burning Issues - An Interdisciplinary Account,* édité par E. H. Hovy and D. R. Scott, Springer Verlag, pp. 161-194.

Passonneau Rebecca J. & Litman Diane J. 1993. *Intention-based segmentation: Human reliability and correlation with linguistic cues,* Actes 31st Annual Meeting of the Association for Computational Linguistics, Colombus, Ohio, USA, pp. 148-155.

Pazzani Michael 1991. A Computational Theory of Learning Causal Relationships, *Cognitive Science,* volume 15, pp. 401-424.

Pazzani Michael 1991. Learning Causal Patterns: Making a transition form data-driven to theory-driven learning, *Machine Learning,* volume 11, pp. 173-194.

Pazzani Michael J. 1988. *Integrating explanation-based and empirical learning methods in OCCAM,* Actes Third European Working Session on Learning, Glasgow, pp. 147-165.

Pearl Judea 1988. *Probabilistic Reasoning in Intelligent Systems,* Morgan Kaufmann.

Pitrat Jacques 1990. *Métaconnaissances - futur de l'intelligence artificielle,* Hermès.

Polanyi L. 1988. A Formal Model of the Structure of Discourse, *Journal of Pragmatics,* volume 12, pp. 601-638.

Popescu-Belis Andrei & Robba Isabelle 1997. *Cooperation between Pronoun and Reference Resolution for Unrestricted Texts,* Actes ACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts, Madrid, Spain.

Propp V. 1970. *Morphologie du Conte,* Collection Points n°12, Seuil, Paris.

Quinlan John R. 1993. *C4.5 : Programs for Machine Learning,* Morgan Kaufmann Publishers, San Mateo, CA.

Rady Mohamed 1983. *L'ambiguïté du langage naturel est-elle la source du non-déterminisme des procédures de traitement?*, Thèse d'état, Université Pierre et Marie Curie.

Ram Ashwin 1993. Indexing, elaboration and refinement: incremental learning of explanatory cases, *Machine Learning,* volume 10, n°3, pp. 7-54.

Rastier François 1989. *Mot, Phrase, Texte : Pour une Sémantique Descriptive Unifiée,* Actes Semantica : les modèles sémantiques pour le traitement automatique du langage, Nanterre, EC2.

Reinhart T. 1984. Principles of Gestalt Perception in the Temporal Organization of Narrative Texts, *Linguistics,* volume 22, pp. 779-809.

Reynar Jeffrey C. 1994. *An Automatic Method of Finding Topic Boundaries*, Actes 32th Annual Meeting of the Association for Computational Linguistics (Student Session), Las Cruces, New Mexico, USA.

Riloff Ellen & Lehnert Wendy 1994. Information Extraction as a Basis for High-Precision Text Classification, *ACM Transactions on Information Systems,* volume 12, n°3, pp. 296-333.

Robba Isabelle 1992. *L'étude de mécanismes de raisonnement par analogie dans le cadre de l'analyse de phrases. Le système MIRA.*, Thèse de doctorat, Paris XI Orsay.

Roget P. 1977. *Roget's International Thesaurus, Fourth Edition,* Harper and Row Publishers Inc.

Rosch E. 1977. Principles of categorization, dans *Cognition and Categorization,* édité par E. Rosch and B. Lloyds, Lawrence Erlbaum, Hillsdale, N.J.

Rosé Carolyn Peinstein 1995. *Conversation acts, interactional structure, and conversational outcomes*, Actes Empirical Methods in Discourse: Interpretation and Generation, Menlo Park, CA, AAAI Press.

Rumelhart D.E. 1977. Understanding and Summarizing Brief Stories, dans *Basic Processes in Reading: Perception and Comprehension,* édité par D. LaBerge and S. J. Samuels, Laurence Erlbaum, Hillsdale, N.J., pp. 265-303.

Rumelhart D. E. & McClelland J. L. (éd.) 1986. *Parallel Distributed Processing: Explorations in Microstructure of Cognition,* volume 1, MIT Press.

Sabah Gérard 1988. *L'intelligence artificielle et le langage - Volume 1 - Représentation des connaissances*, Hermès, Paris.

Sabah Gérard 1978. *Contribution à la compréhension effective d'un récit,* Thèse d'état, Université Pierre et Marie Curie, Paris.

Sabah Gérard & Briffault Xavier 1993. *CARAMEL : A Step Towards Reflection in Natural Language Understanding Systems,* Actes IEEE International Conference on Tools with Artificial Intelligence, Boston.

Sabah Gérard & Vilnat Anne 1991. *Flexible Case Structure Implemented into a Deterministic Parser,* Actes Sixth Annual Workshop on Conceptual Structures, Binghamton, AAAI and Sony, pp. 343-356.

Sabatier Paul 1997. *Évaluer des Systèmes de Compréhension de Textes,* Actes 1ères Journées Scientifiques et Techniques FRANCIL, Avignon, France, pp. 223-226.

Salton Gerard 1989. *Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer*, Reading, Addison-Wesley, MA.

Salton Gerard & Allan James 1993. *Selective Text Utilization and Text Traversal*, Actes Hypertext-93, New York, pp. 131-144.

Salton Gerard, Allan James, Buckley Chris & Singhal Amit 1994. Automatic analysis, theme generation, and summarization of machine-readable texts, *Science*, volume 264, pp. 1421-1426.

Salton Gerard, Singhal Amit, Buckley Chris & Mitra Mandar 1996. *Automatic Text Decomposition Using Text Segments and Text Themes*, Actes Hypertext'96, Seventh ACM Conference on Hypertext, Washington, D.C., pp. 53-65.

Schank Roger C. 1986. *Explanation Patterns - Understanding Mechanically and Creatively*, Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Schank Roger C. 1982. *Dynamic Memory: a theory of reminding and learning in computers and people*, Cambridge University Press, New York.

Schank Roger C. 1972. Conceptual Dependency: A Theory of Natural Language-Understanding, *Cognitive Psychology*, volume 3-4, pp. 552-631.

Schank Roger C. & Abelson Robert P. 1977. *Scripts, plans, goals and understanding*, Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Schank Roger C. & Leake David B. 1989. Creativity and Learning in a Case-Based Explainer, *Artificial Intelligence*, volume 40, n°1-3, pp. 353-385.

Schank Roger C. & Osgood R. 1990. A content theory of memory indexing, Technical Report Northwestern University, Institute for the Learning Sciences, 2.

Schmid Helmut 1994. *Probabilistic Part-of-Speech Tagging Using Decision Trees*, Actes International Conference on New Methods in Language Processing, Manchester, UK.

Schröder M. 1992. *Knowledge based analysis of radiology reports using conceptual graphs*, Actes 7th Annual Workshop on Conceptual Graphs, Las Cruces, NM, pp. 213-222.

Séligman Laurence 1985. *Intégration de la syntaxe, de la sémantique et de la pragmatique dans un analyseur de textes*, Thèse de doctorat, Université Pierre et Marie Curie.

Sidner Candace 1983. Focusing in the comprehension of definite anaphora, dans *Computational Models of Discourse*, édité par M. Brady and R. Berwick, MIT Press, Cambridge, MA, pp. 267-330.

Silberztein Max 1993. *Dictionnaires électroniques et analyse automatique de textes : le système INTEX*, Masson, Paris.

Sowa John F. 1992. Conceptual Graphs Summary, dans *Conceptual Structures: current research and practice*, édité par T. E. Nagle, J. A. Nagle, L. L. Gerholz and P. W. Eklund, Ellis Horwood Workshops, England, pp. 3-51.

Sowa John F. 1984. *Conceptual Structures: Information Processing in Mind and Machine*, Addison Wesley.

Stairmand Mark 1994. *Lexical chains, WordNet and information retrieval*, Master Thesis, Centre for Computational Linguistics, UMIST, Manchester.

Stark Heather 1988. What do paragraph markers do?, *Discourse Processes,* volume 11, n°3, pp. 275-304.

Stein Achim & Schmid Helmut 1995. Étiquetage Morphologique de Textes Français avec un Arbre de Décisions, *Traitement Automatique des Langues,* volume 36, n°1-2, pp. 23-35.

Todorov Tzvetan 1972. Motif, dans *Dictionnaire encyclopédique des sciences du langage*, édité par O. Ducrot and T. Todorov, Points, Éditions du Seuil, pp. 280-285.

Tversky A. 1977. Features of similarity, *Psychological Review,* volume 84, pp. 327-352.

Vazov Nicolay 1997. *Le rôle de l'information temporelle dans le raisonnement et l'identification des différentes structures de texte,* Actes 1ères Journées du Chapitre Français de l'ISKO, Lille, France.

Vendler Z. 1967. *Linguistics and Philosophy*, Cornell University Press, Ithaca.

Véronis Jean & Khouri Liliane 1995. Étiquetage grammatical multilingue : le projet MULTEXT, *TAL,* volume 36, n°1-2, pp. 233-248.

Vygotski L.S. 1962. *Thought and Language*, MIT Press, Cambridge, Massachusetts.

Waltz D. L. & Pollack J. B. 1985. Massively Parallel Parsing: A Strongly Interactive Model of Natural Language Interpretation, *Cognitive Science,* volume 9, pp. 51-74.

Wayne Charles L. 1997. *Topic Detection & Tracking (TDT) - Overview & Perspective,* Actes Topic Detection and Tracking Workshop, College Park, MD.

Wermelinger Michel 1995. *Conceptual Structures Linear Notation: A proposal for PIERCE,* Actes Fourth International Workshop on PIERCE: A Conceptual Graphs Workbench, Santa Cruz, CA, pp. 13-24.

Wilensky R., Chin D., Luria M., Martin J., Mayfield J. & Wu D. 1988. The Berkeley UNIX Consultant Project, *Computational Linguistics,* volume 14, n°4, pp. 35-84.

Wilensky Robert W. 1983. *Planning and Understanding: A Computational Approach to Human Reasoning*, Addison-Wesley, MA.

Yang Yiming, Carbonell Jaime, Allan James & Yamron Jon 1997. *Topic Detection and Tracking - Detection Task,* Actes Topic Detection and Tracking Workshop, College Park, MD.

Youmans Gilbert 1991. A New Tool for Discourse Analysis: The Vocabulary-Management Profile, *Language,* volume 67, n°4, pp. 763-789.

Zweigenbaum Pierre & Bouaud Jacques 1997. *Construction d'une représentation sémantique en Graphes Conceptuels à partir d'une analyse LFG,* Actes TALN'97, Grenoble, pp. 30-39.

Zweigenbaum Pierre, Bouaud Jacques, Habert Benoît & Nazarenko Adeline 1997. *Coopération apprentissage en corpus et connaissances du domaine pour la construction d'ontologies,* Actes 1ères Journées Scientifiques et Techniques FRANCIL, Avignon, France, pp. 501-508.

Zweigenbaum Pierre & MENELAS Consortium 1995. MENELAS: coding and information retrieval from natural language patient discharge summaries, dans *Advances in Health Telematics*, édité par M. F. Laires, M. J. Ladeira and J. P. Christensen, IOS Press, Amsterdam, pp. 82-89.

Annexe A

La notation linéaire des graphes conceptuels

1. Grammaire de la notation linéaire des graphes conceptuels

La grammaire que nous donnons est directement sous la forme requise pour l'outil *Tgen* que nous avons utilisé afin de construire le compilateur de cette notation (seules les actions accompagnant les règles ont été supprimées). *Tgen* est un outil comparable à l'association de Lex et Yacc et permet d'engendrer en Smalltalk des compilateurs à partir d'une grammaire exprimée sous forme déclarative. Les conventions d'écriture des grammaires dans *Tgen* sont très proches de la Backus Norm Form (BNF). La principale différence avec cette notation réside dans l'inversion entre terminaux et non-terminaux : les terminaux dans *Tgen* apparaissent encadrés par "< >" alors que cette marque désigne les non-terminaux dans le cadre de la BNF.

Par ailleurs, ce préambule s'applique également aux grammaires des annexes B et C.

```
"-------------------------------------------------------------"
"-- Notation étendue des graphes conceptuels"
"---- spécification des differents types de graphes"
"-------------------------------------------------------------"
CGFile :        CGDeclarations;

CGDeclarations  :       CGDeclarations Statement
                |;

Statement       :       GraphDefinition
                |       ConceptTypeDefinition
                |       RelationTypeDefinition
                |       ConceptTypeCreation
                |       RelationTypeCreation
                |       CanonDefinition
                |       CanonicalGraphDefinition
                |       PrototypeDefinition
                |       SchemaDefinition;

GraphDefinition :       GraphName GraphNameSep CGraph
                |       CGraph;

ConceptTypeCreation     :       CONCEPTTYPE TypeField InheritsFrom TypeDefiningList;

RelationTypeCreation    :       RELATIONTYPE TypeField InheritsFrom TypeDefiningList;

TypeDefiningList        :       TypeDefiningList TypeListSep TypeField
```

```
                        |       TypeField;
ConceptTypeDefinition   :       TYPE TypeField Argument IS CGraph;
RelationTypeDefinition  :       RELATION TypeField Arguments IS CGraph;
Argument        :       BeginArguments VariableDef EndArguments;
Arguments       :       BeginArguments VariableList EndArguments;
VariableList    :       VariableList VariableSep VariableDef
                |       VariableDef;
CanonicalGraphDefinition        :       CANONICAL GRAPH FOR TypeField IS CGraph;
PrototypeDefinition     :       PROTOTYPE FOR TypeField IS CGraph;
SchemaDefinition        :       SCHEMA FOR TypeField IS CGraph;
CanonDefinition         :       CANON CanonName BeginCanon GraphList EndCanon;
CanonName       :       ContextName CanonNameSep CanonName
                |       ContextName;
GraphList       :       GraphDefinition GraphList
                |       GraphDefinition;

"-----------------------------------------------------------"
"-- Notation adoptée pour un graphe conceptuel"
"---- structure générale d'un graphe"
"-----------------------------------------------------------"
CGraph          :       ConceptBranch
                |       Relation ConceptLink
                |       Relation ConceptList;
ConceptBranch   :       Concept
                |       Concept RelationLink
                |       Concept RelationList;
RelationBranch  :       Relation
                |       Relation ConceptLink
                |       Relation ConceptList;
ConceptLink     :       Arc ConceptBranch;
RelationLink    :       Arc RelationBranch;
ConceptList     :       BeginList ConceptListBody EndList;
ConceptListBody         :       ConceptListBody ListSep ConceptLink
                        |       ConceptLink;
RelationList    :       BeginList RelationListBody EndList;
RelationListBody        :       RelationListBody ListSep RelationLink
                        |       RelationLink;
Arc     :       RightArc
        |       LeftArc;

"-----------------------------------------------------------"
"---- définition des entités de base des graphes"
"-----------------------------------------------------------"
Concept         :       BeginConcept ConceptBody EndConcept
                |       BeginConcept ConceptBody AnnotationDelimiter Annotation EndConcept;
Relation        :       BeginRelation RelationBody EndRelation
```

```
                |       BeginRelation RelationBody AnnotationDelimiter Annotation EndRelation;

ConceptBody     :       TypeField
                |       TypeField FieldSep ReferentField
                |       GraphSetBody;

RelationBody    :       TypeField;

TypeField       :       TypeLabel
                |       Not TypeLabel;

Annotation      :       AnnotationElement ListSep Annotation
                |       AnnotationElement;

AnnotationElement       :       AnnotationType FieldSep AnnotationValue;

"-----------------------------------"
"------ définition des référents"
"-----------------------------------"
ReferentField   :       SetReferent
                |       IndividualReferent;

SetReferent     :       BeginSetAttributes SetDefinition EndSetAttributes;

BeginSetAttributes      :       SetType
                        |;

EndSetAttributes        :       Variable
                        |       Cardinality
                        |       Variable Cardinality
                        |;

IndividualReferent      :       IndividualElement
                        |       GenericReferent
                        |       GraphElement;

IndividualElement       :       IndividualIdentifier IndividualAttributes
                        |       IndividualIdentifier
                        |       IndividualAttributes;

GraphElement    :       CGraph GraphAttributes
                |       CGraph;

GraphAttributes :       Variable;

GenericReferent :       GenericMark;

IndividualAttributes    :       Name
                        |       Name Variable
                        |       Name Measure
                        |       Name Variable Measure
                        |       Variable Measure
                        |       Measure
                        |       Variable;

Cardinality     :       MeasureMark Integer;

Measure         :       MeasureMark Number Unit;

SetDefinition   :       BeginSet SetBody EndSet;

SetBody         :       GenericReferent
                |       IndividualSetBody
                |       GraphSetBody;

IndividualSetBody       :       IndividualElement ListSep IndividualSetBody
                        |       IndividualElement ListSep GenericMark
```

```
                    |       IndividualElement;
GraphSetBody    :   GraphElement ListSep GraphSetBody
                |   GraphElement ListSep GenericMark
                |   GraphElement;


"---------------------------------"
"-- délimiteurs et symboles"
"---------------------------------"
BeginConcept    :   '[';                TypeLabel       :   <identifier>;
BeginRelation   :   '(';                AnnotationType  :   <identifier>;
BeginList       :   '{';                AnnotationValue :   <identifier>;
BeginSet        :   '{';                Name    :   <identifier>;
BeginArguments  :   '(';                ContextName     :   <identifier>;
BeginCanon      :   '{';                GraphName       :   <identifier>;
EndConcept      :   ']';                IndividualIdentifier    :   <individual>;
EndRelation     :   ')';                SetType         :   <identifier>;
EndList         :   '}';                Variable        :   <variable>;
EndSet  :   '}';                        VariableDef     :   <variable>;
EndArguments    :   ')';                Unit    :   <identifier>;
EndCanon        :   '}';                Number          :   <integer>
AnnotationDelimiter     :   ';';                        |   <real>;
ListSep         :   ',';                Integer :   <integer>;
FieldSep        :   ':';                CONCEPTTYPE     :   'ConceptType';
GraphNameSep    :   ':';                RELATIONTYPE    :   'RelationType';
VariableSep     :   ',';                TYPE    :   'Type';
TypeListSep     :   ',';                RELATION        :   'Relation';
CanonNameSep    :   '/';                CANONICAL       :   'Canonical';
RightArc        :   '->';               GRAPH           :   'graph';
LeftArc  :   '<-';                      FOR     :   'for';
Not     :   '~';                        IS      :   'is';
InheritsFrom    :   '<';                PROTOTYPE       :   'Prototype';
GenericMark     :   '*';                SCHEMA          :   'Schema';
MeasureMark     :   '@';                CANON           :   'Canon';
```

2. Exemples

Concepts et référents

- générique : [Homme: *] ou [Homme]
- avec coréférence : [Humain: *x1]
- avec mesure : [Taille: @180 cm]
- nommé: [Homme: Jean]
- individuel : [Femme: #34]
- avec annotations : [Couper; prédicat: vrai; aspect: perfectif]
- ensembliste générique : [Humain: {*}]
- ensembliste avec coréférence : [Homme: {*} *x2]
- ensembliste avec contrainte de cardinalité : [Femme: {*} @3]
- ensembliste avec typage : [Humain: dist {*}]

- ensembliste complexe : [Homme: {*x4, Pierre,#67,*}@5 *x3]

Graphes et contextes

- graphe linéaire : [Homme] ← (agent) ← [Casser] → (objet) → [Verre]

- graphe avec factorisation d'un concept :
  [Transporter]
    { → (agent) → [Homme]→ (aPourMétier) → [Déménageur],
    → (objet) → [Armoire],
    → (moyen) → [Camion],
    → (origine) → [Ville: Rouen],
    → (destination) → [Ville: Amiens]
    }

- graphe avec cycle :
  [Donner]
    { → (agent) → [Femme: *x1],
    → (destinataire) → [Humain],
    → (objet) → [Argent] → (appartientA) → [Femme: *x1],
    }

- contexte avec un seul graphe :
  [Contexte : [Homme] ← (agent) ← [Casser] → (objet) → [Verre]] ou
  [[Homme] ← (agent) ← [Casser] → (objet) → [Verre]]

- contexte avec plusieurs graphes et coréférence entre graphes :
  [Contexte : {[Homme: *x1] ← (agent) ← [Casser] → (objet) → [Verre], [Être_en_colère] → (source) → [Homme: *x1]}] ou
  [[Homme: *x1] ← (agent) ← [Casser] → (objet) → [Verre], [Être_en_colère] → (source) → [Homme: *x1]]

Bases de connaissances

- graphe canonique :
  Canonical graph for Casser *is*
    [Être_animé] ← (agent) ← [Casser] → (objet) → [Objet_physique]
  même principe pour les prototypes et les schémas, avec le mot-clé 'Prototype' à la place de 'Canonical graph' dans le premier cas et le mot-clé 'Schéma' dans le second cas.

- graphe de définition :

Type Couper *x *is*

[Couper: *x]

{ → (agent) → [Animé: *y],

→ (objet) → [Entité_concrète: *z] → (caractéristique) → [Solide],

→ (moyen) → [Entité_concrète] → (caractéristique) → [Tranchant],

→ (méthode) → [[Animé: *y] ← (agent) ← [Traverser] → (objet) → [Entité_concrète: *z]]

}

même principe pour les relations mais en remplaçant le mot-clé ‘Type’ par le mot-clé ‘Relation’.

- définition du treillis des types :

ConceptType Animal < Entité_mobile, Animé

même principe pour les relations mais en remplaçant le mot-clé ‘ConceptType’ par le mot-clé ‘RelationType’

- définition d’une base de connaissances :

Canon base/sous-base1/partie3

{ graphe1: [Homme] ← (agent) ← [Casser] → (objet) → [Verre]

graphe2: [Homme] ← (agent) ← [Boire] → (objet) → [Vin]

}

3. Particularités de la notation adoptée par rapport à celle de Sowa

Par rapport aux spécifications que Sowa présente dans [Sowa 1984], notre réalisation se différencie sur trois points : la formalisation des extensions de la notation des graphes conceptuels de base, la modification de certains points de cette notation et enfin, l’ajout de nouvelles extensions. Le premier point concerne la définition des contextes et des différents types de connaissances (graphes canoniques, graphes de définition, prototypes, etc.). Dans [Sowa 1984], la syntaxe de ces différentes définitions est illustrée par des exemples mais n’est pas formalisée rigoureusement par une grammaire. Pour réaliser notre compilateur, nous avons repris les propositions les plus couramment admises sur cette question précise, propositions que l’on trouve notamment dans [Esch 1992].

Le deuxième point s’inspire quant à lui de suggestions faites dans [Wermelinger 1995]. La notation adoptée par Sowa pour les graphes comprenant des factorisations de concept présente plusieurs inconvénients : utilisation du caractère ‘Retour chariot’ au

niveau de la grammaire, nécessité d'inférer le sens des relations, présence de virgules en cascade en fin de graphe. Exemple : [C1] –

(R1)

(R2) → [C2],.

Il nous a paru en l'occurrence plus clair d'adopter une notation proche des blocs d'instructions en langage C. Une accolade ouvrante signale le début de la factorisation tandis qu'une accolade fermante en marque explicitement la fin. Pour chaque relation, le sens est systématiquement indiqué à la fois avant la relation et après, ce qui évite d'avoir à inférer le sens de la relation à partir de la marque suivant la relation. Par ailleurs, les parties des graphes correspondant aux différentes relations concernées par la factorisation sont séparées par des virgules. On n'a pas ainsi à utiliser le caractère 'Retour chariot' comme délimiteur.

Pour ce qui est des nouvelles extensions, nous avons puisé là aussi dans les propositions les plus établies. En l'occurrence, nous avons choisi de mettre en œuvre la notion d'annotation. Celle-ci consiste à ajouter des informations non liées au formalisme des graphes conceptuels au niveau des concepts et des relations. Dans le cadre du traitement des langues naturelles, les annotations permettent par exemple de conserver des informations provenant d'autres analyses que l'analyse sémantique. Dans le cas des représentations de texte que nous manipulons dans MLK, les annotations servent ainsi à identifier le concept d'une proposition ayant le rôle de prédicat. Les annotations prennent ici la forme d'un ensemble de couples attribut-valeur, séparés par des points-virgules.

Exemple : [Couper; prédicat: vrai; aspect: perfectif]

Annexe B

Notation linéaire des schémas et outils de la mémoire pragmatique

1. Grammaire de la notation linéaire des schémas

(cf. préambule de l'annexe A à propos des conventions d'écriture de la grammaire)

Du fait de l'utilisation de graphes conceptuels dans les schémas, il faudrait bien évidemment joindre à cette grammaire celle de la notation des graphes conceptuels. Puisque celle-ci est donnée à l'annexe A, nous nous sommes contenté de faire apparaître en gras les non-terminaux faisant le pont avec cette grammaire.

