

Modélisation en sémantique

Patrice Enjalbert, Bernard Victorri

▶ **To cite this version:**

Patrice Enjalbert, Bernard Victorri. Modélisation en sémantique. Gérard Sabah. Compréhension des langues et interaction, Hermès, pp.71-110, 2006. halshs-00009905

HAL Id: halshs-00009905
https://shs.hal.science/halshs-00009905

Submitted on 2 Apr 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Modélisation en sémantique

Patrice Enjalbert et Bernard Victorri

1. Présentation

1.1. La place de la modélisation entre théorie linguistique et TAL

La sémantique — c'est-à-dire l'accès au sens des énoncés et des textes — est souvent présentée en TAL comme une « ambition » (selon l'expression d'A. Nazarenko dans ce même volume) difficile à atteindre. Pourtant certaines tâches comme le dialogue homme-machine et l'extraction d'information nécessitent clairement une composante de ce type. En outre, un certain nombre d'autres applications posent par nature une question sémantique, même si les techniques courantes tentent souvent d'en faire plus ou moins l'économie : traduction automatique, résumé automatique, procédures avancées de recherche d'information telles que les systèmes de question/réponse… On se heurte ici à des verrous technologiques dont on peut aisément diagnostiquer qu'ils réclament pour être levés le développement de méthodes sémantiques. Mais le croisement entre sémantique linguistique et TAL a aussi une pertinence dans l'autre sens : elle permet de faire progresser notre compréhension des phénomènes sémantiques eux-mêmes à partir de traitements automatiques. Il s'agit alors de construire des modèles, formalisés par des méthodes mathématiques, logiques ou d'Intelligence Artificielle, qui peuvent être rendus opérationnels grâce à des dispositifs informatiques appropriés. Comme dans d'autres disciplines scientifiques, cette mise en œuvre informatique est un moyen pour expérimenter les modèles et valider, ou faire progresser, les théories. On tirera particulièrement profit d'études sur de très gros corpus, aujourd'hui facilement disponibles. Ce type de travaux est crucial pour le développement d'une approche sémantique en TAL, visant à lever les verrous dont nous venons de parler.

1.2. Les paliers de la sémantique

Pour en présenter les différents aspects, nous avons choisi de structurer notre exposé par niveaux, ou paliers (cf. [ENJ 05a]) correspondant à une complexité croissante des objets linguistiques à prendre en compte. Le premier niveau est celui du lexique, c'est-à-dire des constituants élémentaires porteurs de sens. Puis vient le niveau du syntagme et de la phrase, assemblage structuré d'unités en un tout cohérent porteur d'un sens relativement autonome. Et finalement celui du texte ou du discours, suite de phrases plus ou moins fortement liées, qui constitue la forme globale que prend, en général, le message linguistique. Ce point de vue structurel doit être complété par une perspective qui resitue le texte/discours dans un acte de communication, laissant toute sa place au libre jeu de l'interprétation du lecteur/auditeur, et que l'on désigne communément sous le terme de pragmatique. Ce que certains verront comme un au-delà de la sémantique constitue plutôt pour nous un regard particulier opérant aux trois niveaux. Il faut d'ailleurs insister sur le fait que ces paliers sont tout sauf étanches. La langue n'est pas un jeu de poupées russes ! Par exemple la sémantique d'un item lexical fait intervenir le contexte d'énonciation et le cotexte qui lèvent les ambiguïtés et en ajustent en quelque sorte le sens. À l'inverse, bien des mécanismes que nous décrirons ici au niveau du texte apparaissent en fait dès celui de la phrase.

1.3. Une double problématique : représentations/opérations

Les questions que nous nous poserons sont les suivantes : Quelles valeurs sémantiques peuvent être associées aux objets linguistiques relevant de ces différents niveaux ? Quelles structures les organisent ? Comment décrire les unes et les autres, c'est-à-dire comment les représenter d'un point de vue formel dans une optique de TAL ? Et d'un autre côté, quelles opérations permettent de créer ces valeurs et ces structures dans ce que l'on appelle la construction du sens ? C'est cette double problématique

représentation/opérations qui sera le fil conducteur de ce chapitre. Pour chaque palier, nous présenterons d'abord cette double problématique avant de montrer un exemple de modèle sur un point particulier, à titre d'illustration.