On trouvera un exemple de représentation d'un schéma avec cette notation linéaire au niveau de la figure 4.10 du chapitre 4.

```
"-------------------------------------------------------"
"---- Notation adoptée pour la représentation"
"sous forme textuelle d'une liste de schémas"
"-------------------------------------------------------"
ReseauSchemas  :        ListeSchemas;

ListeSchemas   :        ListeSchemas Schema
               |;


"---------------------------------------"
"---- Représentation d'un schéma"
"---------------------------------------"
Schema         :        DebutSchema CorpsSchema FinSchema;

DebutSchema    :        MotCleDebutSchema IdentificateurSchema;

FinSchema      :        MotCleFinSchema IdentificateurSchema;

CorpsSchema    :        CorpsSchema EntiteSchema
               |;

EntiteSchema   :        AttributSchema
               |        RolesSchema
               |        Champ;

ListeChamps    :        ListeChamps Champ
               |;

Champ  :        NomChamp SepChamp ValeurChamp FinChamp;

ValeurChamp    :        CGraph
               |        Number
               |        Texte
```

```
                |       Symbole;

Texte   :       DebutTexte ListeMotsEtPonctuation FinTexte;

ListeMotsEtPonctuation  :       ListeMotsEtPonctuation Mot
                        |       ListeMotsEtPonctuation Ponctuation Mot
                        |;

"-------------------------------------------------------"
"---- Représentation des rôles d'un schéma"
"-------------------------------------------------------"
RolesSchema     :       MotCleDebutRolesSchema CorpsRolesSchema MotCleFinRolesSchema;

CorpsRolesSchema        :       CorpsRolesSchema RoleSchema
                        |;

RoleSchema      :       DefinitionRoleSchema FinRoleSchema
                |       DefinitionRoleSchema FinChamp ListeChamps DernierChampRoleSchema
                        FinRoleSchema;

DernierChampRoleSchema          :       NomChamp SepChamp ValeurChamp;

DefinitionRoleSchema    :       TypeLabel SepDefRolSchem Concept;

"-------------------------------------------------------"
"---- Représentation d'un attribut de schéma"
"-------------------------------------------------------"
AttributSchema  :       DebutAttribut CorpsAttribut FinAttribut;

DebutAttribut   :       MotCleDebutAttribut IdentificateurAttribut;

FinAttribut     :       MotCleFinAttribut IdentificateurAttribut;

CorpsAttribut   :       CorpsAttribut Reference
                |;

"-------------------------------------------------------------------------------"
"---- Représentation d'une référence à un schéma ou à un graphe"
"-------------------------------------------------------------------------------"
Reference       :       EnteteReference CorpsReference FinReference;

EnteteReference :       TypeObjetReference IdentificateurObjetReference;

CorpsReference  :       ListeChamps CGraph;


"-------------------------------------"
"---- Délimiteurs et symboles"
"-------------------------------------"
MotCleDebutSchema       :       'Schema';
MotCleDebutAttribut     :       'Attribut';
MotCleDebutRolesSchema          :       'Roles';
MotCleFinSchema         :       'FinSchema';
MotCleFinAttribut       :       'FinAttribut';
MotCleFinRolesSchema    :       'FinRoles';
DebutTexte      :       '<';
FinReference    :       '.';
FinChamp        :       ';';
FinRoleSchema   :       '.';
FinTexte        :       '>';
SepChamp        :       ':';
SepDefRolSchem          :       '->';
Ponctuation     :       ','
                |       '.';
```

```
IdentificateurAttribut      :       <identifier>;
IdentificateurSchema        :       <identifier>;
IdentificateurObjetReference        :       <identifier>;
NomChamp        :       <identifier>;
Symbole         :       <identifier>;
Mot     :       <identifier>;
TypeObjetReference      :       'Schema'
                        |       'Graphe';
```

2. Outils de gestion de la mémoire pragmatique

Le principal outil concernant la mémoire pragmatique est le gestionnaire illustré par la figure B.1. Celui-ci permet essentiellement de :

- créer de nouveaux schémas à partir de leur définition sous forme de fichiers texte (cf. notation linéaire du §1) ;
- supprimer des schémas existant ;

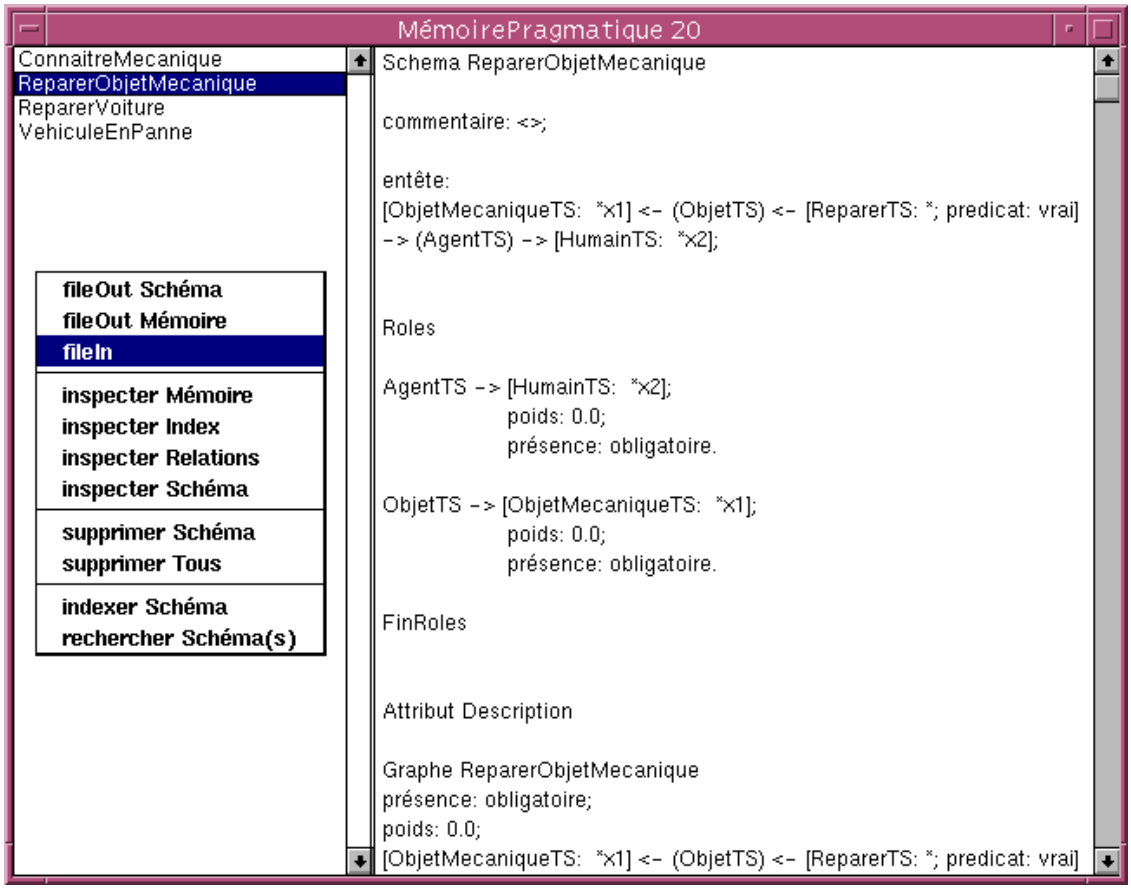


Fig. B.1 - Gestionnaire de la mémoire pragmatique

- visualiser le contenu des schémas et leurs relations, soit sous forme linéaire, soit sous forme d'objets Smalltalk. Dans ce dernier cas, il est possible de lancer un inspecteur spécialisé dédié aux schémas (cf. figure B.3), inspecteur faisant apparaître en accès direct toute la structure d'un schéma ;

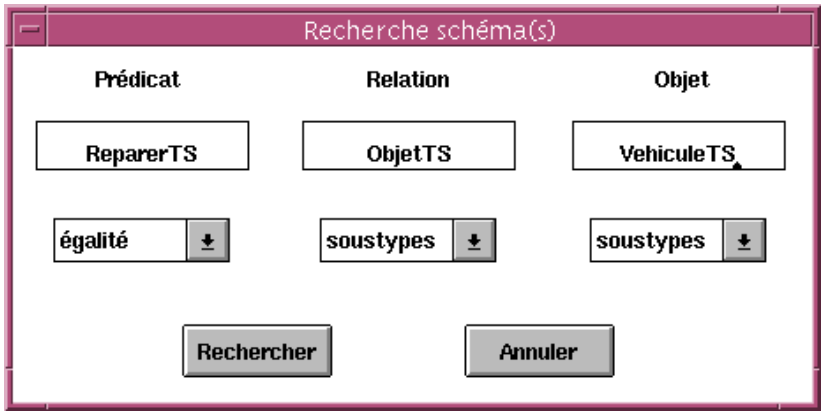


Fig. B.2 - Interface de recherche d'un schéma dans la mémoire pragmatique

- indexer un nouveau schéma ou changer l'indexation d'un schéma déjà indexé. Chaque schéma est indexé par deux concepts, supposés appartenir à son entête, ainsi que par la relation qui les lient au sein de celui-ci ;

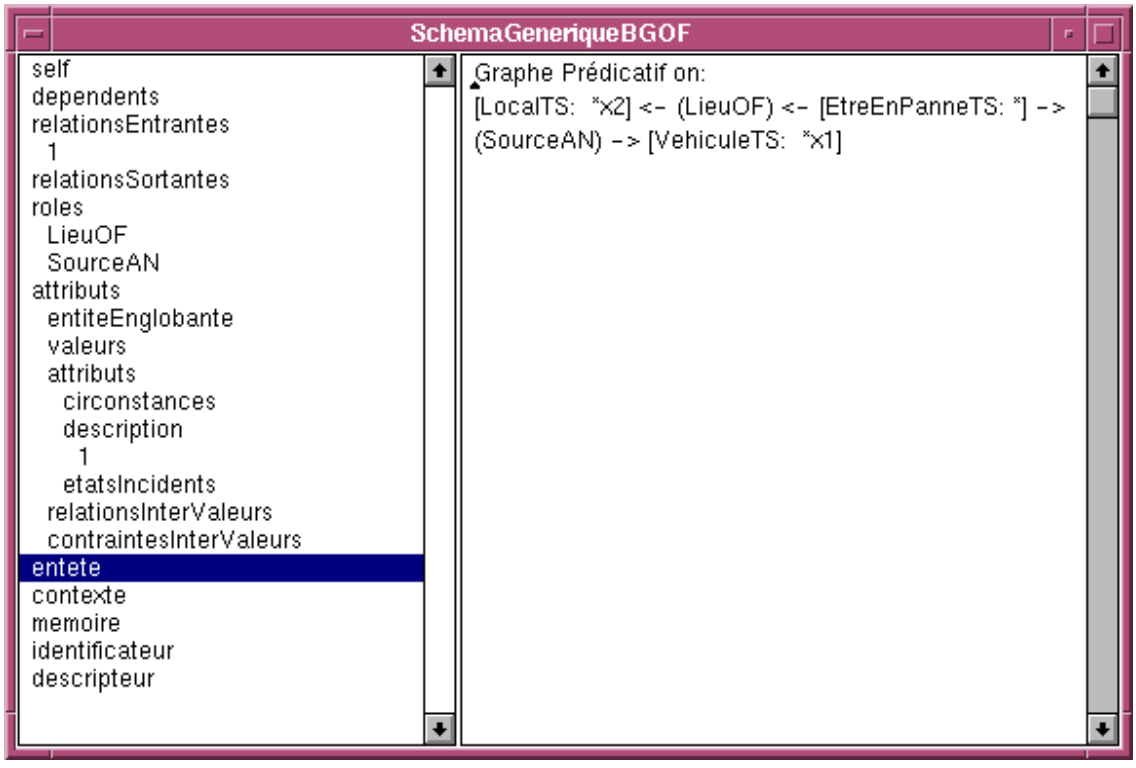


Fig. B.3 - Inspecteur spécialisé dédié aux schémas

- rechercher un schéma au sein de la mémoire à partir des index. La figure B.2 montre l'interface spécifiquement dédiée à cette fonction. Pour chaque index, on peut ainsi spécifier un certain nombre de conditions permettant de définir le champ de la recherche : égalité stricte des types, recherche également parmi les sous-types, recherche également parmi les sur-types, recherche également parmi les sous-types et les sur-types.

Il est à noter que l'interface d'indexation d'un nouveau schéma correspond à la partie supérieure de l'interface de recherche de la figure B.2.

Annexe C

Notation linéaire et outils de manipulation des représentations de texte

1. Grammaire de la notation linéaire des représentations de texte

(cf. préambule de l'annexe A à propos des conventions d'écriture de la grammaire)

Du fait de l'utilisation de graphes conceptuels dans les représentations de texte, il faudrait bien évidemment joindre à cette grammaire celle de la notation des graphes conceptuels. Puisque celle-ci est donnée à l'annexe A, nous nous sommes contenté de faire apparaître en gras les non-terminaux faisant le pont avec cette grammaire.

On trouvera un exemple d'une représentation de texte sous cette forme au niveau de la figure 5.5 du chapitre 5.

```
"-------------------------------------------------------------------------"
"---- Notation adoptée pour la représentation"
"sous forme textuelle d'une liste de représentations de texte"
"-------------------------------------------------------------------------"
RepresentationsDeTexte   :        ListeEpisodes;

ListeEpisodes    :        ListeEpisodes Episode
                 |;


"----------------------------------------"
"---- Représentation d'un épisode"
"----------------------------------------"
Episode          :        DebutEpisode CorpsEpisode FinEpisode;

DebutEpisode     :        MotCleDebutEpisode IdentificateurEpisode;

FinEpisode       :        MotCleFinEpisode IdentificateurEpisode ;

CorpsEpisode     :        CorpsEpisode ChampEpisode
                 |;

ChampEpisode     :        UniteThematique
                 |        RelationsInterUTs
                 |        Champ;

Champ  :         NomChamp SepChamp ValeurChamp FinChamp;

ValeurChamp      :        Texte
                 |        Symbole;
```

```
Texte    :        DebutTexte ListeMotsEtPonctuation FinTexte ;

ListeMotsEtPonctuation  :       ListeMotsEtPonctuation Mot
                        |       ListeMotsEtPonctuation Ponctuation Mot
                        |;

"---- Représentation des relations inter unites thématiques"
RelationsInterUTs       :       DebutRelsInterUTs CorpsRelsInterUTs FinRelsInterUTs ;

CorpsRelsInterUTs       :       CorpsRelsInterUTs RelationInterUTs
                                        |;

RelationInterUTs        :       EnteteRelInterUTs CorpsRelInterUTs FinRelInterUTs;

EnteteRelInterUTs       :       MotCleRelInterUTs IdentificateurRelInterUTs;

CorpsRelInterUTs        :       ListeChamps DefRelInterUTs;

ListeChamps     :       ListeChamps Champ
                |;

DefRelInterUTs  :       IdentificateurUT SepSourceCible IdentificateurUT SepDefRel
                        TypeRelationInterUTs
                |       IdentificateurUT SepObjetAttribut DesignationAttribut SepObjetAttribut
                        IdentificateurGraphe SepSourceCible IdentificateurUT SepDefRel
                        TypeRelationInterUTs;

DesignationAttribut     :       IdentificateurAttribut
                        |       DebutGrphSsAttrib;

"---- Représentation d'une unité thématique"
UniteThematique         :       DebutUT CorpsUT FinUT;

DebutUT         :       MotCleDebutUT IdentificateurUT;

FinUT    :      MotCleFinUT IdentificateurUT;

CorpsUT         :       CorpsUT ChampUT
                |;

ChampUT         :       AttributUT
                |       GraphesSansAttribut
                |       RelationsInterGraphes
                |       Champ;

"---- Représentation d'un attribut d'unité thématique"
AttributUT      :       DebutAttribut CorpsAttribut FinAttribut;

DebutAttribut   :       MotCleDebutAttribut IdentificateurAttribut;

FinAttribut     :       MotCleFinAttribut IdentificateurAttribut;

CorpsAttribut   :       CorpsAttribut Graphe
                |;

GraphesSansAttribut     :       DebutGrphSsAttrib CorpsAttribut FinGrphSsAttrib;

"---- Représentation d'un graphe"
Graphe   :      EnteteGraphe CorpsGraphe FinGraphe;

EnteteGraphe    :       MotCleGraphe IdentificateurGraphe;

CorpsGraphe     :       ListeChamps CGraph;
```

```
"---- Représentation des relations inter-graphes"
RelationsInterGraphes   :   DebutRelsInterGraphes CorpsRelsInterGraphes FinRelsInterGraphes;
CorpsRelsInterGraphes   :   CorpsRelsInterGraphes RelationInterGraphes
                        |;
RelationInterGraphes    :   EnteteRelInterGraphes CorpsRelInterGraphes FinRelInterGraphes;
EnteteRelInterGraphes   :   MotCleRelInterGraphes IdentificateurRelInterGraphes;
CorpsRelInterGraphes    :   ListeChamps DefRelInterGraphes;
DefRelInterGraphes      :   DesignationAttribut SepObjetAttribut IdentificateurGraphe
                            SepSourceCible DesignationAttribut SepObjetAttribut
                            IdentificateurGraphe SepDefRel TypeRelationInterGraphes;

"----------------------------------------"
"---- Délimiteurs et symboles"
"----------------------------------------"
MotCleDebutEpisode      :   'Episode';
MotCleDebutUT           :   'UT';
DebutRelsInterUTs       :
        'RelationsInterUTs';
DebutRelsInterGraphes   :
        'RelationsIntraUT';
DebutTexte      :   '<';
MotCleDebutAttribut     :   'Attribut';
DebutGrphSsAttrib       :   'Autres';
MotCleRelInterGraphes   :   'Relation';
MotCleRelInterUTs       :   'Relation';
MotCleGraphe    :   'Graphe';
MotCleFinEpisode        :   'FinEpisode';
MotCleFinUT     :   'FinUT';
FinRelsInterUTs         :
        'FinRelationsInterUTs';
FinRelsInterGraphes     :
        'FinRelationsIntraUT';
FinRelInterUTs  :   '.';
FinRelInterGraphes      :   '.';
FinGraphe       :   '.';
FinTexte        :   '>';
MotCleFinAttribut       :   'FinAttribut';
FinGrphSsAttrib         :   'FinAutres';
FinChamp        :   ';';
SepChamp        :   ':';
SepDefRel       :   ':';
SepSourceCible  :   '->';
SepObjetAttribut        :   '.';
Ponctuation     :   ','
                |   '.' ;
IdentificateurEpisode   :   <identifier>;
IdentificateurUT  :   <identifier>;
IdentificateurAttribut  :   <identifier>;
IdentificateurGraphe    :   <identifier>;
IdentificateurRelInterUTs :   <identifier>;
IdentificateurRelInterGraphes   :
        <identifier>;
NomChamp        :   <identifier>;
Symbole         :   <identifier>;
Mot     :   <identifier>;
TypeRelationInterUTs    :   <identifier>;
TypeRelationInterGraphes        :
        <identifier>;
```

2. Outils de manipulation des représentations de textes

Afin de faciliter les tests, nous avons développé la notion de base de textes. Celle-ci a pour objectif de centraliser les différentes représentations construites pour un ensemble de textes. Cette centralisation est pour le moment extrêmement rudimentaire, notamment parce que la mise en correspondance des différentes analyses est presque inexistante. Seule a été prévue pour le moment la mise en relation des graphes de la représentation sémantique et de la représentation situationnelle (i.e. représentation thématique) avec la forme de surface.

Sur le plan pratique, le gestionnaire de base de textes illustré par la figure C.1 se contente de remplir des fonctions de base :

- ajout d'un nouveau texte à la base. L'ajout peut se faire à partir de la forme de surface du texte, de la forme linéaire de sa représentation sémantique ou de la forme linéaire de sa représentation situationnelle ;

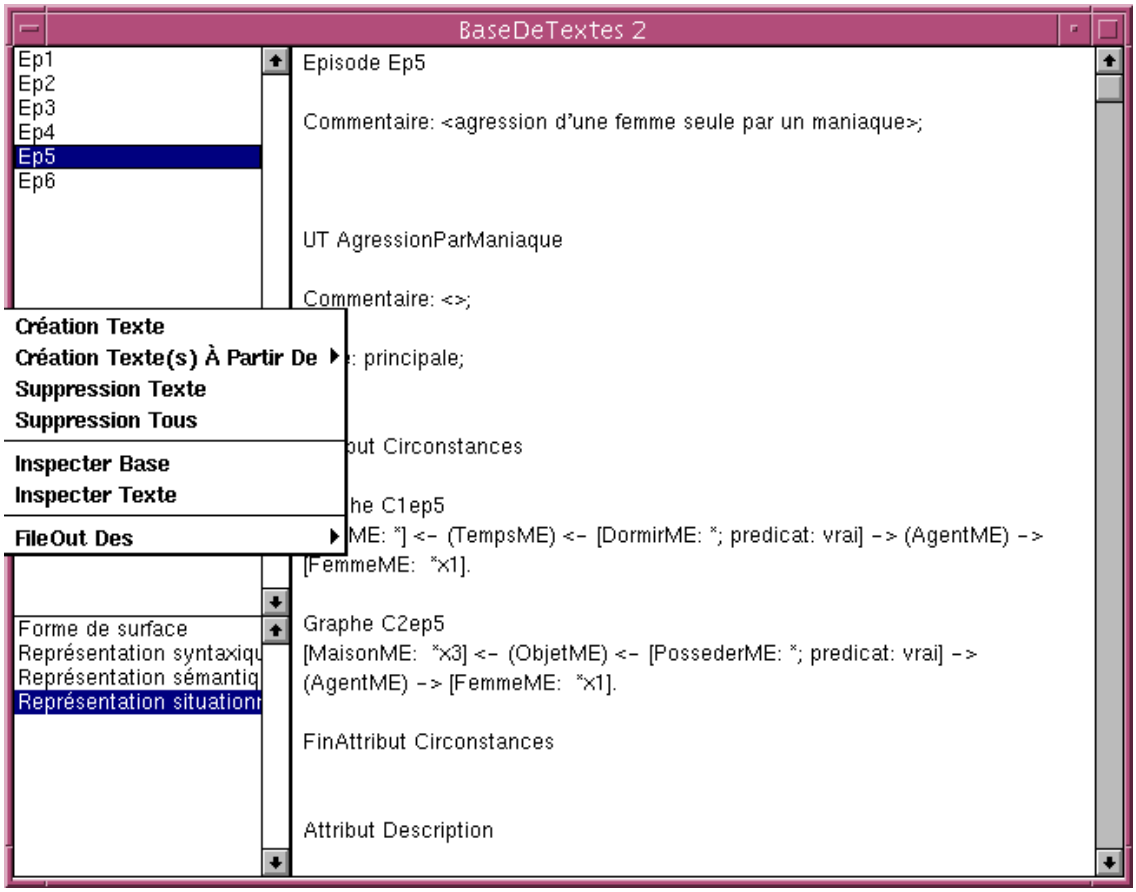


Fig. C.1 - Gestionnaire d'une base de texte

- suppression d'un texte de la base ;
- visualisation d'un texte de la base. Il s'agit plus précisément de visualiser ses différentes représentations, lorsqu'elles sont disponibles ;

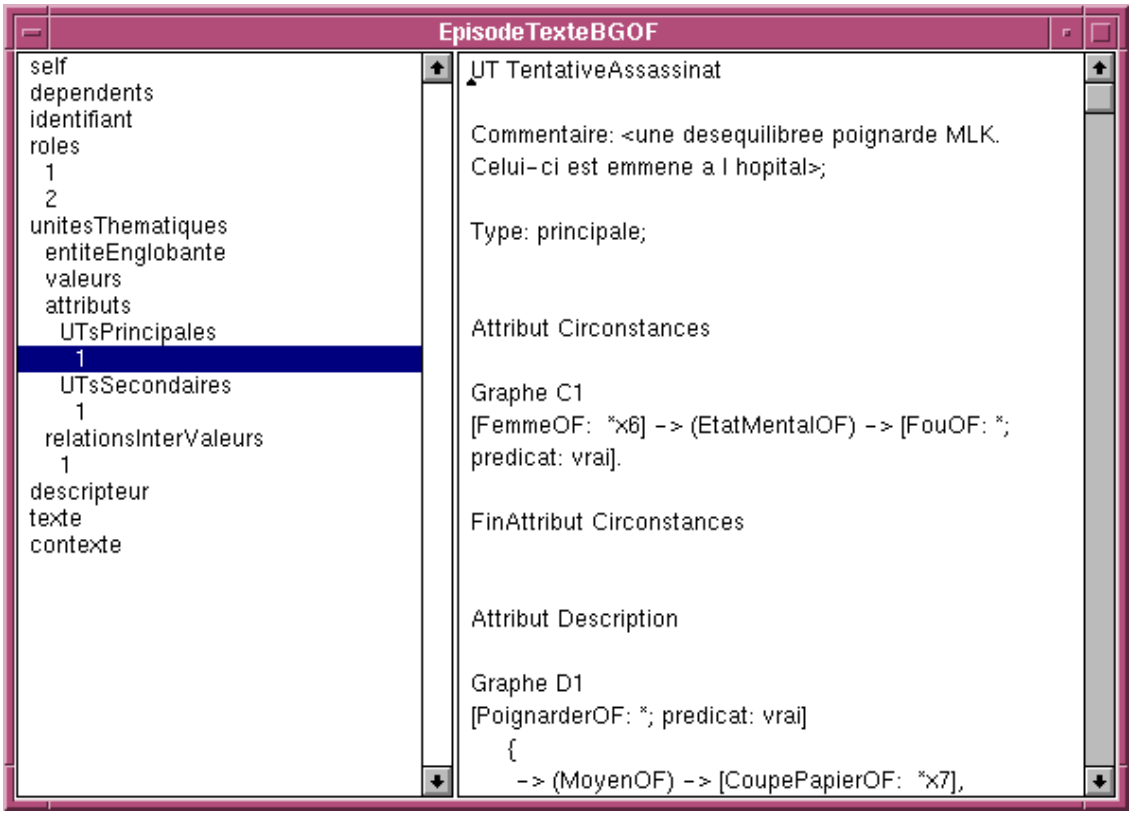


Fig. C.2 - Inspecteur spécialisé dédié aux épisodes

- lancement d'un inspecteur spécialisé sur une représentation spécifique d'un texte. En l'occurrence, nous ne disposons que d'un inspecteur spécialisé dédié aux représentations situationnelles des textes. La figure C.2 en donne un aperçu. Il étend les inspecteurs présents par défaut dans l'environnement Smalltalk en faisant apparaître de façon directement accessible toute la structure d'un épisode, ce qui permet d'accéder rapidement à l'un de ses éléments et de l'inspecter, quel que soit son degré d'enchâssement.

  Le même type d'inspecteur, illustré par la figure C.3, existe pour les Unités Thématiques.

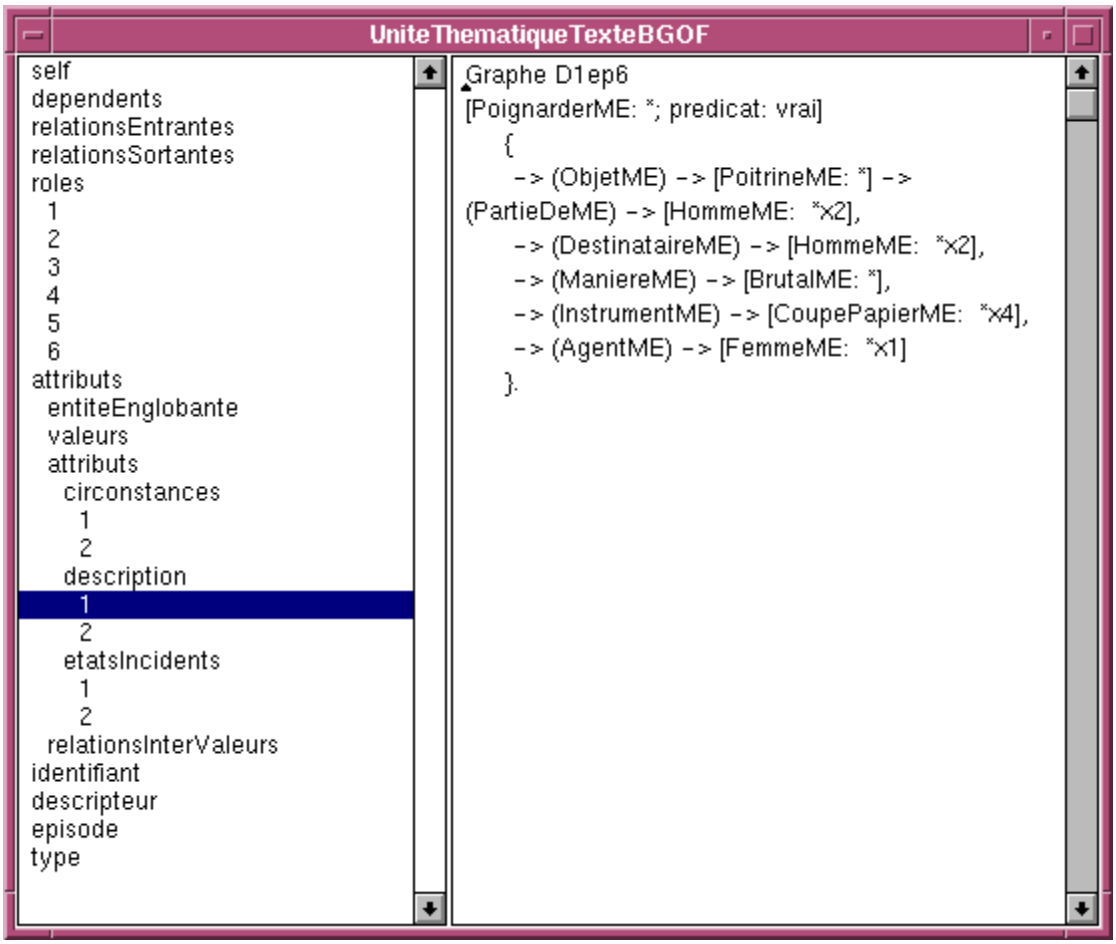


Fig. C.3 - Inspecteur spécialisé dédié aux Unités Thématiques

Annexe D

Exemples d'unités thématiques

Épisode 1
Commentaire: un soldat blesse un chef d'état à son arrivée à l'aéroport

UT ep1
Type: principale;

Attribut Circonstances

Graphe C1
[Être_localisé; prédicat: vrai]
{→ (objet) → [Événement: *x1],
→ (lieu) → [Aéroport: *x2]
}.

FinAttribut Circonstances

Attribut Description

Graphe D1
[Attaquer; prédicat: vrai]
{→ (agent) → [Soldat: *x3],
→ (patient) → [Chef_d'état: *x4],
→ (manière) → [Soudain]
}.

Graphe D2
[Trébucher; prédicat: vrai] → (agent) → [Soldat: *x3].

Graphe D3
[Poignarder; prédicat: vrai]
{→ (agent) → [Soldat: *x3],
→ (destinataire) → [Chef_d'état: *x4],
→ (objet) → [Bras: *x5] → (partie_de) → [Chef_d'état: *x4],
→ (instrument) → [Baïonnette: *x6]
}.

Graphe D4
[Arrêter; prédicat: vrai]
{→ (agent) → [Policier: *x7],
→ (patient) → [Soldat: *x3]
}.

FinAttribut Description

Attribut ÉtatsIncidents

Graphe E1
[Être_blessé; prédicat: vrai]
{→ (source) → [Chef_d'état: *x4],
→ (manière) → [Léger]
}.

Graphe E2
[Être_emprisonné; prédicat: vrai]→ (source) → [Soldat: *x3].

FinAttribut ÉtatsIncidents

RelationsIntraUT

Relation D4E2
Origine: analyseCausale;
Description.D4 → ÉtatsIncidents.E2: causalité.

Relation D3E1
Origine: analyseCausale;
Description.D3 → ÉtatsIncidents.E1: causalité.

FinRelationIntraUTs

FinUT ep1

FinÉpisode 1

Épisode 2
Commentaire: un jeune homme en poignarde un autre à la suite d'une dispute

UT ep2
Type: principale;

Attribut Circonstances

Graphe C1
[SeQuereller; prédicat: vrai]
{→ (agent) → [Jeune_homme: *x1],
→ (co-agent) → [Jeune_homme: *x2],
→ (objet) → [Argent]
}.

FinAttribut Circonstances

Attribut Description

Graphe D1
[Frapper; prédicat: vrai]
{→ (agent) → [Jeune_homme: *x2],
→ (patient) → [Jeune_homme: *x1]
}.

Graphe D2
[Poignarder; prédicat: vrai]
{→ (agent) → [Jeune_homme: *x1],
→ (objet) → [Ventre] → (partie_de) →
[Jeune_homme: *x2],
→ (destinataire) → [Jeune_homme: *x2],
→ (instrument) → [Cran_d'arrêt: *x3]
}.

Graphe D3
[Arrêter; prédicat: vrai]
{→ (agent) → [Humain: *x4],
→ (patient) → [Jeune_homme: *x1]
}.

FinAttribut Description

Attribut ÉtatsIncidents

Graphe E1
[Être_mort; prédicat: vrai] → (source) →
[Jeune_homme: *x2].

Graphe E2
[Être_emprisonné; prédicat: vrai] → (source) →
[Jeune_homme: *x1].

FinAttribut ÉtatsIncidents

RelationsIntraUT

Relation D3E2
Origine: analyseCausale;
Description.D3 → ÉtatsIncidents.E2: causalité.

Relation D2E1
Origine: analyseCausale;
Description.D2 → ÉtatsIncidents.E1: causalité.

FinRelationIntraUTs

FinUT ep2

FinÉpisode 2

Épisode 3
Commentaire: un assassinat politique

UT ep3
Type: principale;

Attribut Circonstances

Graphe C1
[Menacer; prédicat: vrai]
{→ (agent) → [Homme_politique: *x1],
→ (patient) → [Homme_politique: {*} *x2]
}.

Graphe C2
[Croire; prédicat: vrai]
{→ (agent) → [Femme: *x3],
→ (objet) → [Idée: {*} *x4]
}.

Graphe C3
[Soutenir; prédicat: vrai]
→ (agent) → [Homme_politique: {*} *x2],
→ (objet) → [Idée: {*} *x4]
}.

Graphe C4
[Habiter; prédicat: vrai]
{→ (agent) → [Homme_politique: *x1],
→ (lieu) → [Appartement: *x5]
}.

FinAttribut Circonstances

Attribut Description

Graphe D1
[Pénétrer; prédicat: vrai]
{→ (agent) → [Femme: *x3],
→ (destination) → [Appartement: *x5],
→ (manière) → [Clandestin]
}.

Graphe D2
[SeBaigner; prédicat: vrai]
{→ (agent) → [Homme_politique: *x1],
→ (lieu) → [Baignoire: *x6]
}.

Graphe D3
[Poignarder; prédicat: vrai]
{→ (agent) → [Femme: *x3],
→ (destinataire) → [Homme_politique: *x1],
→ (instrument) → [Couteau: *x7]
}.

Graphe D4
[Arrêter; prédicat: vrai]
→ (agent) → [Policier: *x8],
→ (patient) → [Femme: *x3]
}.

FinAttribut Description

Attribut ÉtatsIncidents

Graphe E1
[Être_mort; prédicat: vrai] → (source) →
[Homme_politique: *x1].

Graphe E2
[Être_guillotiné; prédicat: vrai] → (source) →
[Femme: *x3].

FinAttribut ÉtatsIncidents

RelationsIntraUT

Relation D3E1
Origine: analyseCausale;
Description.D3 → ÉtatsIncidents.E1: causalité.

FinRelationIntraUTs

FinUT ep3

FinÉpisode 3

Épisode 4
Commentaire: l'assassinat du commandant d'une armée à l'issue d'une défaite

UT ep4
Type: principale;

Attribut Circonstances

Graphe C1
[Commander; prédicat: vrai]
{→ (agent) → [Homme: *x1],
→ (objet) - > [Armée: *x2]
}.

FinAttribut Circonstances

Attribut Description

Graphe D1
[Perdre; prédicat: vrai]
{→ (agent) → [Armée: *x2],
→ (objet) → [Bataille: *x3]
}.

Graphe D2
[Poignarder; prédicat: vrai]
{→ (agent) → [Homme: *x4],
→ (destinataire) → [Homme: *x1],
→ (instrument) → [Épée: *x5]
}.

FinAttribut Description

Attribut ÉtatsIncidents

Graphe E1
[Être_mort; prédicat: vrai] → (source) →
[Homme: *x1].

FinAttribut ÉtatsIncidents

RelationsIntraUT

Relation D2E1
Origine: analyseCausale;
Description.D2 → ÉtatsIncidents.E1: causalité.

FinRelationIntraUTs

FinUT ep4

FinÉpisode 4

Épisode 5
Commentaire: l'agression d'une femme seule par un maniaque

UT ep5
Type: principale;

Attribut Circonstances

Graphe C1
[Dormir; prédicat: vrai]
{→ (agent) → [Femme: *x1],
→ (temps) → [Nuit: *x2]
}.

Graphe C2
[Habiter; prédicat: vrai]
{→ (agent) → [Femme: *x1],
→ (lieu) → [Maison: *x3]
}.

FinAttribut Circonstances

Attribut Description

Graphe D1
[Pénétrer; prédicat: vrai]
{→ (agent) → [Homme: *x4],
→ (destination) → [Maison: *x3],
→ (manière) → [Par_effraction]
}.

Graphe D2
[Attacher; prédicat: vrai]
{→ (agent) → [Homme: *x4],
→ (patient) → [Femme: *x1]
}.

Graphe D3
[Déchirer; prédicat: vrai]
{→ (agent) → [Homme: *x4],
→ (objet) → [Chemise_de_nuit: *x5]
}.

Graphe D4
[Poignarder; prédicat: vrai]
{→ (agent) → [Homme: *x4],
→ (destinataire) → [Femme: *x1],
→ (instrument) → [Couteau_de_chasse: *x6],
→ (manière) → [Sauvage]
}.

FinAttribut Description

Attribut ÉtatsIncidents

Graphe E1
[Être_blessé; prédicat: vrai] → (source) →
[Femme: *x1].

FinAttribut ÉtatsIncidents

RelationsIntraUT

Relation D4E1
Origine: analyseCausale;
Description.D4 → ÉtatsIncidents.E1: causalité.

FinRelationIntraUTs

FinUT ep5

FinÉpisode 5

Annexe E

Un environnement de test des réseaux à propagation d'activité : MALCOM

Cette annexe est destinée à donner un aperçu rapide de l'environnement de test des réseaux propagation d'activité que nous avons construit afin de mettre au point le mécanisme de rappel de la mémoire épisodique décrit au chapitre 6.

1. Structure et activité d'un réseau à propagation d'activité dans MALCOM

1.1. Structure

Dans MALCOM, un réseau à propagation d'activité se compose de quatre types d'objets différents :

- la *couche*. La notion de couche dans MALCOM ne possède pas d'autre mission que de structurer l'ensemble des unités composant un réseau en fonction de ce qu'elles représentent. Dans l'exemple illustré par la figure E.2, les unités font référence à un ensemble de schémas et de concepts en inter-relation. Le réseau est donc divisé en deux couches : l'une contient les unités représentant des schémas tandis que l'autre contient les unités représentant des concepts.
  En aucun cas, la notion de couche n'implique que les unités d'une couche ne doivent avoir que des relations vers des unités de la même couche. Dans le même esprit, une relation unissant deux unités appartenant à deux couches distinctes n'est pas différente d'une relation unissant deux unités appartenant à la même couche. En revanche, la couche détermine le type des unités qui la constitue. Ce type est unique pour toutes les unités de la couche. Il définit la fonction d'activation de chaque unité, c'est-à-dire la fonction calculant le niveau d'activité de l'unité en fonction de la valeur de ses entrées ;

- l'*unité*. L'unité est l'élément constitutif principal d'un réseau à propagation d'activité. Elle est chargée de représenter les entités concernées par la propagation d'activité et de porter en particulier le niveau d'activité que possède celles-ci. Ainsi que nous l'avons

indiqué lors de la présentation de la notion de couche, chaque unité possède un type, déterminant la fonction d'activation que cette unité utilise pour évaluer à tout moment son niveau d'activité ;

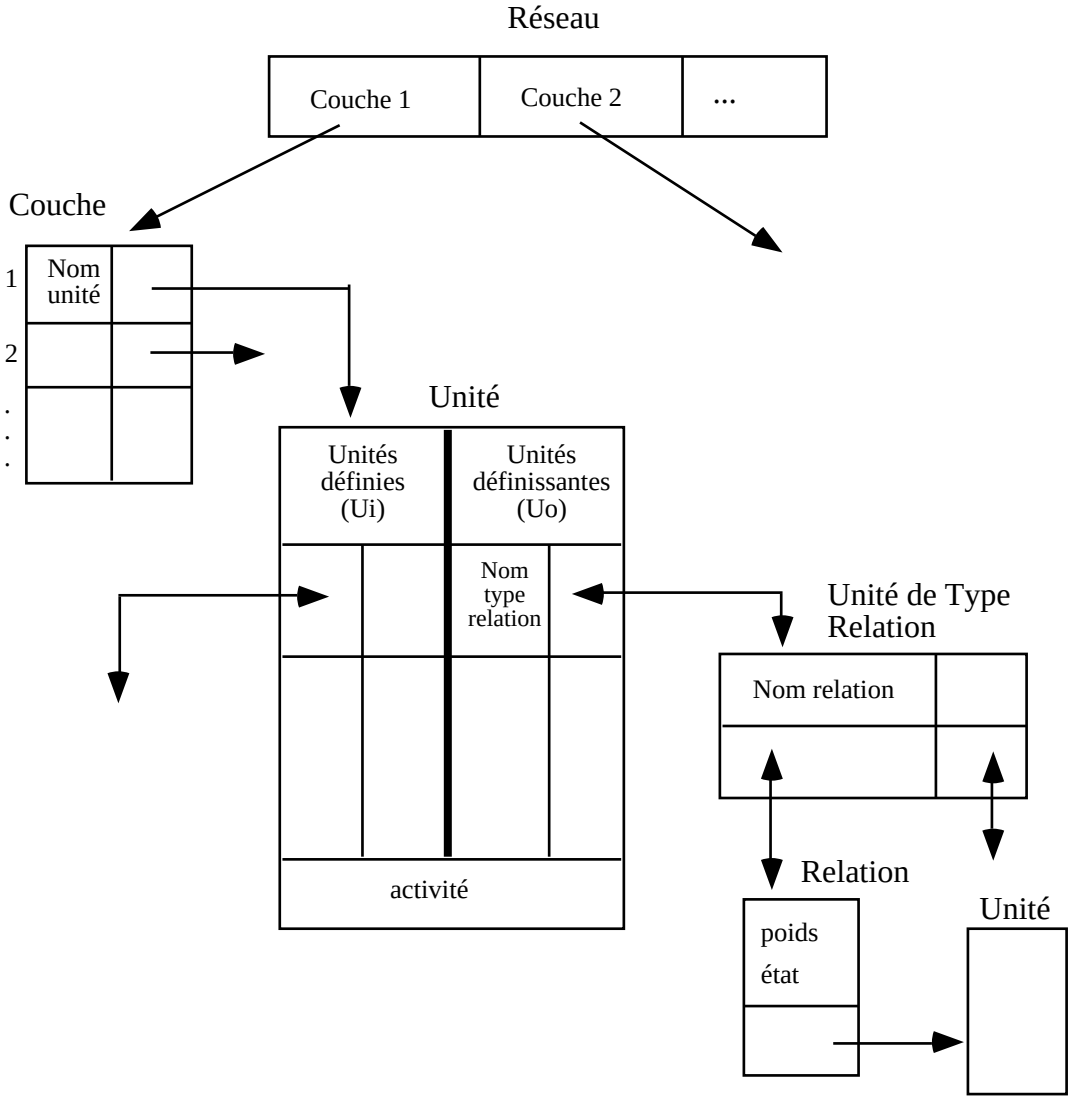


Fig. E.1 - Structure d'un réseau à propagation d'activité

- l'*unité type de relation*. L'unité type de relation est un élément de structuration apportée aux relations qu'une unité entretient avec d'autres unités du réseau. Son rôle est de servir de point de convergence à toutes les relations d'un certain type impliquant une unité donnée. À chaque type de relation est associée une fonction d'activation, qui est appliquée au niveau de toutes les unités type de relation relevant de ce type. Cette fonction d'activation représente une sorte de pré-traitement des entrées opérant sur une partie du flux d'activité parvenant à une unité. Dans un réseau comme celui décrit dans [Lange & Dyer 1989], ce type d'unité sert par exemple à moyenner l'activité provenant

des relations d'un même type, afin d'éviter qu'une différence du nombre de relations ne défavorise un type de relation par rapport à un autre.
Les réseaux considérés ici sont délibérément influencés par la relation de composition du fait des entités qu'ils représentent (cf. notamment les UTs agrégées). C'est pourquoi toutes les unités type de relation d'une unité n'ont pas le même statut. On distingue celles intervenant dans la définition de l'unité, allant vers des unités dites définissantes, i.e. des composants, (cf. Uo de la figure E.1), et celles traduisant l'utilisation de l'unité pour en définir d'autres, appelées unités définies, i.e. composés, (cf. Ui de la figure E.1). Cette distinction n'a pas d'influence à proprement parler au niveau de la propagation d'activité. Lorsque l'activité d'une unité est réévaluée, toutes ses relations sont en effet considérées comme des entrées. Elle intervient en revanche lors de la construction du réseau. L'ajout d'un composant, représenté par l'unité Ua, à une entité représentée par l'unité Ue provoque en effet la création d'une relation *Rea* allant des unités définissantes (Uo) de Ue vers Ua et d'une relation *Rae* allant des unités définies (Ui) de Ua vers Ue ;

- la *relation.* Contrairement à l'unité type de relation, la relation est un objet partagé entre l'unité source de la relation et son unité cible. Elle est le support servant de passage à l'activité d'une unité à l'autre. Elle est également le lieu où cette activité peut être modulée, voire bloquée. La modulation repose sur la présence d'un poids. Le blocage est conditionné quant à lui par un indicateur d'état de la relation. Celui-ci permet en pratique de définir le sens de parcours de l'activité. En reprenant l'exemple donné ci-dessus à propos des unités type de relation, si l'on souhaite que l'activité ne puisse aller que de Ua vers Ue, on bloquera l'activité au niveau de la relation *Rae* tandis qu'on la laissera libre de passer au niveau de la relation *Rea.*

1.2. Activité

Le mode de fonctionnement premier des réseaux à propagation d'activité mis en œuvre par MALCOM est le mode de fonctionnement synchrone : à chaque cycle, l'activité de l'ensemble des unités actives du réseau est réévaluée en fonction de l'activité des unités actives au cycle précédent. Par défaut, les unités actives rassemblent l'ensemble des unités du réseau. On peut néanmoins restreindre cet ensemble en fonction de critères spécifiques. Cette restriction s'opère via le marquage des unités. Suivant les cas, ce marquage peut être statique, c'est-à-dire ne pas changer durant l'ensemble des cycles constituant une phase de propagation ou bien être modifié au cours d'une même phase. Le premier cas correspond à la phase de sélection dans MLK : l'activation ne touche que la partie de la mémoire délimitée par la première phase de sélection. Cette dernière constitue

pour sa part un exemple du second cas : à chaque cycle, les unités actives se restreignent aux unités touchant par le flux d'activité au cycle précédent.

2. Outils de MALCOM

Les outils de MALCOM s'articulent autour du gestionnaire de réseau à propagation d'activité illustré par la figure E.2. Celui-ci permet d'avoir une vue analytique des différentes couches d'un réseau, des unités qui les composent ainsi que des relations qu'elles entretiennent. On peut ainsi visualiser l'activité d'une unité ou bien les caractéristiques propres à une relation. Ce gestionnaire est en outre responsable des opérations de création et de suppression des réseaux ou d'une partie d'entre eux. Il assure en particulier le passage d'une représentation sous forme d'objets Smalltalk à une représentation sous forme texte et inversement : création d'un réseau à partir d'un fichier texte, sauvegarde de la structure d'un réseau sous forme texte, sauvegarde de l'activité d'un réseau, restitution de cette activité.