2. Le palier du lexique

Le premier niveau de la modélisation sémantique concerne donc les constituants élémentaires porteurs de sens, mais, comme souvent en linguistique, cette définition en apparence simple n'est pas si facile à mettre en œuvre. Des lexies comme *belle-mère*, *œil-de-perdrix* ou *pied-à-terre* sont composées d'unités plus simples porteuses de sens, mais celles-ci ne permettent pas de calculer de sens de ces expressions, qui sont dites *non compositionnelles*. On aura donc tout intérêt à les traiter comme des unités lexicales à part entière, au même titre que les unités indécomposables. Le problème se pose alors des cas intermédiaires : il existe en effet toute une gradation d'expressions partiellement compositionnelles, de *pomme de terre* à *chou-fleur* et de *porte-parole* à *porte-plume*. Un problème analogue se pose pour des préfixes (*re-*, *anti-*, *hyper-*, etc.) et des suffixes (*-able*, *-eur*, *-ment*, etc.) qui restent très productifs en français contemporain : *reschtroumpfer* signifie 'schtroumpfer de nouveau' et *schtroumpfable* 'qui peut être schtroumpfé', quel que soit le sens du verbe schtroumpfer. On doit donc les considérer comme des unités porteuses de sens, mais il faut alors prendre garde à la non-compositionnalité de certaines formes : *regarder* ne veut pas dire 'garder de nouveau' et *expert-comptable* ne signifie pas 'expert qu'on peut compter', ni d'ailleurs 'expert sur lequel on peut compter' !

Parmi les constituants élémentaires porteurs de sens, on distingue généralement les unités *lexicales* et les unités *grammaticales*[1]. Les unités grammaticales, appelées aussi mots outils, ont un rôle plus structurel dans la langue : il s'agit des déterminants, prépositions, conjonctions, etc. mais aussi des morphèmes flexionnels comme les marques de temps sur les verbes ou les marques de singulier et pluriel sur les noms. Les unités lexicales sont plus directement référentielles au sens où elles évoquent à elles seules des catégories d'objets du monde, de concepts, d'événements ou de qualités de ces objets ou événements : ce sont en général des noms, des verbes, des adjectifs, ou encore des adverbes. Nous allons nous centrer dans cette section sur les unités lexicales, qui posent des problèmes spécifiques de représentation sémantique et de calcul du sens. En ce qui concerne les unités grammaticales, leur sens est entièrement lié à des calculs qui ne peuvent s'effectuer qu'au niveau de la phrase, voire du texte, comme nous aurons l'occasion de le voir plus en détail pour les marques de temps verbal (§ 5).

2.1. Représentation du sens lexical

Sans aucune prétention à une vue universelle et univoque de ce vaste champ linguistique (voir [NYC 98]), nous proposerons ici de distinguer trois catégories de modèles, correspondant à la fois à trois manières de concevoir le lexique et à trois modes de représentation formalisées.

2.1.1. Représentations relationnelles

On peut regrouper sous cette appellation deux types de modèles, issus de traditions et de méthodologies différentes, mais qui utilisent le même type d'outil mathématique : les graphes. La première tradition vient de l'intelligence artificielle : c'est le modèle des *réseaux sémantiques*. Conçus, à l'origine comme un modèle de description du lexique en psychologie cognitive (voir [CAR 89] p. 103 sq.), les réseaux sémantiques ont surtout été utilisés par la suite pour représenter des relations entre concepts. De nombreuses variantes ont été proposées, le caractère commun étant une représentation sous forme de graphes dont les nœuds sont étiquetés par des mots et les arcs par des relations entre les concepts associés à ces mots. La forme la plus aboutie des réseaux sémantiques est sans doute le modèle des graphes conceptuels de Sowa [SOW 84], qui s'étend, d'ailleurs, bien au-delà du lexique, à la représentation du sens au niveau de la phrase, voire du texte.

1. Voir notamment Talmy [TAL 00], p. 21, qui définit deux « sous-systèmes » aux fonctions sémantiques nettement différenciées : le sous-système grammatical détermine la structure de la représentation cognitive évoquée par un énoncé, et le sous-système lexical son contenu.