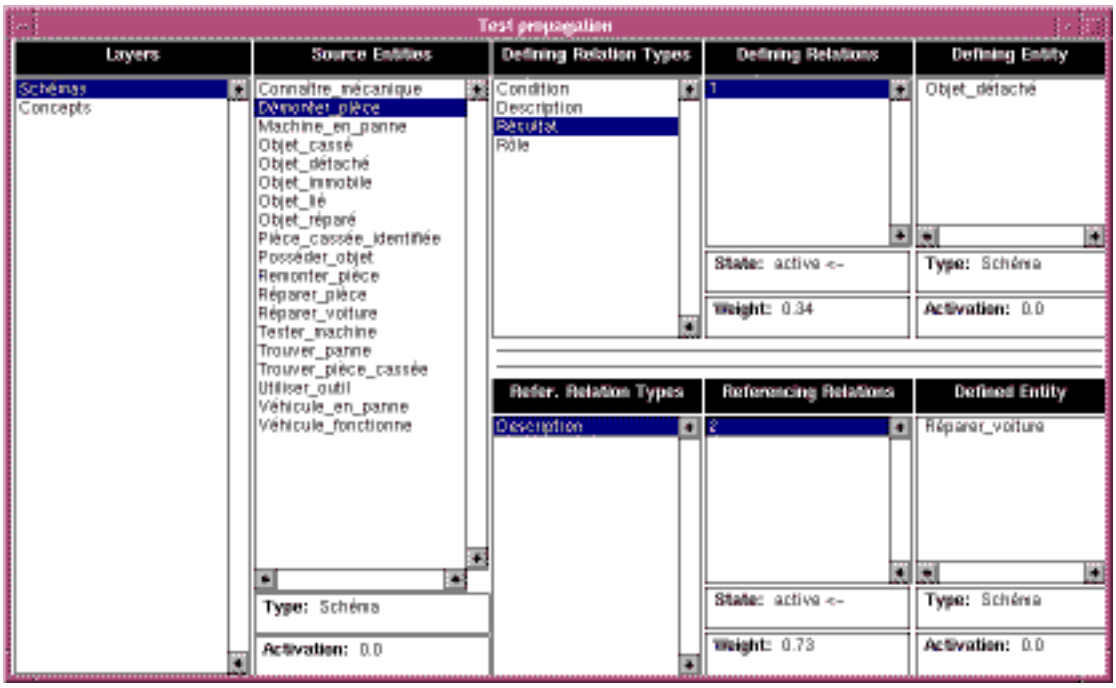


Fig. E.2 - Gestionnaire d'un réseau à propagation d'activité

Format d'un fichier de définition de la structure d'un réseau à propagation d'activité

```
// Fichier de définition de la structure d'un réseau à propagation d'activité
// 20 May 1997 6:27:50 pm

--Couches

--Couche Schémas
--TypeCouche schémas
--TypeUnité schéma
```

```
réparer_voiture
trouver_panne
démonter_pièce
réparer_pièce
        …
objet_lié
--FinCouche Schémas

--Couche Concepts
--TypeCouche concepts
--TypeUnité concept
réparer
personne
voiture
outil
        …
                …
moyen_de_transport
universel
domaine
--FinCouche Concepts

--FinCouches

--TypesRelation
TypeRelationBase condition
TypeRelationBase description
TypeRelationBase résultat
TypeRelationBase rôle
TypeRelationBase sorte_de
TypeRelationBase partie_de
--FinTypesRelation

--Connexions
// Schéma réparer_voiture

// Rôle
RelationBase UU réparer_voiture rôle prédicat réparer :: -0.58 :: -0.31
RelationBase OO réparer_voiture rôle agent personne :: 0.44 :: 0.47
RelationBase UU réparer_voiture rôle objet voiture :: 0.14 :: -0.57
RelationBase OO réparer_voiture rôle instrument outil :: 0.78 :: -0.47
RelationBase OU réparer_voiture rôle lieu local :: 0.03 :: -0.95

// Description
RelationBase UU réparer_voiture description 1 trouver_panne :: 0.98 :: 0.94
RelationBase UU réparer_voiture description 2 démonter_pièce :: 0.17 :: 0.73
RelationBase UU réparer_voiture description 3 réparer_pièce :: -0.45 :: -0.55
RelationBase OO réparer_voiture description 4 remonter_pièce :: 0.52 :: -0.9

// Résultat
RelationBase OU réparer_voiture résultat 1 véhicule_fonctionne :: -0.1 :: -0.82

// Condition
RelationBase UU réparer_voiture condition 1 véhicule_en_panne :: -0.08 :: 0.01
RelationBase UO réparer_voiture condition 2 connaître_mécanique :: 0.25 :: 0.71
        …
// Relations au sein des concepts

RelationBase OU réparer sorte_de 1 action :: 0.13 :: 0.84
RelationBase OU personne sorte_de 1 être_animé :: -0.22 :: -0.22
RelationBase OU voiture sorte_de 1 véhicule :: 0.5 :: -0.68
        …
RelationBase OU domaine sorte_de 1 universel :: -0.7 :: -0.37
RelationBase UU objet_fabriqué sorte_de 1 objet :: -0.6 :: -0.52
--FinConnexions

--FonctionsActivation

--Unités
Schéma identityWithThreshold 1 -1
Concept identityWithThreshold 1 -1
--FinUnités

--TypesRelation
ParDéfaut average
--FinTypesRelation

--FinFonctionsActivation
```

La définition de la structure d'un réseau à propagation d'activité se décompose en trois grandes sections :

- une section de définition des différentes couches du réseau ainsi que des unités constituant ces couches. Dans le cas présent, le réseau dispose de deux couches : une

première contenant des unités de type schéma (réparer_voiture, objet_lié, etc.) ; une seconde contenant des unités de type concept (réparer, outil, etc.). Cette section liste également l'ensemble des types de relation utilisés ;

- une section de définition des relations existant entre les unités composant le réseau. Chaque relation (ou plus exactement chaque couple de relations) est définie par la séquence : code du sens de parcours de la relation (UU, OU, UO ou OO)[1], nom de l'unité définie, nom du type de relation, nom de la relation (les deux relations effectivement créées portent le même nom), nom de l'unité définissante, poids de la relation *Rea* (de l'unité définie vers l'unité définissante), poids de la relation *Rae* (de l'unité définissante vers l'unité définie) ;

- une section de définition des fonctions d'activation des différents types d'unités ainsi que des différents types d'unités type de relation. Pour chaque fonction d'activation, on peut préciser si nécessaire une liste de paramètres.

Format d'un fichier de définition de l'activité d'un réseau à propagation d'activité

```
// Fichier de définition de l'activation d'un réseau à propagation d'activité
// 20 May 1997 4:57:17 pm

// Activation des schémas
réparer_voiture :: 3.3937761783599853515e-4
trouver_panne :: -5.8332371711730957031e-4
démonter_pièce :: -3.3965647220611572265e-4
réparer_pièce :: 1.4993740081787109375e-4
remonter_pièce :: -1.7330313682556152343e-7
véhicule_fonctionne :: 0.0
véhicule_en_panne :: 4.8849253654479980468e-4
        ...

// Activation des concepts
réparer :: 1.1240582466125488281e-4
personne :: -2.7014610767364501953e-8
voiture :: -9.1173429489135742187e-5
outil :: 0.0
local :: 0.0
etre_en_panne :: -6.8871850967407226562e-5
véhicule :: 3.2694694995880126953e-4
        ...
```

Chaque unité étant désignée par un nom, la structure d'un fichier de définition de l'activité d'un réseau à propagation d'activité est très simple. Elle se compose en effet d'un ensemble de couples :

nom_d'unité :: valeur_d'activité

Outre les fonctions énumérées ci-dessus, le gestionnaire de la figure E.2 a aussi pour rôle d'être le point d'accès aux autres outils de MALCOM. Ceux-ci s'articulent autour de deux fonctions principales : d'une part, la visualisation globale de la structure d'un réseau à propagation d'activité ; d'autre part, le lancement de la propagation d'activité et le suivi de l'état d'activité du réseau. La première fonction est assumée en tant que telle par

[1] La première lettre représente l'unité définie tandis que la seconde représente l'unité définissante. Le code 'U' indique que l'unité considérée émet effectivement de l'activité vers son homologue. En revanche, un 'O' signifie une absence d'émission d'activité (blocage).

le visualisateur de la figure E.3. Compte tenu de l'importance de la relation de composition et de la notion de hiérarchie dans les réseaux considérés, cette visualisation se fait sous la forme d'un arbre, avec une notation des coréférences faisant apparaître les manifestations d'une même unité dans différents sous-arbres. La figure E.4 montre le

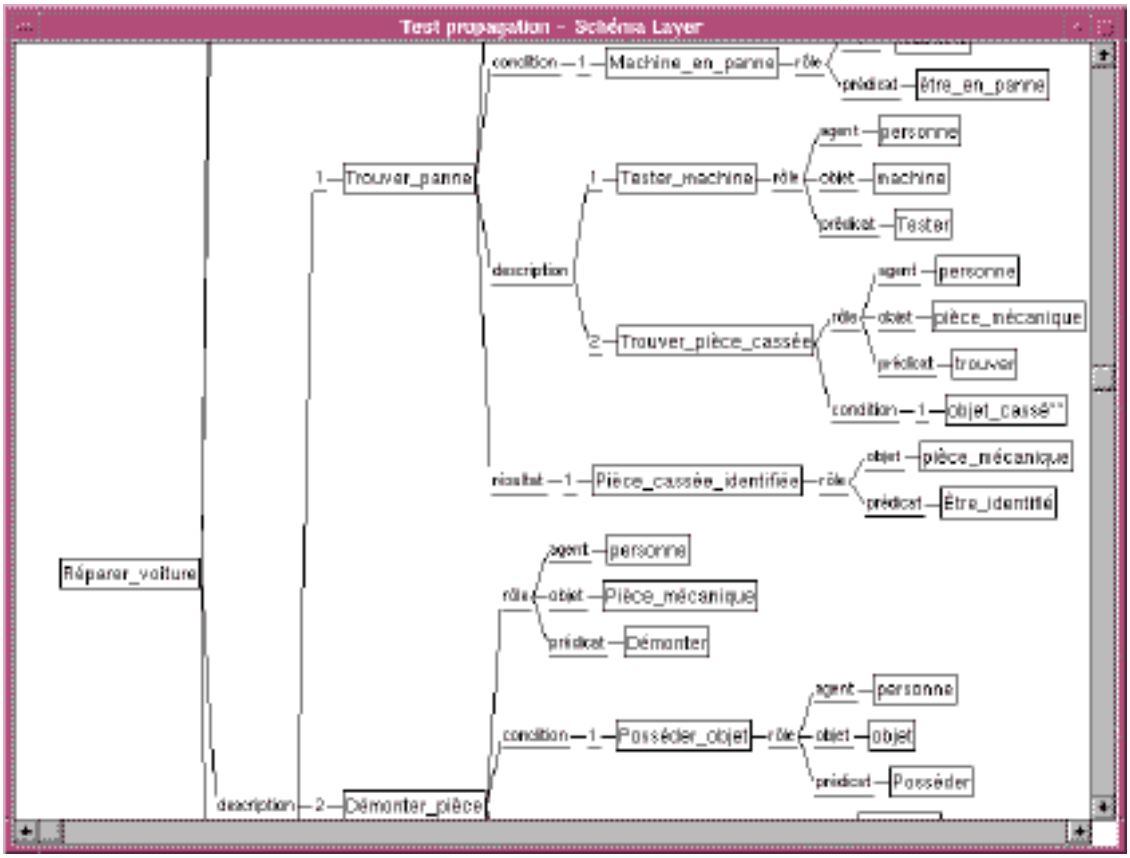


Fig. E.3 - Visualisateur de la structure d'un réseau de propagation d'activité

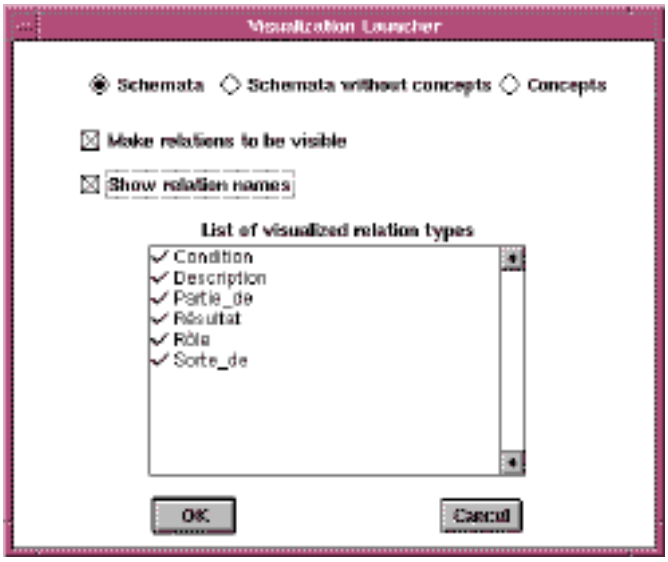


Fig. E.4 - Interface de paramétrage de la visualisation de la structure d'un réseau

panneau de configuration permettant de spécifier les entités que l'on souhaite voir présentes au niveau de cette visualisation.

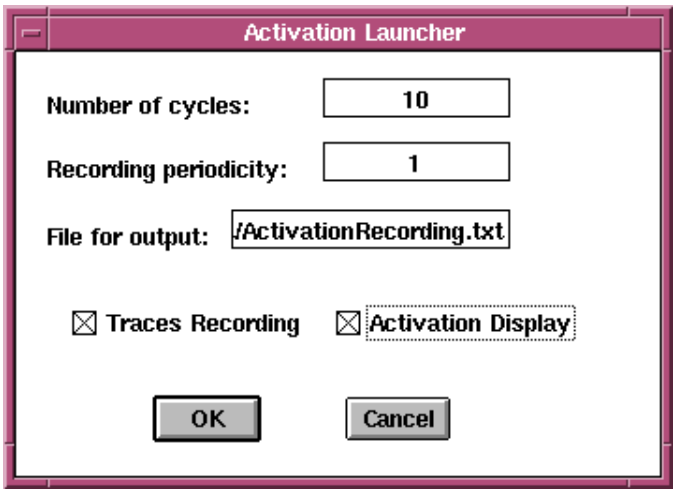


Fig. E.5 - Interface de paramétrage d'une session de propagation d'activité

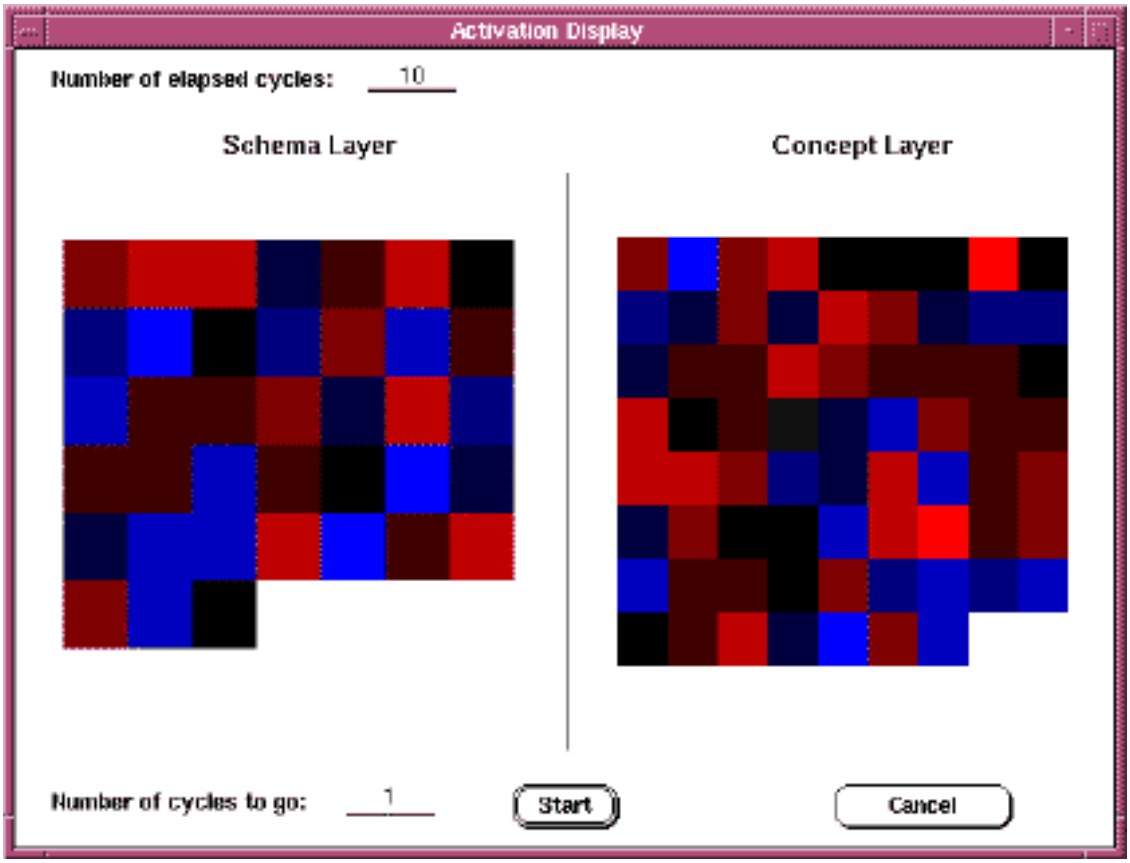


Fig. E.6 - Visualisateur de l'activité d'un réseau

Le lancement de la propagation d'activité dans un réseau s'effectue quant à lui par l'intermédiaire soit de l'interface de lancement de la figure E.4, soit du visualisateur de la

figure E.5. Le premier mode de lancement est plutôt dédié à une propagation sur un grand nombre de cycles. Dans cette optique, il offre la possibilité d'enregistrer l'état d'activité du réseau dans un fichier texte suivant une périodicité définie par l'utilisateur. Le second mode est pour sa part plutôt orienté vers un suivi pas à pas de l'activité du réseau (le pas est là aussi définissable par l'utilisateur), suivi rendu possible par une visualisation graphique de cette activité.

Annexe F

Interfaces de la mémoire épisodique

Cette annexe expose un ensemble d'interfaces permettant d'interagir avec la mémoire épisodique sans avoir recours à la programmation. La première d'entre elles, illustrée par la figure F.1, offre la possibilité de contrôler l'ensemble du processus d'intégration d'une nouvelle UT textuelle au sein de la mémoire. Pour une UT textuelle donnée, elle visualise de façon synthétique son niveau de similarité avec chacune des UTs agrégées constituant la mémoire. Elle montre le niveau de similarité par attribut et met en évidence les règles de similarité applicables. Dans le cas de la figure F.1 par exemple, on constate ainsi que l'UT textuelle *TentativeAssassinatMLK* et l'UT agrégée *AssassinatChefDEtat* ont des attributs *Circonstances* et *ÉtatsIncidents* fortement similaires et des attributs

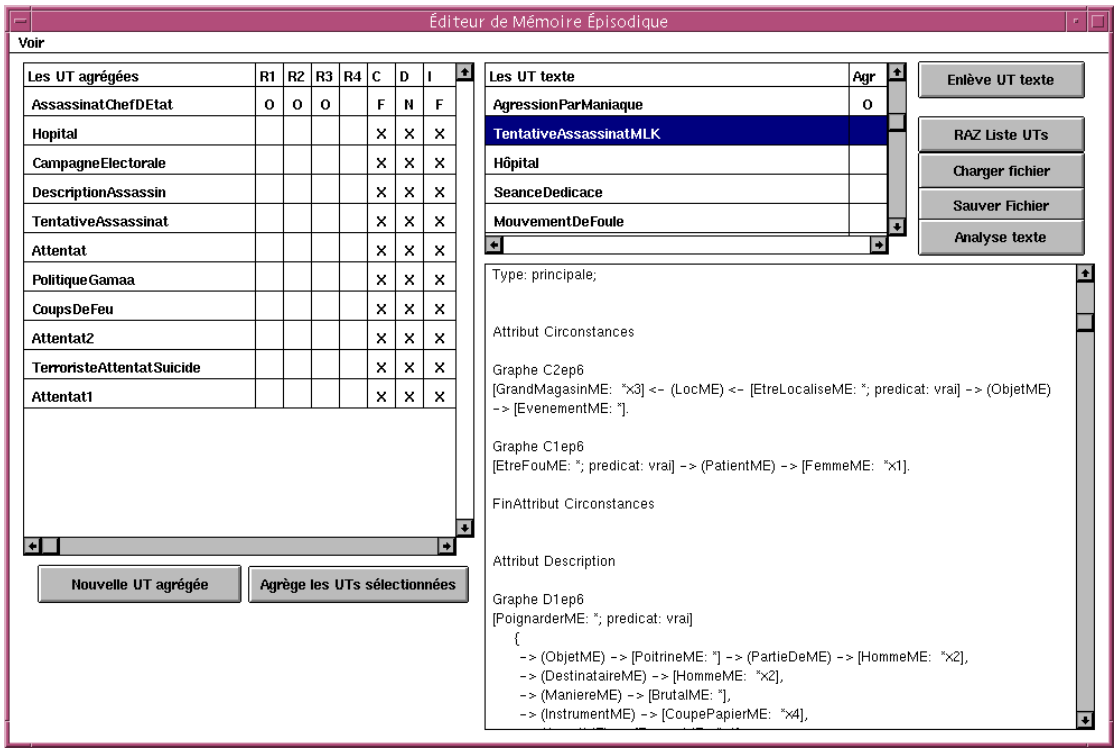


Fig. F.1 - Outil d'évaluation de la similarité des Unités Thématiques (UTs) et de contrôle de l'intégration dans la mémoire épisodique de nouvelles UTs[1]

[1] Les interfaces apparaissant dans les figures F.1 à F.4 ont été développées par Gaël de Chalendar dans le cadre de son travail de DEA.

Description présentant une similarité normale. Les règles R1, R2 et R3 (cf. §4.2.2 du chapitre 6) sont donc applicables. En revanche, aucun des attributs de cette même UT textuelle n'est similaire à l'attribut correspondant d'une autre UT agrégée de la mémoire.

L'interface de la figure F.1 permet également de définir à quelle UT agrégée une nouvelle UT textuelle doit être agrégée, celle-ci pouvant aussi être mémorisée en tant que nouvelle UT agrégée.

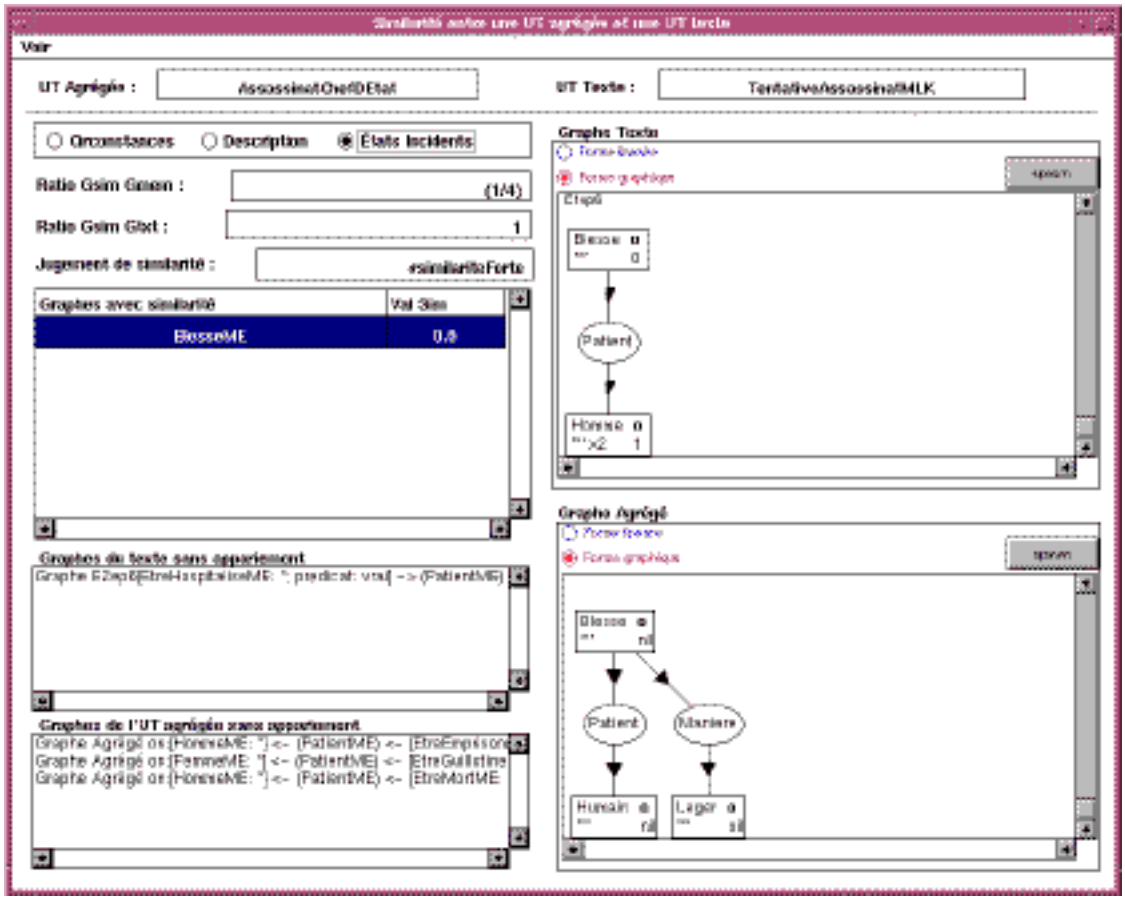


Fig. F.2 - Interface détaillant la similarité d'une UT textuelle et d'une UT agrégée

Si l'on souhaite examiner de plus près la similarité d'une UT textuelle et d'une UT agrégée, on peut lancer l'interface de la figure F.2 à partir de celle de la figure F.1. Il est alors possible de voir pour chaque attribut quels sont les graphes similaires et quels sont les graphes non similaires de part et d'autre. Plus globalement, on peut aussi visualiser la valeur des ratios de graphes similaires, lesquels déterminent le niveau de similarité des attributs. Les graphes sont affichés au choix sous forme graphique ou sous forme linéaire.

La mémoire épisodique possède également des outils de visualisation indépendants des fonctions d'intégration des nouvelles UTs textuelles. La figure F.3 montre ainsi

l'interface d'un visualisateur donnant une vue globale de l'ensemble des UTs agrégées de la mémoire et des relations thématiques que celles-ci entretiennent entre elles.

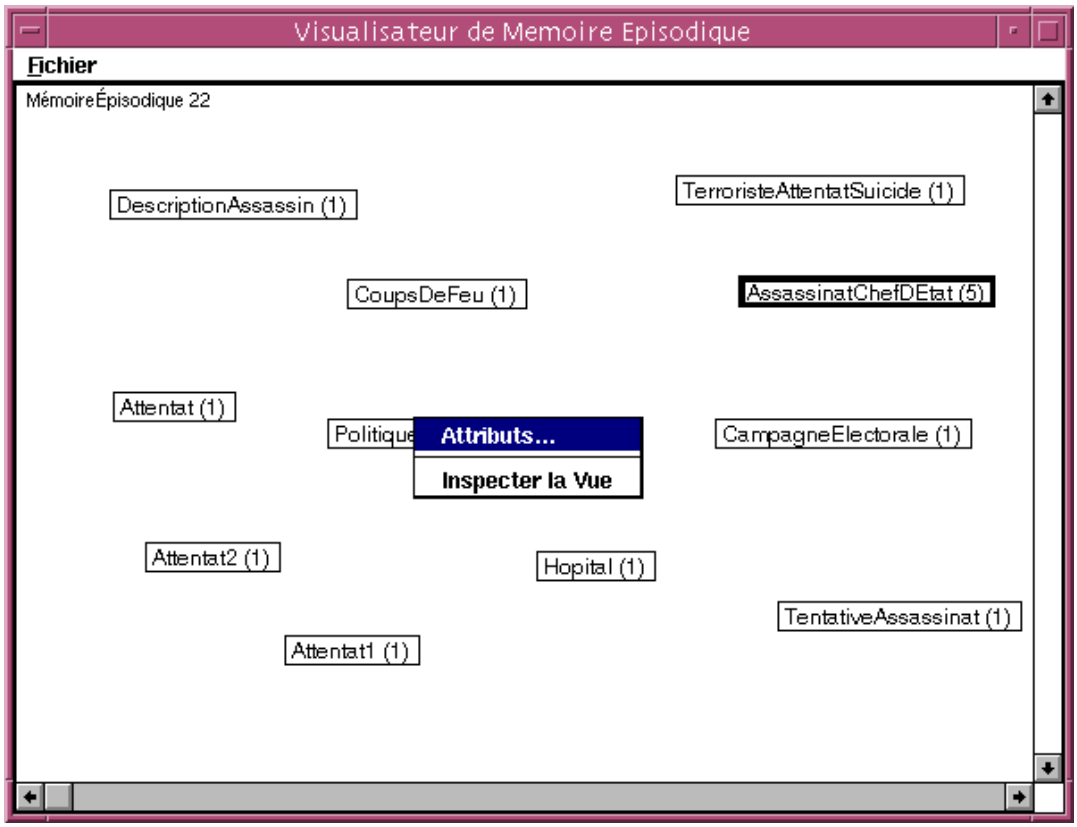


Fig. F.3 - Visualisateur de mémoire épisodique

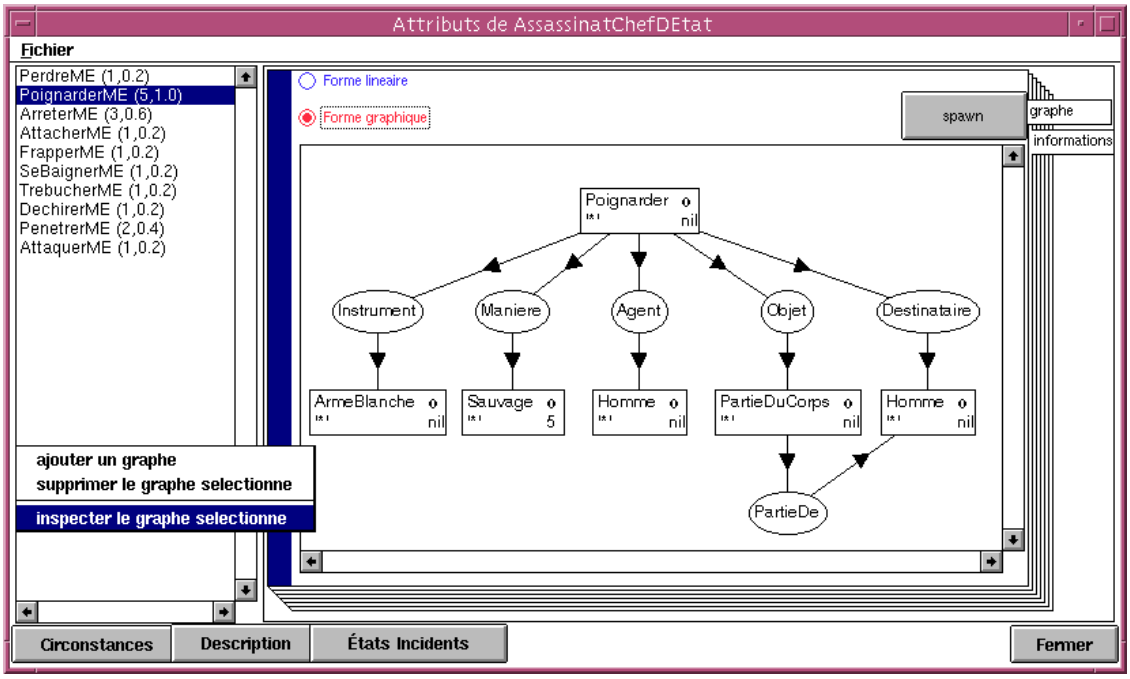


Fig. F.4 - Visualisateur d'une UT agrégée

À partir de cette vue globale, chaque UT agrégée peut être inspectée de façon plus précise. L'interface de la figure F.4 permet d'examiner le contenu d'une UT agrégée en faisant apparaître pour chacun de ses attributs les graphes agrégés qui le composent. Ces graphes sont là aussi visualisables sous forme graphique ou sous forme linéaire. L'inspecteur spécialisé de la figure F.5 offre quant à lui un accès complet, bien que plus orienté vers le programmeur, à tous les constituants d'une UT agrégée.

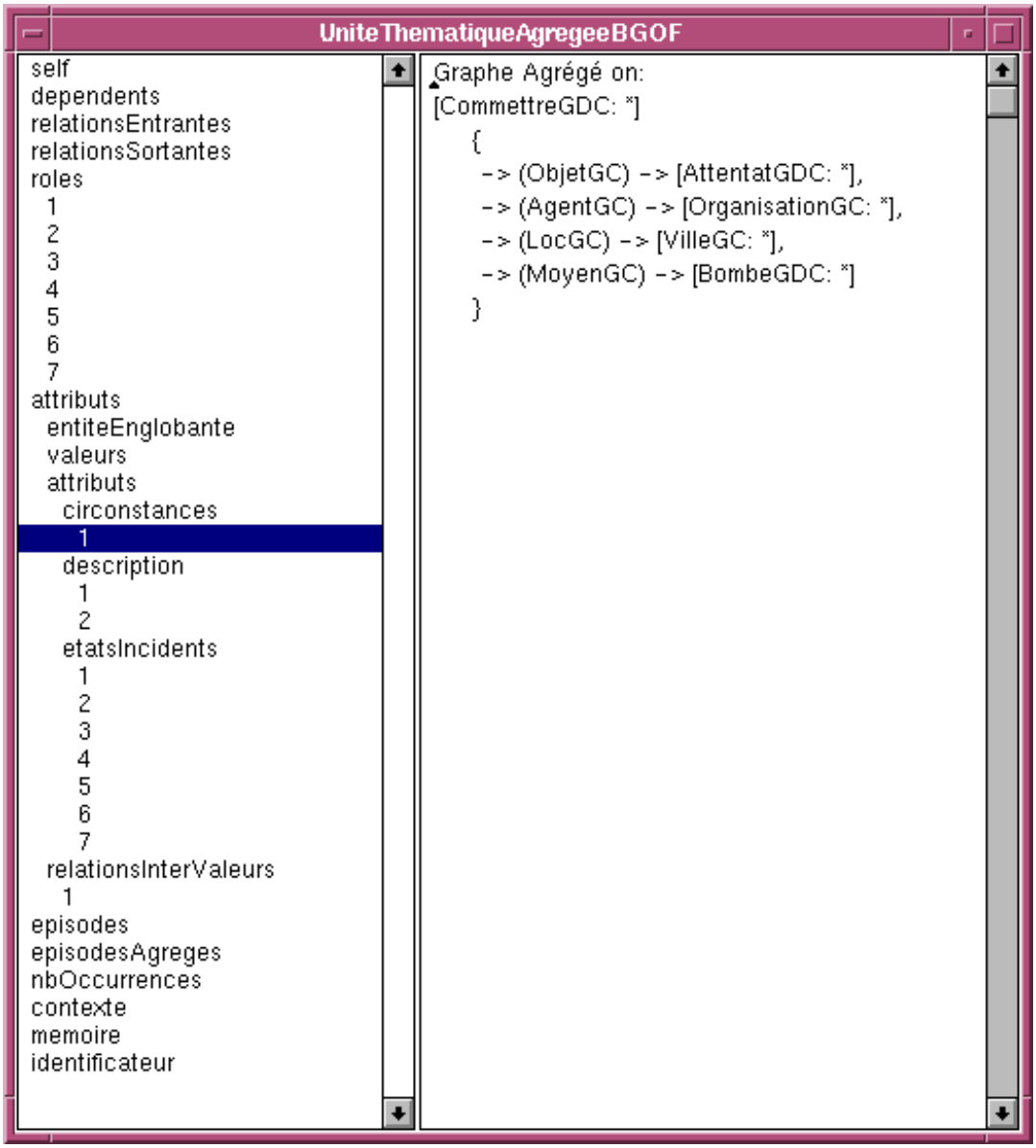


Fig. F.5 - Inspecteur spécialisé dédié aux UTs agrégées

Annexe G

Formats et pré-traitement des textes

1. Corpus

Le texte ci-dessous illustre la forme d'origine des textes traités, aussi bien lorsqu'ils proviennent du journal *Le Monde* que lorsqu'il s'agit de dépêches de l'AFP, comme c'est le cas ici. Un ensemble de balises de type SGML permettent de délimiter les paragraphes, de repérer le titre de chaque article et d'apporter des informations de nature diverse sur ce dernier ou ses constituants : l'identificateur de l'article, sa date de parution, le numéro des paragraphes, éventuellement l'auteur de l'article, etc. On pourra trouver davantage de précisions sur la façon dont ce balisage a été réalisé dans [Adda et alii 1997][1].

<head type=PRINCIPAL id=afp_mai1994_142.1 n=1>
Un mort et trois blessés dans un attentat de l'IRA à la voiture piégée
</head>

Un homme a été tué et trois personnes blessées, dont deux enfants, dans l'explosion vendredi matin d'une voiture piégée à Lurgan, à 30 kilomètres au sud-ouest de Belfast, a-t-on appris auprès de la police.

L'attentat a été revendiqué dans la matinée par l'Armée républicaine irlandaise (IRA).

La bombe a explosé à 08H20 locales (07H20 GMT) au moment où l'homme, sa femme et ses enfants montaient dans le véhicule.

L'homme était un civil employé dans la station de police de la ville où il s'occupait du nettoyage.

L'un des enfants, une fille de trois ans, souffre de plusieurs fractures et de blessures au visage. Sa mère et son frère de neuf ans ont été également hospitalisés pour des blessures plus légères.

Jeudi soir, un catholique de 23 ans avait été tué par balles lors d'un attentat loyaliste dans le nord de Belfast, alors qu'il tenait dans ses bras un bébé d'un an.

Dimanche, une septuagénaire catholique avait été tuée chez elle, alors qu'elle regardait la télévision, par un commando de l'UVF (Ulster Volunteer Force), une milice paramilitaire protestante à Dungannon (ouest).

[1] Nous remercions Gilles Adda d'avoir mis à notre disposition les corpus du journal *Le Monde* et de l'AFP sur lesquels nous avons travaillé.

2. Chaîne de pré-traitement

La chaîne de pré-traitement des textes précédant la segmentation thématique ou le calcul des cooccurrences lexicales est formée comme nous l'avons décrit au chapitre 9 de trois opérations principales : la segmentation des textes opérée par le segmenteur *Mtseg*, leur étiquetage morpho-syntaxique, réalisé par le *TreeTagger*, et la sélection ou le repérage des mots dits pleins. Le résultat obtenu à l'issue de chacune de ces trois opérations est illustré sur le texte exemple ci-dessous.

Texte exemple

Ces 5 pays, qui avaient été élus le 8 novembre dernier par l'Assemblée générale des Nations unies, siégeront au Conseil comme membres non-permanents pour une durée de deux ans (jusqu'au 31 décembre 1997). Ils remplacent l'Argentine, Oman, le Nigeria, le Rwanda et la République tchèque.

Dans le cas des textes tels que celui donné en exemple au §1, la segmentation est précédée par un élagage des balises : on ne retient qu'une version minimale (c'est-à-dire avec suppression de toutes les informations présentes dans les balises autre que la balise elle-même) des balises de début et de fin de texte ainsi que des balises de début et de fin de paragraphe. Par ailleurs, on supprime les éléments, tels que les titres ou les auteurs, ne faisant pas partie du corps des textes.

Texte exemple à la suite de la segmentation par *Mtseg*

```
        [CHUNK      <DIV FROM="1">
        (PAR        <P FROM="1">
        (SENT       <S>
1\1     TOK         Ces
1\5     DIG         5
1\7     TOK         pays
1\11    PUNCT       ,
1\13    TOK         qui
1\17    TOK         avaient
1\25    TOK         été
1\29    TOK         élus
1\34    TOK         le
1\37    DATE        8_novembre
1\48    TOK         dernier
1\56    TOK         par
1\60    LSPLIT      l'
1\62    COMP        Assemblée_générale
1\81    TOK         des
1\85    COMP        Nations_unies
1\98    PUNCT       ,
1\100   TOK         siégeront
1\110   TOK         au
1\113   TOK         Conseil
1\121   TOK         comme
1\127   TOK         membres
1\135   TOK         non-permanents
1\150   TOK         pour
1\155   TOK         une
1\159   TOK         durée
1\165   TOK         de
1\168   TOK         deux
1\173   TOK         ans
1\177   OPUNCT      (
1\178   COMP        jusqu'au
1\187   DATE        31_décembre_1997
1\203   CPUNCT      )
1\204   PTERM_P     .
        )SENT       </S>
        (SENT       <S>
1\206   TOK         Ils
1\210   TOK         remplacent
1\221   LSPLIT      l'
1\223   TOK         Argentine
1\232   PUNCT       ,
1\234   TOK         Oman
1\238   PUNCT       ,
1\240   TOK         le
1\243   TOK         Nigeria
1\250   PUNCT       ,
1\252   TOK         le
```

```
1\255   TOK     Rwanda                  1\286   PTERM_P     .
1\262   TOK     et                              )SENT   </S>
1\265   TOK     la                              )PAR    </P>
1\268   TOK     République                      )PAR    </P>
1\279   TOK     tchèque                         ]CHUNK          </DIV>
```

Comme on peut le constater au travers du balisage ci-dessus, *Mtseg* fait apparaître plusieurs niveaux de segmentation : segmentation en paragraphes, en phrases et en mots. Dans le cas présent, nous nous intéressons essentiellement à la segmentation en mots. Celle-ci permet plus spécifiquement de faire la distinction entre les mots, simples ou composés (TOK ou COMP), les signes de ponctuation (xPUNCT) et des entités spécifiques comme les chiffres (DIG) ou les dates (DATE). Ces distinctions sont particulièrement utiles lors de la sélection finale des mots dits pleins.

La segmentation est suivie d'une étape de transformation en vue de l'étiquetage par le *TreeTagger*. Cette transformation consiste d'abord à supprimer l'intégralité du balisage introduit par la segmentation tout en conservant la présentation, requise par le *TreeTagger*, d'une entité (mot ou signe de ponctuation) par ligne. Elle se poursuit par l'éclatement des noms composés[1] mis en évidence par la segmentation. Cet éclatement est rendu nécessaire par leur absence au niveau du vocabulaire du *TreeTagger*. On est ainsi obligé d'étiqueter séparément leurs différents constituants et de les reconstituer après cet étiquetage, avec bien entendu une certaine variabilité sur la forme finale d'un mot composé induit par les différences dans la façon dont sont étiquetés ses constituants en fonction du contexte plus large dans lequel ils sont plongés. Il y a donc une certaine perte dans la reconnaissance des mots composés.

Aussi bien dans le cas du balisage que dans celui des mots composés, on conserve les informations supprimées afin de les faire réapparaître après l'étiquetage. Dans le cas des balises, on se contente de retenir l'étiquette associée à chaque entité en laissant de côté toutes les informations de position ainsi que les marques de délimitation des unités au delà du mot ajoutées par le segmenteur (on conserve les délimitations de paragraphe pré-existantes et les limites des phrases restent accessibles au travers de l'étiquette assignée par le segmenteur aux signes de ponctuation).

[1] Il s'agit plus précisément des mots composés dans lesquels aucun tiret ne marque l'attachement des mots entre eux.

Texte exemple après étiquetage morpho-syntaxique par le *TreeTagger*

```
<tt>
Ces PRO:demo:attr ce
5 ADJ:num:card @card@
pays NOM pays
, PON:comma ,
qui PRO:rela qui
avaient VER:aux:impf avoir
été VER:aux:pper être
élus VER:pper élire
le DET:def le
8 ADJ:num:card @card@
novembre NOM novembre
dernier ADJ dernier
par PRE par
l' DET:def le
Assemblée NOM assemblée
générale ADJ général
des PRE:det de+le
Nations NOM nation
unies ADJ uni
, PON:comma ,
siégeront VER:futu siéger
au PRE:det à+le
Conseil NOM conseil
comme ADV comme
membres NOM membre
non-permanents ADJ permanent
pour PRE pour
une DET:indef une
durée NOM durée
de PRE de
deux ADJ:num:card deux
ans NOM an
( PON (
jusqu'au NPR jusqu'au
31 ADJ:num:card @card@
décembre NOM décembre
1997 ADJ:num:card @card@
) PON )
. PON:sep .
Ils PRO:pers:conj il
remplacent VER:pres remplacer
l' DET:def le
Argentine ADJ argentin
, PON:comma ,
Oman NPR Oman
, PON:comma ,
le DET:def le
Nigeria NPR Nigeria
, PON:comma ,
le DET:def le
Rwanda NPR Rwanda
et CON:coo et
la DET:def le
République NOM république
tchèque ADJ tchèque
. PON:sep .
</tt>
```

Comme on peut le constater au niveau du résultat ci-dessus de l'étiquetage du texte exemple, le *TreeTagger* assure à la fois la levée des ambiguïtés morpho-syntaxiques et la lemmatisation. Il permet également de reconnaître dans une certaine mesure les noms propres[1] (étiquette NPR).

Cet étiquetage est suivi de deux opérations inversant l'effet de celles l'ayant précédé. On commence ainsi par reconstituer les mots composés et ajouter ensuite à chaque entité l'étiquette avec laquelle elle avait été marquée lors de la segmentation. On réalise donc une fusion entre les résultats de la segmentation et ceux de l'étiquetage morpho-syntaxique. Le produit de cette fusion est illustrée ci-dessous pour le texte exemple.

[1] Cette reconnaissance est sans doute l'un des points faibles de l'étiqueteur dans la mesure où cette reconnaissance est tributaire de son vocabulaire. Il n'offre pas la possibilité de compléter une liste de noms propres définie de façon déclarative et ne possède pas de mécanisme permettant de les détecter. Les manques sont donc importants en ce qui concerne ce point précis.

Texte exemple après fusion des résultats de la segmentation et de l'étiquetage

TOK <tt>
TOK Ces PRO:demo:attr ce
DIG 5 ADJ:num:card @card@
TOK pays NOM pays
PUNCT , PON:comma ,
TOK qui PRO:rela qui
TOK avaient VER:aux:impf avoir
TOK été VER:aux:pper être
TOK élus VER:pper élire
TOK le DET:def le
DATE 8_novembre ADJ:num:card NOM @card@ novembre
TOK dernier ADJ dernier
TOK par PRE par
LSPLIT l' DET:def le
COMP Assemblée_générale NOM ADJ assemblée général
TOK des PRE:det de+le
COMP Nations_unies NOM ADJ nation uni
PUNCT , PON:comma ,
TOK siégeront VER:futu siéger
TOK au PRE:det à+le
TOK Conseil NOM conseil
TOK comme ADV comme
TOK membres NOM membre
TOK non-permanents ADJ permanent
TOK pour PRE pour
TOK une DET:indef une
TOK durée NOM durée
TOK de PRE de
TOK deux ADJ:num:card deux
TOK ans NOM an
OPUNCT (PON (
COMP jusqu'au NPR jusqu'au
DATE 31_décembre_1997 ADJ:num:card NOM ADJ:num:card @card@ décembre @card@
CPUNCT) PON)
PTERM_P . PON:sep .
TOK Ils PRO:pers:conj il
TOK remplacent VER:pres remplacer
LSPLIT l' DET:def le
TOK Argentine ADJ argentin
PUNCT , PON:comma ,
TOK Oman NPR Oman
PUNCT , PON:comma ,
TOK le DET:def le
TOK Nigeria NPR Nigeria
PUNCT , PON:comma ,
TOK le DET:def le
TOK Rwanda NPR Rwanda
TOK et CON:coo et
TOK la DET:def le
TOK République NOM république
TOK tchèque ADJ tchèque
PTERM_P . PON:sep .
TOK </tt>

Texte exemple à l'issue de la sélection des mots dits pleins (format long)

<tt>
<MV ce>
<MV 5>
<NOM pays>
<MV qui>
<MV avoir>
<MV être>
<VER élire>
<MV le>
<MV 8_novembre>
<ADJ dernier>
<MV par>
<MV le>
<COMP assemblée_général>
<MV de+le>
<COMP nation_uni>
<VER siéger>
<MV à+le>
<NOM conseil>
<MV comme>
<NOM membre>
<ADJ permanent>
<MV pour>
<MV une>
<NOM durée>
<MV de>
<MV deux>
<NOM an>
<MV jusqu'au>
<MV 31_décembre_1997>
<MV PTERM>
<MV il>
<VER remplacer>
<MV le>
<ADJ argentin>
<NPR oman>
<MV le>
<NPR nigeria>
<MV le>
<NPR rwanda>
<MV et>
<MV le>
<NOM république>
<ADJ tchèque>
<MV PTERM>
</tt>

La dernière étape consiste à faire apparaître la distinction entre les mots qui seront considérés lors des traitements ultérieurs et ceux qui seront laissés de côté. Selon la nature de ces traitements, deux formats ont été définis. Le format court ne fait apparaître que la liste des mots retenus, sans aucun marquage de la nature de ces mots. Il est utilisé dans le cadre du calcul des cooccurrences lexicales. Le nombre important de textes à traiter dans le cadre de cette tâche oblige en effet à privilégier la concision au détriment de l'informativité.