La deuxième tradition est plus directement lexicographique : il s'agit de construire un réseau de relations purement lexicales : synonymiques, antonymiques, hyperonymiques, méréologiques (partie-tout), etc. Cela a donné des outils informatiques tels que *WordNet* [MIL 90], ou, pour le français, le *Sémiographe* [DUT 03]. En raison du phénomène de la polysémie, sur lequel nous reviendrons ci-dessous, un même mot occupe plusieurs positions dans un réseau de ce type dès qu'il possède plusieurs sens : autrement dit, les nœuds de ces réseaux ne sont pas les unités lexicales en tant que telles, mais plutôt les *sens* de ces unités. Cela explique que la différence entre ces réseaux lexicographiques et les réseaux sémantiques soit en pratique peu sensible, voire délibérément ignorée dans certains travaux : si un mot n'est pas de même nature qu'un concept, on identifie plus facilement, du moins dans certains cadres théoriques, un sens d'un mot et un concept étiqueté par ce mot.

Ce qui caractérise ces représentations, c'est qu'il n'y a pas à proprement parler de définition du sens ou du concept associé à une unité : chaque concept n'est caractérisé que par les relations qu'il entretient avec les autres concepts. Du point de vue linguistique, cela n'est pas sans rappeler la théorie saussurienne selon laquelle le signe n'a de valeur que par sa position dans le système de l'ensemble des signes. Mais cela se révèle aussi assez efficace pour stocker des connaissances, y compris de nature encyclopédique, sur les objets du monde désignés par ces signes. Ainsi, pour reprendre un exemple classique, il suffira d'indiquer que 'plume' est dans une relation partie-tout avec 'oiseau' et que 'canari' est en relation d'hyponymie avec 'oiseau' pour en déduire que les canaris ont des plumes. Pour la psychologie cognitive, c'est d'ailleurs en tant que modèle de *mémoire sémantique* que les réseaux sémantiques ont été utilisés, pour rendre compte de la plus ou moins grande facilité avec laquelle les sujets associent des mots et sont capables de répondre à des questions du type « les canaris ont-ils des plumes ? »

2.1.2. Représentations définitoires

On peut regrouper sous cette appellation des modèles très divers, qui ont en commun de chercher à donner une définition formelle des unités, la diversité des modèles provenant de la diversité des formalismes utilisés – le terme « formalisme » étant d'ailleurs à prendre dans son sens le plus large.

À une extrémité du spectre on trouvera les formalismes logiques dans lesquels les définitions se présentent sous la forme d'un ensemble, nécessaire et suffisant, de conditions à satisfaire. Pour *canari*, par exemple, cela peut consister par exemple à ajouter à la propriété 'être un oiseau' (propriété générique) des caractéristiques qui différencient le canari des autres oiseaux, comme la couleur, etc. (propriétés spécifiques). Ce type de modèle a l'avantage de permettre des calculs rigoureux au niveau de la phrase et du texte. Mais cela s'avère trop rigide pour traiter la langue, car la plupart des propriétés qui caractérisent un terme ne sont pas vraiment des conditions nécessaires : tous les oiseaux ne volent pas et tous les corbeaux ne sont pas noirs. Cela a conduit à affaiblir les définitions en utilisant la notion de *valeurs par défaut*, qui résout le problème mais au prix d'une plus grande complexité des formalismes logiques et d'une moins grande capacité calculatoire, qui représentait le principal avantage de ces approches logiques.

Une approche différente a connu plus de succès : la *théorie du prototype*. Introduite par la psychologue Elenaor Rosch [ROS 75], elle consiste à remarquer que, au plan cognitif, l'appartenance d'un élément à une catégorie n'est pas traitée de manière binaire, en tout ou rien, mais plutôt de manière graduelle. Il existe, comme le confirment les études expérimentales, des « bons » exemplaires et des moins bons, pour la catégorie des oiseaux, des meubles, des couleurs, des formes, etc. Le canari et le moineau, par exemple, sont considérés comme des exemplaires plus prototypiques de la classe des oiseaux que l'autruche ou le pingouin. Cela conduit à définir les sens d'un mot comme un ensemble de traits, plus ou moins caractéristiques, dont aucun n'est strictement nécessaire : simplement, plus une entité possède de traits associés au mot, plus elle a vocation à être désignée par ce mot (voir [NYC 98] p. 303, [CAR 89] p. 99). La formalisation de cette approche utilise des outils de nature géométrique, plus aptes que la logique à traiter l'aspect continu de ces représentations. Le sens du mot est représenté par une région, au centre de laquelle se trouvent les meilleurs exemplaires (ceux qui possèdent le maximum de traits) : plus on s'éloigne de ce centre, qui représente le prototype, moins le mot est pertinent. Les frontières de cette

région sont floues, ce qui correspond aux hésitations des locuteurs (et aux désaccords entre locuteurs) sur la désignation d'exemplaires particulièrement atypiques.