Le format long est utilisé pour sa part dans le cadre de la segmentation thématique. Il conserve tous les mots du texte, ainsi que les délimitations de paragraphe et de phrases. La mise en évidence des mots à retenir se fait par un marquage de type SGML : la balise MV sert à désigner les mots devant être laissés de côté ; les mots à retenir sont caractérisés quant à eux par une balise reprenant leur étiquette morpho-syntaxique. On conserve de cette façon une information que l'on peut ensuite exploiter pour mettre en œuvre par exemple des filtres sans avoir à relancer la chaîne de pré-traitement.

Texte exemple à la suite de la sélection des mots dits pleins (format court)

<tt>
pays
élire
dernier
assemblée_général
nation_uni
siéger
conseil
membre
permanent
durée
an
remplacer
argentin
oman
nigeria
rwanda
république
tchèque
</tt>

Stop-list

abréviations		verbes auxiliaires et modaux	composés non significatifs	erreurs de la chaîne à filtrer
etc	bd	être	d'autre_part	autre
mm	boul	faire	jusqu'à_présent	même
me	doc	pouvoir	suite_à	fois
mgr	ex	devoir	d'une_part	d'antan
mme	messrs	falloir	avoir_priori	l'on
mr	mlle	vouloir	grâce_à	l'un
dr	mrs	permettre	avoir_posteriori	jusqu'à
st	ms	avoir	tel_quel	l'une
tel	no		avoir_fortiori	jusqu'au
tél	pr		avoir_contrario	jusqu'aux
avcf	prof		vingt_et_un	d'emblée
kg	ste			d'affilée
km	vol			
ap	c.-à-d			
apr	ets			
ave	civ			
	c-à-d			

La phase finale de sélection ou de marquage des mots à retenir passe par l'application d'une stop-list, c'est-à-dire par la suppression d'un ensemble de mots considérés comme non significatifs à des titres divers vis-à-vis de la tâche considérée. Dans le cas présent, cette liste rassemble : un ensemble d'abréviations, les verbes auxiliaires et les verbes modaux les plus courants, des mots composés n'étant pas des noms composés et des erreurs de la chaîne de pré-traitement, c'est-à-dire des mots ne devant pas être retenus mais n'apparaissant pas comme tels.

Annexe H

Outils de SEGCOHLEX

1. Calcul des collocations

Le calcul des collocations est assuré par un outil spécifique permettant de gérer de manière automatique toute une chaîne de traitements. Compte tenu du nombre de collocations à enregistrer, il est impossible de faire tenir en mémoire vive toutes les collocations enregistrées sur 24 mois du journal *Le Monde*. Par ailleurs, il est également impossible d'assurer cet enregistrement uniquement dans un fichier car les temps d'accès

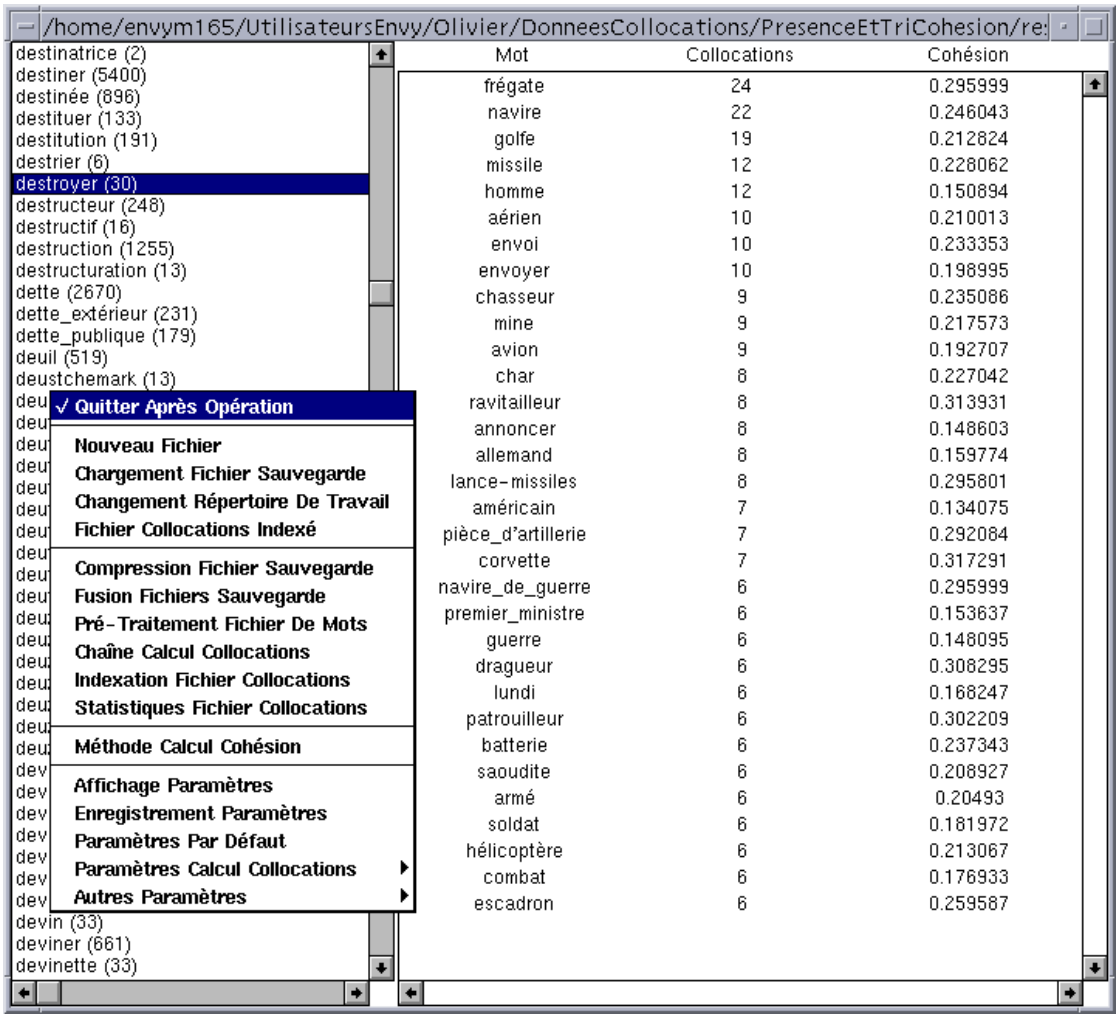


Fig. H.1 - Interface de visualisation des collocations et de contrôle des outils assurant leur recueil

à un disque dur sont trop élevés pour le nombre de collocations à traiter. Le recueil de ces dernières nécessite donc la mise en œuvre de toute une chaîne de traitements. L'interface de la figure H.1 constitue à la fois le centre de contrôle de ces traitements et l'outil de visualisation des collocations recueillies.

Les collocations présentant la propriété intéressante d'être additives, nous avons opté pour une stratégie consistant à découper les fichiers en sous-fichiers d'une taille suffisamment réduite pour les traiter en ayant recours principalement à la mémoire vive comme instrument de stockage (taille en moyenne égale à une moitié d'un mois du journal *Le Monde*). Plus précisément, nous utilisons des fichiers de débordement pour stocker les collocations les moins fréquentes, ce qui permet de travailler à partir de plus gros fichiers tout en ne pénalisant pas trop les temps de traitement.

La chaîne de traitement gère donc de manière automatique le découpage des fichiers, le recueil des collocations pour chacun d'entre eux, leur enregistrement sur fichier puis la fusion de ces différents résultats pour obtenir le réseau final de collocations. À côté de cette chaîne, on trouve également les outils permettant de filtrer les collocations de faible fréquence ainsi que d'effectuer le compactage du réseau que ce filtrage implique (suppression des mots auxquels il n'est plus fait référence). À ceux-ci s'ajoutent également les outils permettant en final d'indexer le réseau de collocations afin d'y assurer un accès suffisamment performant.

La chaîne de traitement décrite ci-dessus ainsi que les outils venant la compléter peuvent être lancés à partir de l'interface de la figure H.1. Il est même possible de lancer les différents outils composant la chaîne de traitement indépendamment les uns des autres si on le souhaite. Cette interface est également utilisable pour fixer les paramètres du calcul des collocations (taille de la fenêtre, mesure de cohésion, etc.).

La visualisation des collocations ne présente quant à elle pas de particularités notables en dehors de ce que montre la figure H.1 et d'une fonction de tri permettant de afficher les collocations soit par ordre alphabétique, soit par ordre décroissant de leur nombre d'occurrences ou de leur mesure de cohésion.

2. Segmentation thématique

La méthode de segmentation thématique de SEGCOHLEX est une méthode dite quantitative. Comme toutes les méthodes de ce type, elle comporte un ensemble de paramètres dont le réglage donne lieu à une phase de mise au point assez longue et délicate. L'interface de contrôle et de visualisation de la figure H.2 a été construite afin de faciliter ce travail de mise au point. Elle permet tout d'abord d'avoir un accès direct à

l'ensemble des paramètres de la méthode, soit directement, soit par l'intermédiaire d'un menu : taille de la fenêtre de calcul de la cohésion, taille de la fenêtre de lissage, type des mots retenus (nom, adjectif et/ou verbe), nombre de mots de la fenêtre nécessaire pour sélectionner un mot du réseau de collocations, etc.

Cette interface permet ensuite de lancer les différentes opérations de la méthode de segmentation et de visualiser directement leur résultat de façon graphique sous la forme d'une courbe. La présence des mots du texte avec leur position assure par ailleurs la mise en correspondance directe de la courbe avec le texte. Le lancement des opérations peut avoir lieu pas par pas ou en séquence. Enfin, on peut à tout moment rappeler les valeurs de cohésion obtenues à l'issue d'une des opérations effectuées et les faire apparaître sous forme de courbe. Cette visualisation peut être éventuellement dupliquée afin de pouvoir comparer plusieurs courbes différentes.

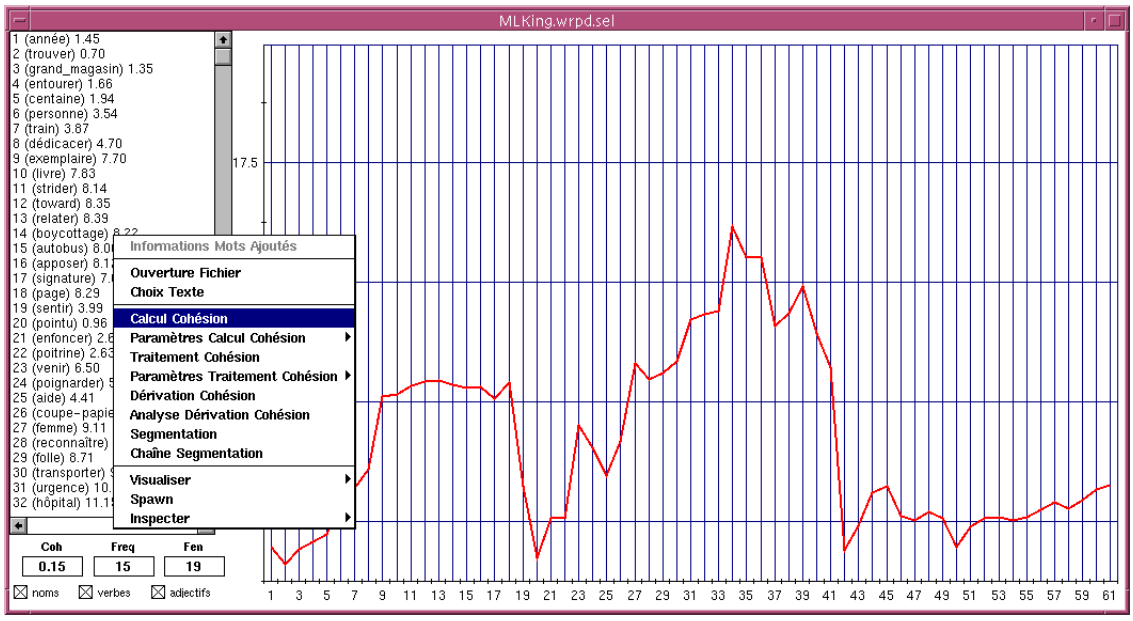


Fig. H.2 - Interface de contrôle de la segmentation thématique de SEGCOHLEX

Il est à noter que nous utilisons le même outil, légèrement adapté, pour la mise au point de la segmentation thématique de SEGAPSITH.

Annexe I

Les méthodes quantitatives de segmentation thématique

Dans cette annexe, nous présentons un panorama des principales méthodes quantitatives de segmentation thématique venant compléter les travaux présentés au cours du chapitre 9.

1. La segmentation thématique sans utilisation de connaissances

1.1. Vocabulary Management Profile

La méthode décrite par Youmans dans [Youmans 1991] n'est pas à proprement parler une méthode de segmentation thématique mais plutôt une méthode d'analyse de l'information lexicale présente dans les textes. Son principe consiste tout simplement à évaluer l'information apportée par les différentes parties d'un texte en comptant le nombre de mots nouveaux que celles-ci introduisent. Plus précisément, on utilise une fenêtre rassemblant un nombre fixe de mots que l'on déplace sur le texte. À chaque station de cette fenêtre, on compte le nombre de mots à l'intérieur de celle-ci correspondant à des mots nouvellement introduits dans le texte. On obtient en final une courbe traduisant l'évolution des apports en mots nouveaux en fonction de la position dans le texte. Cette courbe est appelée Vocabulary Management Profile (VMP). Dans son expérimentation, Youmans a choisi une taille de fenêtre de 35 mots et un pas de déplacement de 1 mot. Un mot nouvellement introduit sera donc comptabilisé dans 35 stations de la fenêtre consécutives.

L'intérêt du travail de Youmans du point de vue de la segmentation thématique réside dans la corrélation[1] que celui-ci a observé entre les évolutions présentes au niveau du VMP d'un texte et les évolutions du contenu informationnel de celui-ci. En particulier, certains motifs pourraient être rapprochés des bornes de segments textuels délimitant des zones thématiquement homogènes. L'hypothèse sous-jacente est que l'apparition d'un nouveau thème s'accompagne de l'introduction d'un nouveau vocabulaire qui lui est lié.

[1] Il s'agit d'une corrélation et non d'une relation directe.

Aucune procédure n'est cependant proposée pour mettre en évidence ces bornes de façon automatique à partir de la courbe initiale. Par ailleurs, les valeurs des paramètres adoptées par Youmans n'ont pas été spécifiquement fixées pour la tâche de segmentation. Nomoto et Nitta rapportent dans [Nomoto & Nitta 1994] que l'utilisation de cette méthode pour retrouver les frontières entre des articles ayant été concaténés a été un échec avec une fenêtre assez large, d'une taille de 300 mots, le VMP devenant alors trop plat. Hearst a quant à elle adapté avec plus de succès cette méthode pour segmenter des textes suivant leurs différents thèmes en utilisant une taille de fenêtre voisine de celle de Youmans (40 mots) mais en retenant un pas de déplacement plus important (20 mots).

L'avantage de cette méthode réside son extrême simplicité. Comme toutes celles fondées sur la distribution des mots, elle est tributaire de la présence d'un vocabulaire suffisamment spécifique vis-à-vis des thèmes traités. Au delà de ce facteur général, sa précision concernant la délimitation des segments est plus particulièrement dépendante de la présence de ce vocabulaire spécifique au début de chaque segment. Si les mots thématiquement significatifs sont trop répartis sur l'ensemble d'un segment, on n'observera pas de pic au niveau de la courbe d'introduction du vocabulaire au début de ce segment et sa détection ne sera pas possible, ou sera tout du moins très imprécise. Par ailleurs, la méthode VMP ne permet intrinsèquement pas de repérer plusieurs segments faisant référence à un même thème. En effet, lorsque le vocabulaire propre à un thème a été introduit par un segment, il n'est plus inédit et ne peut donc plus servir à détecter un éventuel autre segment situé dans la suite du texte et traitant du même thème.

1.2. TextTiling

(cf. §1.3 du chapitre 9)

1.3. Nomoto et Nitta

Le travail exposé dans [Nomoto & Nitta 1994] est très proche de la méthode de Hearst fondée sur la mesure du cosinus. Une des principales différences entre les deux travaux réside dans la façon dont les unités élémentaires constituant les blocs, les pseudo-phrases dans le cas de Hearst, sont déterminées. Nomoto et Nitta exploitent sur ce point l'existence en japonais d'une structure du type thème/rhème explicitement marquée dans laquelle l'exposé du sujet est séparé du commentaire qui lui est associé par une marque linguistique particulière, appelée le *wa*. Précisons que cette structure se situe à une échelle plus large que l'habituelle distinction entre thème et rhème puisqu'elle s'étend sur plusieurs phrases. Nomoto et Nitta définissent ainsi la notion de segment de discours

comme étant un bloc de phrases délimité par une marque de fin de texte et/ou la présence d'un *wa*. Au lieu de comparer des blocs de pseudo-phrases, il comparent donc des blocs de segments de discours.

Le processus de segmentation est globalement le même que celui de Hearst :

- normalisation des textes. Celle-ci comporte deux étapes. La première consiste à ne sélectionner que les noms présents dans les textes, ce qui est réalisé en utilisant un analyseur morphologique. La seconde effectue un premier découpage des textes en segments de discours, au sens défini ci-dessus. Suivant les genres textuels, on obtient des segments de discours allant de quelques noms jusqu'à une soixantaine de noms;
- calcul d'une courbe de cohérence par comparaison des blocs adjacents de segments de discours;
- analyse de la courbe de cohérence pour détecter les bornes des segments de texte thématiquement homogènes. Cette analyse est réalisée ici manuellement. Comme dans le cas de Hearst, on recherche pour ce faire les minima les plus significatifs de la courbe.

La deuxième étape est menée de la même façon que dans TextTiling. On calcule une valeur de cohérence à chaque frontière entre segments de discours en comparant les deux blocs formés par *k* segments de discours situés de part et d'autre de la frontière considérée. *k* est aussi égal à 10 dans le cas présent. Le processus est équivalent à la comparaison des deux volets d'une fenêtre d'une taille de k*2 segments de discours que l'on déplace sur les textes avec un pas de un segment de discours.

Les différences avec TextTiling résident dans la mesure utilisée pour comparer les blocs de segments de discours ainsi que dans la politique de pondération des mots formant ces blocs. Concernant cette dernière, la simple utilisation du nombre d'occurrences des mots est remplacée par une pondération suivant le facteur *tf.idf* [Salton 1989]. Celui-ci, habituellement utilisé en Recherche d'Informations pour rendre compte de l'importance d'un terme dans un document, est défini de la façon suivante :

$$w_{ij} = tf_{ij} \log \frac{N}{df_j}$$

où tf_{ij} est le nombre d'occurrences du terme T_j dans le document D_i, N est le nombre total de documents considérés et df_j est le nombre de documents contenant le terme T_j. On remarque que ce facteur défavorise fortement les termes que l'on retrouve dans tous les documents, considérés comme peu discriminants, alors qu'il favorise au contraire les

termes caractéristiques d'un petit sous-ensemble de documents. Dans le contexte du travail de Nomoto et Nitta, ce facteur est calculé en considérant qu'un document s'identifie à un bloc de segments de discours et que N correspond donc à l'ensemble des blocs formés par le déplacement de la fenêtre de calcul de la courbe de cohérence.

La mesure de similarité utilisée pour comparer ces blocs deux à deux et calculer ainsi la cohérence thématique est une mesure proche du cosinus utilisé dans TextTiling. Il s'agit en l'occurrence du cœfficient de Dice, qui s'écrit de la façon suivante pour deux vecteurs X ($x_1,\ldots,x_t$) et Y ($y_1,\ldots,y_t$), représentant chacun un bloc de segments de discours :

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

où $w(x_i)$ représente le poids affecté au mot x_i en utilisant la pondération *tf.idf*.

Le calcul de la courbe de cohérence est suivie d'une opération de lissage permettant de faire apparaître plus nettement ses grandes évolutions. Ce lissage est réalisé par une fenêtre de moyennage local d'une taille de 5 segments de discours.

Une évaluation de la méthode a été réalisée sur une tâche de redécouverte des frontières d'un ensemble de textes concaténés, comme dans le cas de la seconde évaluation présentée par Hearst. L'ensemble à segmenter était formé ici d'une vingtaine d'éditoriaux d'un journal économique. Différentes tailles régulières de segment de discours ont été testées, entre 5 et 35 mots, en plus du découpage réalisé en fonction des critères linguistiques exposés ci-dessus, lequel donne des segments de taille irrégulière. Les meilleurs résultats en termes de moyenne de la précision et du rappel ont été obtenus pour le découpage en segments linguistiquement marqués : autour de 0,81 pour le rappel et de 0,5 pour la précision. On constate en outre que les résultats les plus proches pour des segments de taille régulière sont obtenus pour des segments de 10 et de 15 mots, ce qui est proche de la moyenne de 13,7 mots des segments linguistiquement fondés.

1.4. Reynar

Alors que les méthodes proposées par Hearst et Nomoto et Nitta reposent sur des comparaisons locales de blocs adjacents, celle développée par Reynar [Reynar 1994] aborde les textes et le problème de leur segmentation de façon plus globale. Son principe général consiste en effet à placer des bornes une par une, soit à des frontières de phrases, soit à des frontières de paragraphes, de façon à minimiser un critère global relatif à la similarité des différents segments ainsi construits. Le processus s'arrête soit lorsque le

placement d'une nouvelle borne conduit à une augmentation du critère, quelle que soit sa position, soit lorsqu'un nombre de bornes fixé a priori est atteint.

Comme dans les cas précédents, les textes subissent au préalable un pré-traitement. Ils sont lemmatisés et leurs mots grammaticaux ainsi qu'un ensemble de mots très fréquents et peu significatifs (comme les verbes être et avoir par exemple) sont filtrés. Au stade initial de l'algorithme, le texte ne comporte aucune borne en dehors de celle de début de texte et de celle de fin de texte. Lorsqu'une nouvelle borne est ajoutée, on considère qu'elle crée une partition de l'espace s'étendant de la borne précédente (vers le début du texte) à la fin du texte. Sur le principe, la borne est placée de telle façon que la similarité entre les deux segments ainsi créés soit minimale. Pour ce faire, on cherche à minimiser le produit scalaire des vecteurs représentant ces deux segments : $V_{(j-1,j)} \cdot V_{(j,n)}$, où j-1 représente la borne précédente, j, la borne ajoutée et n, la borne de fin de texte. V(i,j) représente le vecteur formé par le nombre d'occurrences de chaque lemme apparaissant entre les bornes i et j. Le critère global que l'on cherche à minimiser lorsque l'on ajoute une nouvelle borne correspond à la somme des produits scalaires associés aux différentes partitions créées :

$$\sum_{j=2}^{b} \frac{V_{(j-1,\,j)} \;\; V_{(j,n)}}{NM(j-1,j) \;\; NM(j,n)}$$

où b est le nombre total de bornes posées et NM(j-1,j) est le nombre de mots composant le segment allant de la borne j-1 à j.

Une évaluation de la méthode a été réalisée comme précédemment sur une tâche de mise en évidence des frontières d'un ensemble de textes concaténés. L'expérimentation a été réalisée dans ce cas sur 660 articles du *Wall Street Journal*, chaque article ayant une taille moyenne de 600 mots. La tolérance dans la mise en correspondance entre borne placée et frontière de texte était comme pour Hearst de 3 phrases. Les bornes ont de plus été placées de telle façon que les segments obtenus aient une taille minimale. Une précision et un rappel moyens ont été calculés sur 150 concaténations différentes des 660 articles. Les chiffres donnés laissent apparaître une disparité assez importante entre la précision et le rappel et une sensibilité assez grande vis-à-vis de la politique de placement des bornes : lorsque l'on fait correspondre les bornes avec des frontières de phrase, on obtient une précision de 0,304 et un rappel de 0,803 tandis que l'utilisation des frontières de paragraphe donne une précision de 0,916 mais un rappel de 0,3.