La théorie du prototype a notamment été adoptée par les tenants des grammaires cognitives nord-américaines (cf. [VIC 04]), comme George Lakoff [LAK 87], qui a cherché à prendre en compte la polysémie dans ce cadre en créant ce que Georges Kleiber [KLE 90] a appelé la « version étendue de la sémantique du prototype ». Pour rendre compte du fait qu'une même unité lexicale peut évoquer des entités appartenant à des catégories conceptuelles nettement distinctes, même si elles entretiennent des relations de voisinage sémantique, on accepte qu'il puisse y avoir plusieurs prototypes différents associés à une même unité. En termes géométriques, on représente le sens d'un mot par une région connexe décomposable en un ensemble de sous-régions (cf. figure 1), dont chacune possède son propre prototype (A, B, C,… sur la figure) : on peut ainsi décrire la diversité des sens d'une unité polysémique, en représentant à la fois ce qui les relie et ce qui les différencie.

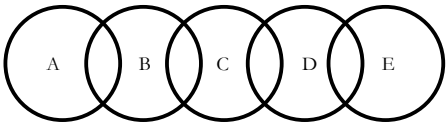


Figure 1. *Représentation de la version étendue de la sémantique du prototype*

D'autres modèles s'éloignent plus radicalement des approches logiques par lesquelles nous avons commencé cette section. Il s'agit d'approches assez diversifiées qui considèrent que le sens se construit de manière dynamique, à partir d'un *noyau de sens* associé à l'unité, qui interagit avec le contexte pour prendre un sens plein dans chacun de ses emplois. L'appellation et la nature précise de ce noyau de sens change suivant les auteurs et les théories, de même d'ailleurs que le degré de formalisation : *formes schématiques* chez Culioli [CUL 90], *archétypes cognitifs* chez Desclés [DES 85], *motifs* chez Cadiot et Visetti [CAD 01], *schémas* chez Langacker [LAN 87], etc. L'introduction de ce niveau abstrait permet de rendre compte de la polysémie : à partir d'un noyau de sens unique, un mot peut prendre des sens différents dans différents contextes.

2.1.3. Représentations componentielles

Ce troisième mode de représentation est directement issu de la linguistique structurale. Il s'agit de décomposer le sens d'une unité lexicale en un ensemble de traits différentiels, appelés *sèmes*. La méthode générale consiste à examiner les différences de sens entre unités d'un même champ sémantique pour en déduire les traits qui caractérisent chaque unité relativement aux autres. Ainsi, dans le champ sémantique des moyens de transport, l'opposition entre *autocar* et *autobus* conduit à introduire un sème 'interurbain' pour le premier et 'intra-urbain' pour le second.

Sous sa forme traditionnelle la plus classique, l'analyse sémique ne produit pas des résultats très différents des approches définitoires vues ci-dessus, même si les *a priori* théoriques et méthodologiques sont très éloignés les uns des autres. En effet, dans les deux cas, on aboutit à un ensemble de traits caractéristiques (on distingue aussi les sèmes génériques et les sèmes spécifiques), et on retrouve bien sûr les mêmes difficultés si l'on cherche à les traiter comme des conditions nécessaires…

Mais Rastier a donné une tout autre ampleur à cette approche, dans le cadre de sa théorie de la sémantique interprétative [RAS 87]. Notamment il introduit la notion de *sèmes afférents*, qui sont aussi des traits différentiels, mais qui définissent des oppositions dans d'autres champs sémantiques que celui auquel appartiennent de manière intrinsèque les unités étudiées. Ces sèmes afférents s'ajoutent donc aux sèmes de l'analyse classique, qu'il appelle *sèmes inhérents*. Ainsi, pour prendre un exemple, *chien* et *loup* s'opposent non seulement par le sème inhérent 'domestique' versus 'sauvage', mais aussi, notamment dans les fables, par le sème afférent 'asservi' versus 'libre'. Les sèmes afférents sont plus ou moins conventionnels et ils sont actualisés ou au contraire neutralisés dans le jeu interprétatif qui porte sur le texte dans sa globalité. Rastier replace donc l'analyse sémique (qu'il appelle la *microsémantique*) dans le cadre plus large de l'analyse textuelle (la *macrosémantique*), mettant ainsi en relation les différents

paliers de la sémantique, notamment avec la notion d'*isotopie*, sur laquelle nous aurons l'occasion de revenir quand nous aborderons le palier du texte.

2.2. Polysémie et calcul du sens lexical

Le problème de calcul du sens se pose dès le niveau des unités linguistiques parce qu'il n'y a pas de correspondance biunivoque entre mots et sens. La règle générale, au contraire, c'est un foisonnement de sens, d'autant plus grand que les mots sont plus fréquents. Cette caractéristique des langues les distingue radicalement des langages formels, et elle rend très problématiques, pour ne pas dire vaines, les tentatives de les traiter de la même manière. L'omniprésence de la polysémie pose un problème fondamental qui ne se pose pas pour les langages formels : comment calculer le sens d'un énoncé composé d'unités dont le sens dépend de cet énoncé ? Comme nous allons le voir, le terme de polysémie recouvre en fait des phénomènes extrêmement variés, et les méthodes de calcul du sens lexical sont donc aussi très diversifiées.

2.2.1. Les mécanismes de changement de sens

Deux grands procédés sont principalement à l'origine de l'acquisition de ces nouveaux sens au cours de l'évolution d'une langue : la *métaphore* et la *métonymie*. La métaphore (cf. chapitre 7) consiste à utiliser un mot qui désigne habituellement une entité ou un événement d'un certain domaine pour évoquer une entité ou un événement qui joue un rôle analogue dans un autre domaine. Par exemple, c'est par métaphore que l'on parle d'*intoxication* à propos de la diffusion de fausses nouvelles : on a transféré ce mot du domaine de la pathologie, où il est né, à celui de l'information, où il a pu prospérer allégrement… Quant à la métonymie, c'est le procédé qui consiste à évoquer une entité (ou un événement) par le mot qui désigne une autre entité (ou événement), liée à la première par un rapport fonctionnel ou structurel. Ainsi c'est par métonymie que *une voile* peut désigner un bateau à voile, ou que l'on dit *faire rire la salle* alors que ce sont les occupants de la salle qui rient.

On parle de *métaphore vive* ou de *métonymie vive* quand on a affaire à des inventions d'un locuteur, créations éphémères de la parole, qui ne sont comprises qu'en contexte. Ainsi dans l'énoncé *Les Dupont viennent-ils avec leur ouragan ce soir ?*, le mot *ouragan* peut désigner métaphoriquement, entre autres, un enfant turbulent aussi bien qu'une voiture particulièrement bruyante. De même une métonymie telle que *L'appendicite de cette nuit fait de la fièvre* ne peut se comprendre que dans le contexte d'un échange entre employés d'un hôpital (cf. [NUN 78] et [FAU 84]). Métaphores et métonymies sont dites *conventionnelles* ou *lexicalisées* quand elles se sont imposées dans la langue, comme par exemple *une bonne bouteille* qui qualifie par métonymie le contenu de la bouteille plutôt que le contenant lui-même. Enfin, on dit qu'elles sont mortes quand l'opération qui a conduit au nouveau sens n'est plus présente dans la conscience des locuteurs. Ainsi dans *la source de ces ennuis*, le mot *source* est perçu simplement comme un synonyme de *cause* et *origine*, sans que les locuteurs l'associent dans cet emploi au domaine de l'hydrologie qui est à l'origine du processus métaphorique.