1.5. Segmentation et Recherche d'Informations

Un certain nombre de travaux en Recherche d'Informations se sont également intéressés à la segmentation des textes, avec deux préoccupations principales. L'une d'elles répond au souci de ne pas noyer l'utilisateur sous une masse trop importante de documents en réponse à une requête. Lorsque les documents sont de taille importante (plusieurs pages), le nombre de ceux qu'un utilisateur peut examiner est en effet réduit. Au contraire, si l'on extrait de ces documents le ou les passages les plus directement en relation avec sa requête, il lui est alors possible de passer en revue un ensemble plus important de réponses.

Le second intérêt de la segmentation des textes réside dans une plus grande efficacité de la recherche en elle-même. Au lieu de calculer une similarité entre une requête et la totalité d'un texte[1], on peut mesurer la similarité entre cette requête et chacun des segments constituant le texte. Chaque segment étant thématiquement plus homogène que l'ensemble du texte, on pourra ainsi trouver à un échelon local une similarité n'existant pas au niveau global et améliorer ainsi le taux de rappel.

Dans ce contexte, la segmentation est utilisée pour cerner les zones de texte abordant un sujet particulier mais sans avoir à tenir nécessairement compte de la linéarité des textes. Un segment thématique peut donc être constitué de plusieurs morceaux de texte non contigus. Salton a particulièrement exploré cette problématique, aussi bien dans la perspective de la Recherche d'Informations que dans celle du résumé automatique de texte [Salton et alii 1996]. Le principe sur lequel reposent ses travaux dans ce domaine consiste à choisir une unité de base, définie sur des critères de forme (organisation typographique), et à établir pour chaque document traité une carte des relations de proximité existant entre toutes les unités qu'il rassemble. La segmentation est ensuite établie suivant la configuration de cette carte.

Dans [Salton et alii 1996], l'unité considérée est le paragraphe. Chacun d'entre eux est classiquement représenté par un vecteur dont chaque dimension correspond à l'un de ses mots. Il s'agit là des mots-types et non de leurs occurrences. On parlera dans ce qui suit de terme. De plus, à l'instar des travaux précédemment exposés, un certain nombre de pré-traitements peuvent être réalisés sur les textes, allant jusqu'au stemming[2]. Toutes les occurrences de mots appartenant à une même classe d'équivalence dérivationnelle sont

[1] La requête et le texte sont représentés par des vecteurs dont les coordonnées représentent les poids des termes qu'ils contiennent. Le calcul de similarité prend de ce fait la forme d'une opération du type produit scalaire.

[2]Le stemming est l'opération permettant de retrouver l'équivalence entre deux mots du point de vue de la morphologie dérivationnelle. Il permet par exemple de ramener à une même forme des mots comme fabriquer, fabrication, fabricant, fabrique et fabricateur.

alors représentées par un même terme. La valeur associée à une dimension d'un tel vecteur est égale au nombre d'occurrences du terme considéré, pondéré par le facteur *tf.idf*. Dans le cas présent, celui-ci est utilisé dans son cadre de définition originel : on caractérise l'importance du mot dans un document en fonction de sa répartition dans l'ensemble de tous les documents disponibles.

En application du principe exposé ci-dessus, une carte des relations entre paragraphes est construite pour chaque texte traité. Cette carte est établie en calculant la similarité entre tous les paragraphes du texte. La similarité entre deux paragraphes est définie par le produit scalaire des vecteurs qui leur sont associés. Celui-ci est normalisé entre 0 et 1. On ne retient comme relation de la carte que les liens entre deux paragraphes possédant une similarité au moins égale à 0,2.

Cette carte est ensuite exploitée afin de faire émerger deux types d'entités : les segments et les thèmes. Les segments sont des unités textuelles formées de plusieurs paragraphes contigus fortement liés entre eux mais globalement peu connectés aux autres paragraphes du texte. Leur cohérence se définit plus particulièrement au niveau rhétorique. Un segment comporte souvent un passage introductif suivi d'un développement et enfin d'une forme de conclusion. Les segments sont construits en simplifiant la carte des relations entre paragraphes de façon à ne faire apparaître que les relations entre des paragraphes proches dans le texte, du point de vue de son organisation linéaire. On masque pour ce faire toutes les relations enjambant plus de 5 paragraphes. On définit alors un segment comme un ensemble de paragraphes liés entre eux (pas nécessairement de façon complète mais le graphe formé doit être connexe) n'entretenant aucune relation avec des paragraphes situés en dehors de ce segment.

Lorsque les textes sont de taille importante, il est possible d'appliquer le principe de définition des segments de façon récursive. En fusionnant les vecteurs représentant les différents paragraphes d'un segment, on obtient un ensemble d'unités qui sont comparables entre elles de la même façon que les paragraphes. On peut donc dresser une carte des relations entre segments et faire éventuellement émerger des macro-segments.

Contrairement aux segments, les thèmes s'affranchissent totalement de l'organisation linéaire des textes. Leur objet est simplement de regrouper les parties d'un texte relatives au même sujet. Leur définition s'effectue suivant une logique proche de celle d'un processus de classification. À partir de la carte initiale des relations entre paragraphes, on considère tous les sous-ensembles de trois paragraphes liés les uns aux autres (relations formant un triangle) et l'on calcule par chacun d'entre eux un vecteur représentant en quelque sorte le centre de gravité des trois paragraphes regroupés. Ces vecteurs sont ensuite fusionnés les uns aux autres lorsque leur similarité dépasse un seuil donné. Encore une fois, la procédure est récursive dans la mesure où le résultat de la fusion de

deux vecteurs peut lui-même fusionné ultérieurement avec un autre vecteur. Le processus s'arrête lorsque la similarité entre vecteurs est inférieure au seuil de référence pour toutes les combinaisons possibles.

L'intérêt du travail de Salton par rapport aux approches présentées précédemment est de dissocier nettement deux dimensions de la segmentation du discours : la définition de segments ayant une cohérence sur le plan rhétorique et celle de segments ayant une cohérence sur le plan thématique. Cette dissociation permet en particulier d'étudier les interactions entre ces deux plans. Les thèmes et les segments sont en effet représentés sous une même forme, en l'occurrence des vecteurs de mots pondérés, et pour juger de la similarité d'un thème et d'un segment, il suffit de calculer une mesure de similarité entre les vecteurs qui les représentent.

Salton montre ainsi que pour certains textes, on observe un parallélisme entre le plan rhétorique et le plan thématique mais que les rapports entre ces deux plans sont pour bon nombre de textes beaucoup plus complexes : un segment peut faire référence à plusieurs thèmes et réciproquement, un thème apparaît souvent dans plusieurs segments. Salton caractérise ainsi quelques grands modes d'organisation textuelle et illustre leur exploitation pour déterminer ce que l'on retient d'un texte dans des tâches comme la Recherche d'Informations ou le résumé automatique de textes.

L'inconvénient essentiel de l'approche présentée réside dans la nécessité de définir a priori des unités textuelles suffisamment grandes. Les phrases sont à cet égard des unités trop petites. La segmentation réalisée n'est donc potentiellement pas aussi précise que celle produite par un système tel que TextTiling. Pour la Recherche d'Informations, adopter le paragraphe, ainsi que le préconise Salton, est tout à fait satisfaisant. C'est en revanche un niveau trop élevé pour d'autres tâches, comme l'Extraction d'Informations, dans lesquelles la détermination du sujet traité doit être réalisé à un niveau plus fin. L'utilisation des paragraphes comme unités de base se heurte par ailleurs à leur manque d'homogénéité tant sur la forme que sur le fond [Stark 1988]. Ils sont en effet de taille assez variable et leur définition obéit à des critères multiples : changement de thème, aération de la présentation, mouvement rhétorique ... Parmi ces variations, il faut citer également le fait que le changement porté par un nouveau paragraphe peut être introduit aussi bien au début de ce paragraphe qu'à la fin du paragraphe précédent.

Sur le plan méthodologique, il faut préciser que processus de segmentation proposé n'a pas fait l'objet d'une évaluation formelle, que ce soit par rapport à des jugements d'experts ou par rapport à des unités textuelles naturelles (textes, sections ...). La validation exposée dans [Salton et alii 1996], qui porte sur des textes extraits d'encyclopédies, est principalement de nature qualitative. La seule évaluation véritable est

indirecte : elle porte sur l'amélioration apportée par ce type de segmentation sur les performances de la recherche de documents en liaison avec une requête (cf. notamment [Salton & Allan 1993]).

1.6. Segmentation et indices linguistiques

Les méthodes exposées dans les sections précédentes segmentent les textes en faisant toutes appel au même grand principe, consistant schématiquement à calculer une mesure de similarité entre des vecteurs représentant de larges passages de texte. L'approche présentée dans cette section suit une logique différente puisqu'elle fonde la segmentation sur la détection d'indices linguistiques précis. Elle est principalement représentée par les travaux de Diane Litman et de Rebecca Passonneau [Litman & Passonneau 1995, Passonneau & Litman 1996, Passonneau & Litman 1993]. Ces travaux s'inscrivent dans la conception de la segmentation du discours définie par Grosz et Sidner [Grosz & Sidner 1986], en vertu de laquelle les segments sont avant tout sous-tendus par une intention.

L'étude réalisée comportent trois volets. Le premier vise à cerner plus précisément la notion de segmentation en analysant la façon dont cette tâche est réalisée par des êtres humains lorsqu'ils s'appuient sur la notion d'intention. Le matériau ayant servi de support à ce travail, et plus généralement à toute l'étude, est un peu différent de celui utilisé par les autres méthodes puisqu'il ne s'agit pas de textes à proprement parler mais de la transcription du récit réalisé oralement par différentes personnes de l'histoire d'un film. Les résultats obtenus montrent qu'en dépit d'une assez grande variabilité dans le nombre de bornes de segment placées et dans leur localisation, des tendances statistiquement significatives se dégagent tout de même et permettent de définir une segmentation de référence.

Le deuxième volet, qui est le plus intéressant pour nous ici, s'est attaché à établir dans quelle mesure cette segmentation humaine de référence peut être corrélée avec les résultats d'un processus automatique de segmentation. Pour ce faire, trois méthodes ont été testées : une fondée sur les chaînes référentielles, une autre s'appuyant sur les connecteurs et enfin, une troisième exploitant les pauses du narrateur. Pour les textes écrits, il est bien évident que cette dernière méthode est inapplicable et que la deuxième méthode doit être adaptée.

L'utilisation des chaînes référentielles pour segmenter les textes repose sur l'hypothèse avancée dans beaucoup de travaux sur la structuration du discours en vertu de laquelle le référent d'une anaphore se trouve dans la grande majorité des cas dans le même segment que celle-ci. En retournant le point de vue, on peut utiliser les chaînes anaphoriques, dans

la mesure bien sûr où l'on sait les mettre en évidence, et leurs interruptions pour localiser les frontières des segments.

L'algorithme présenté ici pour réaliser cette tâche s'inspire de [Passonneau 1993]. Il prend en entrée une succession de quadruplets représentant chacun un groupe nominal ou une expression anaphorique (comme un pronom par exemple). Chaque quadruplet s'écrit de la façon suivante : <LOC, NP, INDEX, INFER>.

LOC fait référence à l'identifiant de la proposition dans laquelle se trouve le groupe nominal et à sa position dans cette proposition. Le terme "proposition" est à prendre dans le cas présent au sens grammatical de "proposition principale". NP représente la forme de surface du groupe nominal. INDEX spécifie quant à lui l'entité qui est désignée par ce groupe nominal. INDEX peut introduire un nouvel identifiant s'il s'agit d'un objet encore jamais rencontré dans le discours ou bien reprendre un identifiant déjà créé pour un autre groupe nominal si les deux groupes font référence à la même entité. INFER rassemble enfin un ensemble de couples (type de relation, identifiant d'entité). Chacun de ces couples caractérise une relation dite inférentielle existant entre l'entité à laquelle renvoie le groupe nominal considéré et une autre entité apparaissant dans le discours. Ces relations, par exemple de type méronymique (partie/tout), rendent compte du fait que la présence d'une des deux entités implique la présence de l'autre.

À partir de ces quadruplets, l'algorithme consiste pour chaque frontière entre deux propositions (P_i, P_{i+1}) à appliquer trois tests. Si l'une au moins de ces trois conditions est remplie, on considère que le segment en cours se poursuit; sinon la frontière (P_i, P_{i+1}) marque une nouvelle borne de segment. Les tests de continuation d'un segment sont les suivants :

- la proposition P_{i+1} contient un quadruplet dont la valeur de INDEX et la même que celle d'un quadruplet de P_i. Autrement dit, les deux propositions font référence à une même entité;
- l'entité désignée par le champ INDEX de l'un des quadruplets de P_{i+1} est liée par une relation référentielle présente dans son champ INFER à une entité désignée par le champ INDEX de l'un des quadruplets de P_i. Les deux propositions font donc référence à deux entités liées par une relation impliquant la présence de l'une quand l'autre est présente;
- un pronom de la troisième personne dans P_{i+1} se voit attribuer comme référent une entité apparaissant dans le champ INDEX de l'un des quadruplets présents dans le segment courant (espace allant de la dernière borne de segment jusqu'à P_i).

Le résultat de l'algorithme est donc une succession de bornes de segments formées chacune d'une frontière entre deux propositions.

La deuxième méthode de segmentation reprend quant à elle en le simplifiant le travail décrit dans [Hirschberg & Litman 1993] sur l'interprétation des connecteurs et introducteurs ("cue words") du point de vue de la structuration du discours. L'hypothèse retenue pour fonder la segmentation pose que tout connecteur, parmi une liste fixée regroupant des mots comme *maintenant, et, parce que, mais ...*, apparaissant en tête d'une proposition marque le début d'un nouveau segment. Comme dans la première méthode, on choisit de placer les bornes de segment au niveau des frontières inter-propositions. La seule différence est que les présentes propositions sont délimitées avant tout sur des critères prosodiques. On parle alors de phrase prosodique. L'algorithme de segmentation consiste alors à parcourir les frontières entre phrases prosodiques et à poser une borne de segment à chaque fois qu'un des connecteurs de la liste établie apparaît en tête de la seconde phrase.

La dernière méthode de segmentation est celle parmi les trois qui est la plus spécifique du discours oral puisqu'elle repose sur la mesure du temps de pause du locuteur à la fin de chaque phrase prosodique. Cette méthode s'inspire en les simplifiant des travaux présentés dans [Hirschberg & Grosz 1992]. L'algorithme de segmentation développé considère ainsi qu'une frontière entre deux phrases prosodiques est une borne de segment lorsqu'il existe une pause du locuteur entre la première et la seconde phrase.

Le tableau de la figure I.1 donne les résultats obtenus par ces trois méthodes ainsi que par le jugement humain par rapport à la segmentation de référence (bornes confirmées par au moins 4 sujets sur 7). On y retrouve la précision et le rappel déjà vus précédemment et on dispose en plus des notions d'erreur et de fallout.

	Rappel	Précision	Fallout	Erreur
Humain	0,74	0,55	0,09	0,11
Anaphores	0,50	0,30	0,15	0,19
Connecteurs	0,72	0,15	0,53	0,50
Pauses	0,92	0,18	0,54	0,49

Fig. I.1 - Résultats des trois méthodes automatiques de segmentation

L'erreur représente le rapport entre le nombre de cas où il y a erreur (bornes oubliées ou au contraire, frontières improprement considérées comme bornes) et le nombre total de cas. Le fallout correspond quant à lui au rapport entre le nombre de frontières identifiées à

tort comme étant des bornes et le nombre de frontières entre propositions ou phrases prosodiques qui ne sont pas des bornes de segment.

On constate globalement que les résultats obtenus par les méthodes automatiques sont assez nettement inférieurs à ceux des juges humains. L'algorithme fondé sur les chaînes référentielles présente les résultats les plus équilibrés, ce qui peut s'expliquer par les connaissances qu'il fait intervenir. Les deux autres méthodes sont de ce point de vue comparables : elles produisent beaucoup de bornes, ce qui induit un bon rappel; mais une part importante d'entre elles ne correspondent pas à aucune borne de la segmentation de référence, fait que l'on peut constater au travers des fortes valeurs de l'erreur et du fallout et de la valeur très faible de la précision.

Passonneau et Litman ont par la suite mené une expérimentation supplémentaire en couplant les méthodes deux à deux, dans le but notamment de réduire le nombre de fausses reconnaissances. Dans cette configuration, une borne de segment n'est posée qu'avec l'accord des deux méthodes associées. Cette façon de faire permet globalement d'améliorer les résultats de manière significative, avec les valeurs les plus intéressantes pour les mesures adoptées obtenues par le couple chaînes référentielles/pauses : 0,47 pour le rappel; 0,42 pour la précision; 0,08 pour le fallout et enfin, 0,13 pour l'erreur. Ces chiffres restent malgré tout en deçà de ceux des juges humains.

La conclusion avancée par Passonneau et Litman est qu'un apport de connaissances supplémentaires devrait permettre à une méthode automatique d'égaler les performances humaines. Cependant, compte tenu de la grande variabilité inter-sujet du jugement humain, il semble nécessaire de toute manière de prévoir des mécanismes d'adaptation aux conditions particulières imposées par chaque locuteur et de considérer que les bornes des segments ont généralement un caractère de flou ne rendant pas toujours possible une localisation très précise.

Le troisième et dernier volet de ce travail sur le problème de la segmentation concerne une autre façon d'améliorer les résultats obtenus par les méthodes automatiques. À partir d'un corpus annoté, c'est-à-dire dans lequel figurent à la fois les bornes de référence et les informations utilisées par les trois méthodes précédentes, un algorithme de classification automatique, en l'occurrence C4.5 [Quinlan 1993], a été appliqué afin de construire un classifieur ayant pour objectif de déterminer si une frontière entre deux phrases prosodiques est ou n'est pas une borne de segment.

Les résultats montrent une amélioration sensible des performances par rapport aux méthodes automatiques précédemment étudiées puisqu'ils deviennent comparables à ceux obtenus par les juges humains. L'utilisation de ces techniques d'apprentissage posent

néanmoins le problème d'une spécialisation sans doute un peu trop grande du segmenteur construit vis-à-vis du corpus d'entraînement.

Dans la dichotomie faite entre les méthodes de segmentation opérant seulement à partir des caractéristiques intrinsèques des textes et celles reposant sur des connaissances extérieures, le travail de Passonneau et Litman se situe un peu à cheval entre les deux. Si l'utilisation des connecteurs et des pauses s'inscrit plutôt dans la première tendance, l'exploitation des chaînes référentielles oblige à faire intervenir des connaissances afin d'établir les relations possibles entre les entités désignées. Un des intérêts de leur travail est à cet égard de montrer qu'une coopération entre des méthodes complémentaires du point de vue du profil de leurs résultats peut faire état de performances dépassant celles dont ces méthodes font preuve lorsqu'elles sont appliquées individuellement.

En revanche, deux points posent problème dans le cadre qui est le nôtre. Le plus évident est que la méthode reposant sur les pauses entre phrases prosodiques n'est pas applicable aux textes écrits. Celle reposant sur les connecteurs demanderait quant à elle des adaptations pour un tel usage : une transposition de la notion de phrase prosodique serait nécessaire, de même qu'un réexamen de la liste des connecteurs et des introducteurs car il existe très certainement des différences dans l'usage de ces mots entre le discours écrit et le discours oral.

Le second problème résulte pour sa part d'une certaine incertitude dans la possibilité de mettre en œuvre la méthode exploitant les chaînes référentielles. Celle-ci suppose en effet que les anaphores soient résolues, ce que l'on ne peut pas tenir comme un fait acquis du point de vue du traitement automatique des langues. Dans le travail de Passonneau et Litman, la part des interventions manuelles est d'ailleurs assez importante, aussi bien pour la mise en évidence des propositions que pour la constitution des quadruplets. Ces interventions manuelles sont acceptables si l'on considère qu'elles sont mieux cernées que les interventions qui consisteraient à placer les bornes des segments et que de ce fait, l'algorithme proposé permet de préciser la notion de segmentation.

Elles soulèvent en revanche des difficultés du point de vue d'un traitement automatique, d'autant plus que la notion de segment peut être utilisée pour la résolution des anaphores. La mise en œuvre complète de cet algorithme va donc sans doute de pair avec l'élaboration d'un mécanisme de coopération étroite avec les processus de résolution des anaphores. Elle reste cependant à réaliser.

2. Les approches à base de connaissances de la segmentation thématique

2.1. Les chaînes lexicales

2.1.1. Morris et Hirst

L'idée des chaînes lexicales a été développée par Jane Morris et Graeme Hirst afin de rendre compte de la cohésion des textes sur le plan lexical. La mise en évidence de cette cohésion peut être utile directement au niveau lexical, en contribuant à la désambiguïsation du sens des mots par la définition d'un contexte local d'interprétation, mais également au niveau de la structuration du discours, en fournissant un moyen de caractériser des zones de texte relatives à un même sujet. Ce second axe est celui qui nous intéresse plus spécifiquement ici et qui a été développé dans [Morris & Hirst 1991]. Il repose sur l'hypothèse stipulant que la cohésion étant un reflet de la cohérence sous-jacente à un discours, il est possible de détecter les ruptures de cette dernière, et donc les frontières entre des segments de ce discours, à partir des ruptures observées dans sa cohésion, notamment au niveau lexical.

La notion de cohésion textuelle en général et celle plus spécifique de cohésion lexicale ont été particulièrement développées par Halliday et Hasan dans [Halliday & Hasan 1976]. Ces deux linguistes ont proposé de faire reposer la cohésion lexicale sur cinq types de relations entre les mots (m1 et m2 représentant deux mots) :

- m1 est identique à m2 et m1 et m2 désignent la même entité;
- m1 est identique à m2 mais m1 et m2 ne désignent pas la même entité;
- m1 (respectivement m2) est un mot désignant un super-ordonné du concept désigné par m2 (respectivement m1). Exemple : m1 = *fruit* et m2 = *pomme*;
- m1 et m2 sont liés par une relation sémantique systématique. Parmi ces relations, on compte principalement les relations de synonymie, d'antonymie, de méronymie ainsi que l'appartenance à une même classe générale (par exemple les chiffres ou les couleurs);
- m1 et m2 sont liés par une relation sémantique non systématique. Des mots comme *hôpital, médecin, infirmière* ou *scalpel* entretiennent une relation de ce type. Celle-ci traduit le fait que les entités désignées par ces mots interviennent dans un même contexte, dans une même situation.

L'hypothèse faite par Morris et Hirst est que la cohésion lexicale se matérialise par l'existence dans les textes de suites de mots liés entre eux par l'une des cinq relations définies ci-dessus. Ces suites de mots sont appelées des *chaînes lexicales*. Pour qu'un mot fasse partie d'une telle chaîne, il faut par ailleurs qu'il ne soit pas éloigné de plus d'une certaine distance, d_{lex}, à la fois du mot qui le suit et de celui qui le précède. Les chaînes lexicales peuvent être de tailles très diverses. Certaines couvrent un texte entier tandis que d'autres ne s'étendent que sur quelques phrases. Au sein d'un même texte, il est d'autre part fréquent que plusieurs chaînes se superposent. Une chaîne rendant compte du thème général d'un texte peut couvrir l'ensemble de ce texte tandis que d'autres chaînes, plus courtes et en phase avec les différents sous-thèmes abordés, viennent se superposer à elle. Enfin, il est possible d'avoir des chaînes lexicales discontinues. Ce phénomène intervient lorsque le début d'une nouvelle chaîne peut se raccrocher à une autre chaîne précédemment mise en évidence dans le texte mais considérée comme close. On parle alors d'un retour de chaîne.

Les travaux tels que ceux de Hearst, Nomoto et Nitta, Reynar ou encore Salton reposent aussi sur la notion de cohésion lexicale mais ils n'exploitent que les deux premières relations parmi les cinq définies par Halliday et Hasan. Cela leur permet de ne pas dépendre d'une source de connaissances mais, à l'inverse, cela limite également leur champ d'intervention à des textes caractérisés par un vocabulaire fortement récurrent. Morris et Hirst vont plus avant dans l'utilisation de la cohésion lexicale en s'appuyant également sur les relations de hiérarchie et les relations sémantiques systématiques. Pour les mettre en évidence, ils ont choisi de faire appel à un thesaurus, en l'occurrence le *Roget's Thesaurus* [Roget 1977].