2.2.2. Le traitement des polysémies systématiques

Certaines polysémies peuvent être qualifiées de *systématiques*, dans la mesure où toutes les unités d'une classe sémantique donnée sont soumises aux mêmes variations de sens. Il en est ainsi notamment pour ce que Kleiber [KLE 94] a appelé *les métonymies intégrées*. Par exemple, toute œuvre peut être désignée par le nom de son auteur : *Il a lu tout Victor Hugo. Ces Matisse sont magnifiques.* etc. De même les mots désignant un contenant peuvent être assez systématiquement utilisés pour désigner le contenu : on peut ainsi boire un verre, un godet, un flacon, une bouteille, une bonbonne, une barrique, etc. On a une propriété analogue entre le support matériel d'un écrit et le contenu informationnel véhiculé par cet écrit : on peut tout aussi bien dire qu'un livre est très petit et qu'il est très intéressant, et il en va de même pour un journal, un carnet, un cahier, etc. On peut donc formuler des règles générales sur le lexique qui permettent de rendre compte de ces phénomènes.

Il faut cependant rester prudent et ne pas généraliser trop vite : dans de nombreux cas, la métonymie ne touche qu'une partie des items lexicaux d'une classe donnée. Ainsi si *une plume* peut désigner un écrivain, *un crayon* ne s'emploie pas pour un dessinateur ni *un pinceau* pour un peintre. Autre exemple : si *de la poire* et *de la mirabelle* désignent les alcools obtenus à partir de ces fruits, *de l'orange* ne désigne pas l'alcool ou le jus d'orange, ni *de l'olive* l'huile d'olive. Ce sont ces exemples, notamment, qu'utilise Kleiber [KLE 99] pour mettre en garde contre l'utilisation abusive de ces mécanismes généraux, tels le « broyeur universel » de Nunberg et Zaenen [NUN 97], censé rendre compte de tous les effets de sens produits par l'emploi du déterminant partitif *du* sur les noms comptables.

Il faut donc à la fois formuler des règles générales et inscrire dans le lexique le comportement spécifique des différentes unités par rapport à ces règles. Le modèle de *lexique génératif* de Pustejovsky [PUS 95] combine ces deux aspects en proposant une description des unités lexicales qui permet de calculer les effets de sens en contexte (voir aussi la notion de *facette* introduite par Cruse dans [CRU 00] et [CRO 04]). Ainsi, pour un mot comme *livre*, Pustejovsky propose une structure de l'entrée lexicale qui intègre les deux types 'objet physique' et 'information', ainsi que d'autres informations, comme son usage (*lire* pour un livre) et son processus de création (*écrire*). L'ensemble de ces informations permet ensuite non seulement de retrouver le sens du mot dans un contexte donné par un mécanisme général de *coercition de type*, mais aussi de calculer correctement le sens d'autres unités utilisées avec le mot *livre* : ainsi *commencer un livre* pourra être interprété, par défaut, comme *commencer à lire un livre* ou *commencer à écrire un livre*.

2.2.3. Calcul de désambiguïsation lexicale

Le traitement des polysémies non systématiques, spécifiques à une unité lexicale donnée, constitue un vaste champ de recherche, plus connu sous le nom de désambiguïsation lexicale (en anglais WSD : *Word sense disambiguisation*), qui s'est fortement développé notamment sous l'impulsion des campagnes SENSEVAL qui ont permis d'évaluer et de confronter les approches de différentes équipes de recherche sur des tâches de désambiguïsation dans plusieurs langues (pour une vue d'ensemble de ces différentes approches, voir [AGI 01] et [IDE 98]).

Le principe général consiste à utiliser différents éléments du cotexte du mot à désambiguïser, notamment (mais pas uniquement) les mots lexicaux les plus proches, et de les comparer aux définitions des différents sens du mot (soit dans un dictionnaire électronique classique, ou, plus fréquemment, dans un réseau tel que *WordNet* cf. § 2.1). Deux grands types de méthodes se dégagent :

– d'une part des systèmes de règles de décision, souvent ordonnées par apprentissage (voir par exemple [YAR 00]), de la forme : « si tel élément est présent dans le cotexte, alors sélectionner tel sous-ensemble de sens ».