Celui-ci est formé de 1042 catégories regroupant chacune un ensemble de mots relatifs à un même sujet de base (la mort, la vue, les moyens de transport, le logement ...). Ces catégories s'insèrent dans une structure hiérarchique dont elles forment le quatrième niveau. Au sommet, on trouve huit grandes classes (relations abstraites, espace, matière, physique, sensations, affects, intellect, volonté). Ces classes sont divisées en sous-classes, elles-mêmes subdivisées en sous sous-classes au sein desquelles les catégories ont été définies par paires antagonistes, lorsque cela était possible. La catégorie *Vie* est par exemple accompagnée de la catégorie *Mort.* Les catégories possèdent elles-mêmes une structure. Les mots liés au sujet de la catégorie sont d'abord regroupés en paragraphes sur le critère de leur catégorie morpho-syntaxique. Au sein d'un paragraphe, des sous-groupes sont distingués suivant les proximités de sens entre les mots. Des relations permettent par ailleurs de lier un sous-groupe aux catégories qui lui sont les plus proches. Le thesaurus est muni d'un index donnant pour un mot l'ensemble

des catégories dans lesquelles il apparaît. La catégorie est dans ce cas accompagnée d'un autre mot donnant une indication sur le sens du mot couvert par la catégorie en question.

La construction des chaînes lexicales d'un texte repose pour l'essentiel sur la mise en évidence entre les mots de ce texte de certains liens existant au sein de ce thesaurus. On distingue cinq liens possibles entre un mot m2 à un mot m1 permettant de rattacher m2 à la chaîne lexicale à laquelle appartient m1 :

- les entrées d'index pour m1 et m2 font apparaître une catégorie commune;
- l'entrée d'index pour m1 fait référence à une catégorie liée (par l'intermédiaire d'un sous-groupe de l'un de ses paragraphes) à une des catégories apparaissant dans l'entrée d'index pour m2;
- m2 (respectivement m1) est un mot accompagnant l'une des catégories de l'entrée d'index de m1 (respectivement m2) ou bien m2 (respectivement m1) est l'intitulé d'une des catégories de l'entrée d'index de m1 (respectivement m2);
- les entrées d'index pour m1 et m2 font apparaître deux catégories appartenant à la même sous sous-classe et constituant une paire d'antagonistes;
- les entrées d'index pour m1 et m2 font apparaître deux catégories pointant (par l'intermédiaire d'un sous-groupe de l'un de leurs paragraphes) vers une même catégorie.

Parmi ces cinq types de liens, les deux premiers recouvrent approximativement 90% des liens trouvés dans les textes étudiés.

La procédure de construction de ces chaînes lexicales commence par un pré-traitement des textes visant, comme dans tous les travaux précédents, à supprimer les mots grammaticaux ainsi que les mots jugés comme non informatifs du fait de leur grande fréquence. Les textes sont ensuite traités mot par mot. Le coeur du processus consiste à tenter de rattacher le mot courant à l'une des chaînes lexicales ou l'un des embryons de chaîne lexicale actifs en mémoire. Les chaînes actives sont les chaînes que l'on considère en cours de construction, c'est-à-dire celles dont le dernier mot est suffisamment proche du mot en cours de traitement pour que celui-ci y soit rattaché. La distance maximale, dl_{ex}, au delà de laquelle un tel rattachement devient impossible, provoquant de fait la clôture de la chaîne, est ici égale à trois phrases. Elle est susceptible de changer afin de s'adapter au style spécifique des textes. Les embryons de chaîne lexicale sont quant à eux les mots n'ayant pu être rattachés à une chaîne lexicale existante. Ils sont considérés comme actifs tant qu'ils ne sont pas situés à une distance supérieure à dl_{ex} du mot en cours de traitement.

Le rattachement d'un mot à une chaîne lexicale s'effectue soit lorsque ce mot est identique au dernier mot de la chaîne, soit lorsque l'on trouve un lien par l'intermédiaire du *Roget's Thesaurus*, parmi les cinq types de liens présentés ci-dessus, entre ce mot et le dernier mot de la chaîne. Dans le cas des embryons de chaîne, la procédure de rattachement est la même sachant que le dernier mot de la chaîne s'identifie alors à l'unique mot composant l'embryon. Le rattachement à une chaîne lexicale bénéficie d'une possibilité supplémentaire. Le lien entre le mot courant, m_{cour}, et le dernier mot de la chaîne, m_{finch}, peut en effet ne pas être direct et passer par un autre mot de la chaîne, m_{intch}. On peut donc avoir le schéma de rattachement m_{cour} — m_{finch} ou le schéma m_{cour} — m_{intch} — m_{finch} (— représentant un lien trouvé par l'intermédiaire du thesaurus).

L'algorithme de construction des chaînes lexicales assure également la mise en évidence des chaînes formées de plusieurs tronçons non contigus. Lorsque toutes les possibilités de rattachement du mot courant ont été passées en revue sans succès, la limite de distance imposée par d_{lex} est levée et un rattachement parmi toutes les chaînes déjà construites est cherchée. S'il est trouvé, le mot courant marque le début d'un nouveau tronçon de la chaîne cible du rattachement. Ce segment peut éventuellement ne comporter qu'un seul mot. En revanche, le premier tronçon d'une telle chaîne doit à l'évidence comporter au moins deux mots.

L'algorithme présenté ci-dessus n'a pas été implémenté du fait de l'absence sous une forme électronique accessible du *Roget's Thesaurus*. En revanche, il a été appliqué manuellement sur 5 textes extraits de magazines généraux représentant un total de 183 phrases.

Dans la seconde partie de leur travail, Morris et Hirst ont étudié la corrélation existant dans les textes entre les chaînes lexicales trouvées par l'algorithme ci-dessus et la structure de ces textes. Le modèle de structuration du discours retenu pour rendre compte de cette structure était dans le cas présent celui de Grosz et Sidner, présenté au chapitre précédent. La détermination de la structure intentionnelle des textes a été réalisée manuellement et comparée aux chaînes lexicales. Morris et Hirst ont tout particulièrement porté leur attention sur le recoupement entre les bornes de segment mises en évidence en utilisant le modèle de Grosz et Sidner et les frontières des chaînes ainsi que celles des tronçons lorsque les chaînes étaient discontinues. Les résultats donnés ne sont que qualitatifs dans la mesure où aucune formalisation de cette procédure de comparaison n'a été faite. Par ailleurs, l'analyse intentionnelle de référence n'a pas fait l'objet d'une validation suivant un protocole tel que celui utilisé par Passonneau et Litman.

Globalement, on observe qu'il existe effectivement une corrélation entre les chaînes lexicales et la structure des textes mais cette corrélation n'est pas forcément simple à exploiter pour segmenter les textes, en particulier de manière automatique, dans la mesure où les points de corrélation ne sont pas systématiques. L'illustration en est donnée par un ensemble de points. Le plus significatif d'entre eux est certainement le fait que certains segments ne peuvent être mis en relation avec aucune chaîne lexicale. Ils ne sont alors détectables que par différence, lorsqu'ils recouvrent des zones de texte elles-mêmes non recouvertes par un autre segment.

Rappelons que dans le modèle de Grosz et Sidner, les segments sont organisés hiérarchiquement. Un segment peut ainsi n'être composé que de segments. Pour retrouver cette structure, il faut donc avoir une chaîne pour ce segment mais également une chaîne pour chacun des segments qui le composent. Or, on observe en pratique que tous ces niveaux n'ont pas forcément leur équivalent en termes de chaînes lexicales. D'autre part, il faut noter que la relation entre chaîne et segment n'est pas forcément univoque : plusieurs chaînes peuvent ainsi faire référence au même segment et à l'inverse, une même chaîne peut recouvrir plusieurs segments. Ce dernier cas est illustratif d'une insuffisance de discrimination, due à un manque de connaissances.

Lorsqu'un segment est effectivement représenté par une chaîne lexicale, il faut également faire face à une certaine variabilité dans la forme de la chaîne, laquelle complique la mise en évidence du segment. La plupart des chaînes lexicales possèdent la même extension que le segment avec lequel elles sont mises en correspondance (moyennant une tolérance de l'ordre de deux phrases). Dans une minorité des cas toutefois, la chaîne étant très courte, elle ne permet que d'identifier la fin du segment. Les plus problématiques sont cependant les chaînes dont l'extension est significativementplus faible que celle du segment auquel elles sont associées, sans être pour autant très courtes. Elles ne permettent en effet de situer les bornes du segment que de façon très approximatives.

En dépit de ces limitations, les chaînes lexicales constituent un indicateur intéressant de la structure des textes, notamment par leur capacité à en appréhender la dimension hiérarchique. Une façon de rendre leur exploitation plus aisée serait certainement d'augmenter l'homogénéité de ces chaînes en s'appuyant sur des connaissances plus étendues. Certains liens intéressants entre les mots relèvent en effet de relations sémantiques non systématiques (liaison par l'appartenance à un même contexte, à une même situation), lesquelles sont absentes d'un thesaurus.

L'utilisation d'un thesaurus pose également un problème sur le plan de la disponibilité de cette source de connaissances, ainsi que l'atteste l'absence d'implémentation de l'algorithme de Morris et Hirst. C'est pourquoi certains travaux [Hirst & St-Onge 1995,

Stairmand 1994] ont exploré la possibilité d'exploiter une ressource telle que WordNet [Miller et alii 1989] pour mettre en évidence des chaînes lexicales. WordNet repose sur des relations sémantiques systématiques entre mots (principalement la synonymie) et présente l'avantage à la fois d'être accessible sans restrictions sous forme électronique et d'avoir été conçu pour ce type d'utilisation. Les travaux évoqués se sont néanmoins intéressés au problème de la désambiguïsation du sens des mots et non à la segmentation du discours. Or, un réseau lexical comme WordNet est moins riche qu'un thesaurus sur le plan de la variété des relations présentes entre les mots. Par ailleurs, il ne remédie pas à l'insuffisance des thesaurus puisqu'il ne comporte pas de relations sémantiques liées aux situations. Il n'est donc pas évident qu'il soit véritablement utilisable pour servir de support à la segmentation thématique des textes.

2.1.2. Okumura et Honda

Le travail d'Okumura et de Honda décrit dans [Okumura & Honda 1994] s'inscrit assez directement dans le prolongement de celui de Morris et Hirst. Okumura et Honda ont en effet implémenté un algorithme de construction de chaînes lexicales proche de celui décrit par Morris et Hirst et évalué ses résultats sur un petit ensemble de textes. Par ailleurs, ils ont également implémenté et évalué un algorithme de segmentation des textes fondé sur les chaînes lexicales ainsi obtenues.

L'algorithme de construction des chaînes lexicales présente la particularité de combiner cette construction avec une désambiguïsation parallèle du sens des mots. Okumura et Honda justifient cette association par la nécessité de lever les ambiguïtés sur le sens des mots le plus rapidement possible afin d'éviter la construction de chaînes lexicales erronées. L'opération s'effectue en deux temps. Un thesaurus[1] est d'abord consulté afin de trouver des relations de cohésion lexicale au sein de chaque phrase, considérée comme une unité privilégiée pour désambiguïser le sens des mots. Les liens possibles ont été simplifiés par rapport à ceux de Morris et Hirst : deux mots sont liés seulement s'ils possèdent dans leur entrée d'index une référence vers une même catégorie. Lorsque deux mots d'une phrase ont plusieurs sens possibles, on choisit pour chacun d'entre eux celui de ses sens renvoyant à une catégorie commune aux deux mots. Ce principe n'est toutefois pas systématiquement applicable car les mots ne renvoient pas nécessairement à une catégorie commune.

[1] Il s'agit d'un thesaurus japonais disponible sous forme électronique et possédant une structure proche de celle du *Roget's Thesaurus.*

Le second temps est la construction proprement dite des chaînes lexicales, qui s'accompagne d'un autre mécanisme de désambiguïsation du sens des mots. La spécificité de la méthode considérée ici est que le rattachement d'un mot à une chaîne ne s'appuie plus seulement sur les liens trouvés dans le thesaurus mais également sur l'ordre décroissant de la saillance des chaînes. La saillance d'une chaîne lexicale est en l'occurrence déterminée par le degré de récence de sa dernière mise à jour ainsi que par sa longueur. Elle est d'autant plus forte que les chaînes sont plus récentes et plus longues. Un mot est donc rattaché à la première chaîne de la liste des chaînes en construction avec laquelle un lien est trouvé par le biais du thesaurus. Ce rattachement provoque une modification de la saillance de la chaîne concernée et donc, un réordonnancement de la liste des chaînes en construction.

La désambiguïsation repose également sur cette liste. Lorsque les liens d'un mot M_{cour} possédant plusieurs sens avec une chaîne interviennent toujours par l'intermédiaire de la même catégorie du thesaurus, c'est elle qui est retenue comme sens de ce mot dans le contexte courant. Autrement, deux cas sont possibles. Si M_{cour} est lié à la chaîne située en tête de la liste des chaînes en construction, il est rattaché à celle-ci; sinon, on ajoute M_{cour} aux différentes chaînes de rattachement possible. On conserve alors l'ambiguïté sur le sens de M_{cour} dans le contexte courant. Cette ambiguïté peut être néanmoins levée par les mots qui suivent. Si l'un d'entre eux n'est pas ambigu et qu'il amène en tête de la liste des chaînes en construction une des chaînes dans lesquelles M_{cour} a été ajouté, alors le sens retenu pour M_{cour} est la catégorie du thesaurus ayant permis le rattachement de M_{cour} à cette chaîne. Dans ce cas, M_{cour} est supprimé des autres chaînes dans lesquelles il avait été ajouté.

À la suite de la construction des chaînes, la segmentation des textes est réalisée en tirant assez directement les conséquences de la correspondance supposée entre les bornes de segment et les bornes des chaînes. À chaque espace entre deux phrases *n* et *n+1,* on comptabilise le nombre de chaînes se terminant à la phrase *n* et le nombre de chaînes commençant à la phrase *n+1*. La valeur de cette mesure donne une indication sur la probabilité d'avoir une borne de segment à chaque espace inter-phrase. La précision et le rappel moyens mesurés pour un ensemble de cinq textes s'élèvent respectivement à 0,25 et 0,52, les valeurs étant assez variables en fonction des textes.

Ces résultats ne sont pas très bons comparés à certains donnés précédemment mais le travail réalisé présente l'avantage d'avoir poussé jusqu'au stade de l'implémentation et de l'évaluation, même si c'est au prix de quelques simplifications, les principes originellement développés par Morris et Hirst. Par ailleurs, la combinaison de la construction des chaînes lexicales et de la désambiguïsation lexicale semble une idée intéressante.

2.2. Lexical Cohesion Profile

(cf. §1.4 du chapitre 9)

Annexe J

Outils et résultats de SEGAPSITH

1. Outils de gestion de la mémoire des signatures thématiques

De façon similaire à ce qui se passe avec la segmentation thématique dans SEGCOHLEX, la mise au point du mécanisme de construction des signatures thématiques est une phase particulièrement importante en raison du nombre de paramètres à ajuster : seuil de similarité, fonction de calcul de la similarité, poids des mots, fonction d'activation des signatures, présence ou non des mots inférés, etc.

C'est pourquoi nous avons élaboré l'outil de gestion de la mémoire des signatures thématiques illustré par la figure J.1. Celui-ci permet de contrôler l'intégration dans la mémoire des signatures d'un ensemble d'Unités Thématiques Lexicales (UTLs) supposées issues de la segmentation thématique de SEGAPSITH. L'outil permet de mener pas à pas toutes les étapes de cette intégration : activation des signatures de la mémoire à partir d'une UTL, calcul de la similarité de cette UTL avec les signatures les plus activées, agrégation de l'UTL avec une signature ou création d'une nouvelle signature. Il autorise en outre la simulation de l'agrégation d'une UTL et d'une signature.

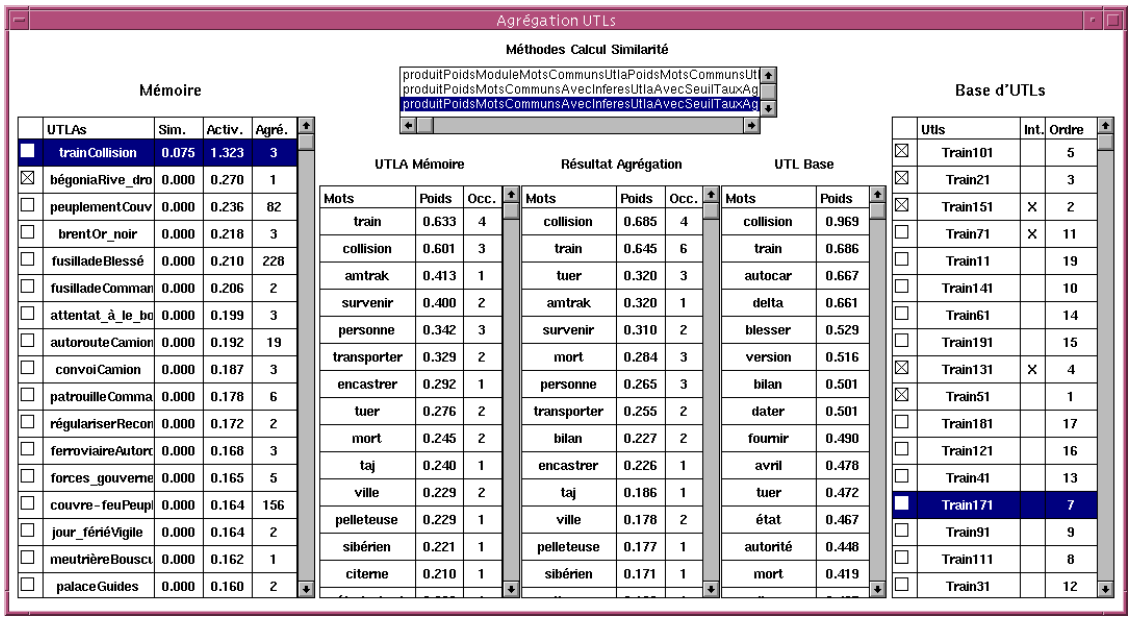


Fig. J.1 - Outil de gestion de la mémoire des signatures thématiques

L'utilisateur peut visualiser le résultat de l'opération et le valider s'il en est satisfait. La modification effective de la signature considérée n'est réalisée qu'après l'intervention de cette validation. L'outil permet également la mémorisation en séquence d'une suite d'UTLs en définissant de plus l'ordre dans lequel ces UTLs sont présentées.

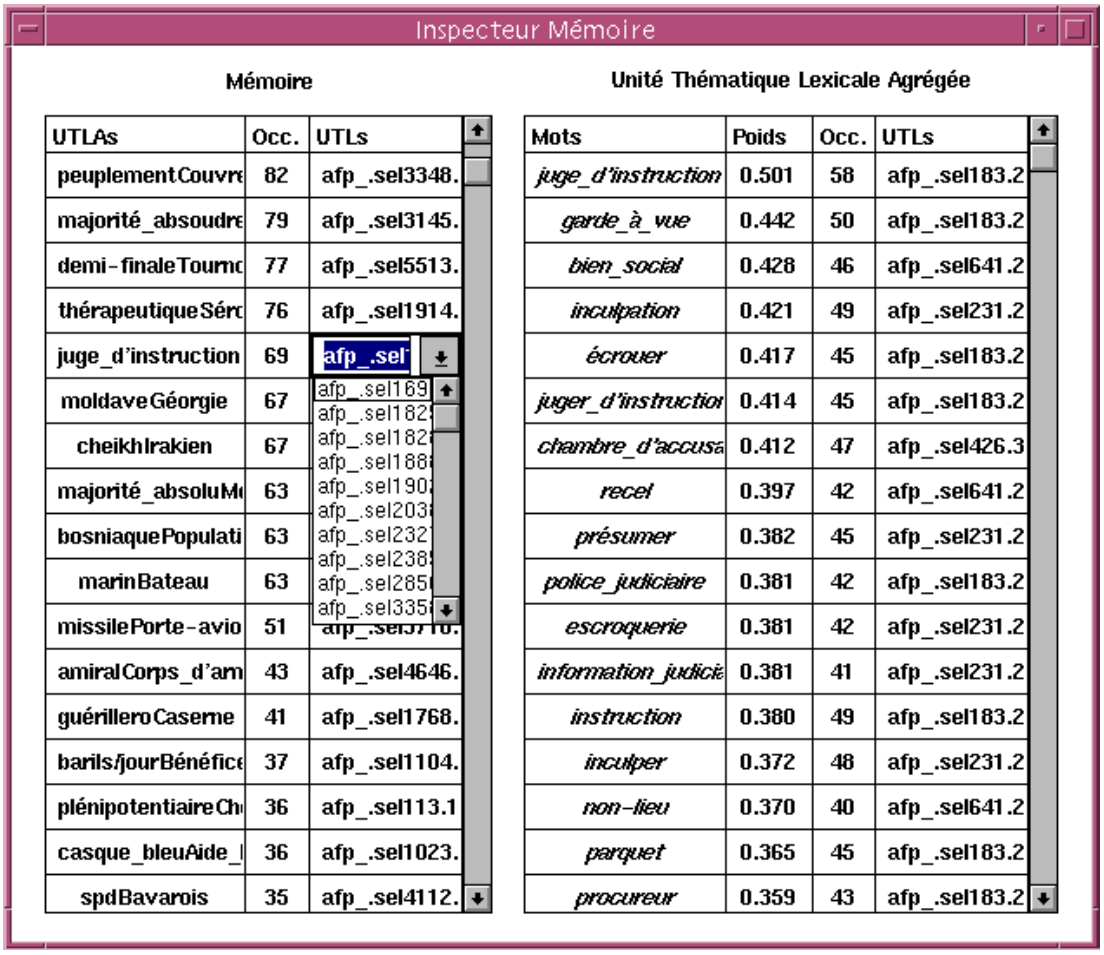


Fig. J.2 - Outil de visualisation de la mémoire des signatures thématiques

À partir de ce gestionnaire, il est possible d'inspecter plus finement la mémoire des signatures thématiques en lançant l'inspecteur représenté par la figure J.2. Grâce à lui, on peut savoir en particulier quels sont les textes, et même plus précisément les UTLs, ayant contribué à la formation d'une signature. Cette information est également disponible au niveau des mots composant une signature.

2. Vue d'ensemble des signatures thématiques obtenues

Afin de donner une vue plus globale des signatures thématiques construites à partir des dépêches de l'AFP du mois de mai 1994, la figure J.3 montre le résultat d'une

classification hiérarchique réalisée sur celles de ces signatures considérées comme stables, c'est-à-dire les signatures dont le nombre d'agrégations est au moins égal à 20. Cette classification hiérarchique a été réalisée suivant l'algorithme classique dans ce domaine :

Fig. J.3 - Hiérarchie des signatures thématiques obtenues après classification hiérarchique[1]

[1] L'algorithme de classification hiérarchique a été implémenté par Brigitte Grau.

à chaque étape, on recherche le couple de signatures les plus proches et on les fusionne. L'algorithme est ainsi appliqué jusqu'à ce qu'il n'y ait plus de fusion à réaliser. Dans le cas présent, la distance retenue est le nombre de mots communs et le critère d'arrêt des fusions est donc l'absence de mots communs. Le résultat de la fusion de deux signatures est l'ensemble des mots communs aux deux signatures.

La figure J.3 montre que le résultat de cette classification est une forêt d'arbres. Chaque nœud est représenté par les deux mots ayant le poids le plus fort parmi ceux constituant le nœud en question. Les feuilles sont les signatures composant la mémoire initialement construite tandis que les nœuds intermédiaires sont les résultats des différentes fusions effectuées. Le chiffre accompagnant le couple de mots de plus fort poids pour chacun de ces nœuds intermédiaires représente le nombre de mots que ce nœud rassemble, c'est-à-dire le nombre de mots communs à ses deux nœuds-fils.

La structuration ainsi faite des signatures thématiques montre que celles-ci peuvent être regroupées en grands domaines thématiques cohérents (sport, politique étrangère, politique nationale, économie, etc.) et que leur hiérarchisation, évoquée au chapitre 10 comme une extension possible, présente effectivement un intérêt certain.