– d'autre part des méthodes géométriques où l'on représente les éléments du cotexte par des vecteurs dans un espace de très grande dimension muni d'une métrique fondée sur la fréquence de co-occurrence dans un corpus (cf par exemple [SCH 98]), suivant le principe de l'*analyse sémantique latente* (en anglais LSA : *Latent Semantic Analysis* cf. [DER 90]). Signalons aussi, parmi les modèles géométriques, le modèle de *construction dynamique du sens* [VIC 96], basé sur la construction d'un espace sémantique associé à chaque unité polysémique [PLO 98], et qui a conduit à la réalisation d'un système de désambiguïsation automatique [JAC 05].

2.3. Un modèle de la sémantique lexicale de l'espace et du mouvement

Après ce survol rapide des principales approches modélisatrices en sémantique lexicale, nous allons présenter un peu plus en détail un modèle particulier, qui va nous permettre d'illustrer de manière plus concrète la plupart des problèmes que nous avons évoqués. Il s'agit du modèle du lexique spatial développé par Yann Mathet (cf. [MAT 00] et [MAT 05]).

Ce modèle permet de représenter un grand nombre d'unités lexicales servant à exprimer des propriétés spatiales et spatio-temporelles, notamment les verbes de mouvement tels que *aller*, *traverser*, *parcourir*, *contourner*, *longer*, *s'approcher*, *entrer*, *quitter*, etc. Notons tout de suite qu'aucun de ces verbes n'est, de fait, limité à l'expression du mouvement : on peut *aller bien ou mal*, *traverser une épreuve*, *parcourir un*

livre, *contourner un problème*, etc. Le modèle laisse délibérément de côté tous les emplois non spatiaux, ce qui ne veut pas dire qu'il élimine toute polysémie. Notamment, comme on le verra, il rend compte de la capacité de ces verbes à prendre un sens dynamique et un sens statique : à côté de *la voiture contourne le rond-point* et *le cycliste longe le parc* (emplois dynamiques), on trouve *l'autoroute contourne la ville* et *les arbres longent la rivière* (emplois statiques).

2.3.1. Les entités

Mathet définit 5 types d'entités : les chemins, les trajectoires, les lieux, les rubans et les entités plurielles.

– un *chemin* est une courbe dans l'espace bi ou tridimensionnel : cela permet notamment de rendre compte de la notion de frontière d'une région surfacique, et plus largement, à un certain niveau de grain, d'entités telles que des routes ou des rivières. On peut définir une *orientation* sur un chemin.

– une *trajectoire* rend compte du déplacement d'un mobile : c'est une fonction qui donne une position dans l'espace au cours du temps On lui associe son support (ensemble des positions parcourues) qui est donc un chemin (orienté par le sens de déplacement du mobile). Un objet immobile est caractérisé par une seule position dans le temps (la trajectoire est une fonction constante).

– un *lieu* est une région de l'espace bi ou tridimensionnel pouvant représenter aussi bien la surface d'une prairie, d'un pays, etc. que l'intérieur d'une maison ou d'une voiture. Certains lieux peuvent être en mouvement, comme l'intérieur d'une voiture, ou subir des modifications : ville en expansion, par exemple. Pour en rendre compte, un lieu est donc aussi défini comme une fonction temporelle, même si cette fonction est le plus souvent constante.

– un *ruban* est un lieu surfacique défini par la donnée d'un chemin et d'une *largeur*, fonction de l'abscisse curviligne sur le chemin. Cela permet de représenter des routes et des rivières de manière plus précise que par un simple chemin.

– enfin, une *entité plurielle* est un ensemble d'entités de même nature.

Une notion importante dans le modèle est celle de *polymorphisme* des entités : une entité peut être conçue de différentes manières et donc relever de plusieurs types. On vient ainsi de voir qu'une route peut être considérée comme un ruban et un chemin, et une voiture comme une trajectoire et un lieu. L'utilisation du polymorphisme offre donc une grande souplesse dans le typage : une grande partie des polysémies systématiques (cf. § 2.2) est prise en compte par ce procédé. De plus, cela permet de simplifier la définition de nombreux verbes. Pour ne donner qu'un exemple, le verbe *longer* sera représenté exactement de la même manière dans *les arbres longent la rivière* et dans *la route longe la rivière*, parce qu'une rangée d'arbres peut être traitée à la fois comme une entité plurielle et comme un chemin.

En revanche, *longer* ne conduit pas à la même représentation dans *la voiture longe la rivière*. Il y a bien prise en compte de la polysémie de ce verbe qui n'a pas, de fait, le même sens dans cet emploi, où il évoque un mouvement, que dans les emplois précédents, où il évoque une relation spatiale statique. Mais cette polysémie est traitée très élégamment, grâce à la correspondance entre trajectoires et chemins : dans l'emploi dynamique, la trajectoire du mobile doit avoir pour support un chemin qui vérifie la relation spatiale évoquée par les emplois statiques du verbe. L'existence d'un « noyau de sens » commun aux deux types d'emplois (cf. § 2.1) est donc bien mise en lumière par cette représentation.

2.3.2. Les relations

Mathet définit cinq classes de relations géométriques pour rendre compte de l'ensemble des expressions spatiales : des relations *topologiques* (intérieur, extérieur, frontière de lieux, intersections entre lieux, chemins, etc., mais aussi extrémités de chemins et de trajectoires), des relations de *distance (*approcher, éloigner, faire un crochet, etc.), des relations d'*orientation* (droite, gauche, haut, bas, etc.), des relations de forme (notamment la notion d'enveloppe convexe, sur laquelle nous allons revenir), et enfin des relations d'*ordre curviligne* (devancer, suivre, doubler, etc.). Il n'est pas possible de détailler ici l'ensemble du dispositif, mais les deux exemples illustrés figures 2 et 3 permettront de se faire une idée de la puissance d'expression obtenue : on peut définir de manière très rigoureuse des « formes schématiques » représentant le sens des verbes (limités toutefois à leurs emplois purement spatiaux).

Ces exemples utilisent la notion d'enveloppe convexe d'une entité, qui est définie comme le plus petit convexe[2] contenant l'entité. La figure 2 représente graphiquement le sens de *passer entre*. Cela revient à exprimer, dans le formalisme adéquat développé par Mathet, que la trajectoire T passe entre les lieux L_1 et L_2 si et seulement si T traverse l'enveloppe convexe de l'union ensembliste $L_1 \cup L_2$ sans pour autant entrer ni dans L_1 ni dans L_2.

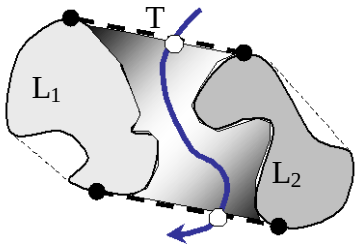


Figure 2. Représentation de *T passe entre L_1 et L_2*

La figure 3 illustre la définition de *contourner* : C contourne L si et seulement si C reste à l'extérieur de L, alors que l'enveloppe convexe de C coupe L.

2.3.3. Implémentation et validation

A des fins de validation, Mathet a conçu et réalisé un logiciel qui implémente les principaux ingrédients du modèle et qui comporte une interface permettant de construire et de visualiser sur écran des lieux surfaciques de toute forme, des chemins, des trajectoires, etc., ce qui fait que l'on peut représenter graphiquement les situations que l'on veut tester, et demander au logiciel si telle ou telle expression linguistique convient ou ne convient pas à la situation présentée.

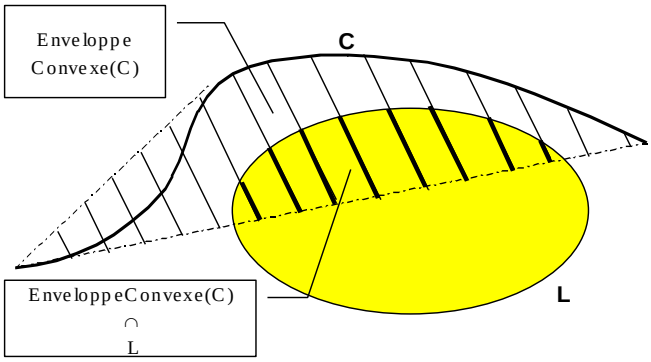


Figure 3. Représentation de *C contourne L*

L'une des grandes qualités du logiciel, qui mérite d'être soulignée, est qu'il ne répond pas de manière binaire (oui ou non), mais qu'il fournit une appréciation graduée de l'acceptabilité de l'expression pour la situation. Il y a donc prise en compte du caractère continu du sens, sur lequel nous avons insisté ci-dessus (§2.2). Si, par exemple, on examine la définition de *contourner* que nous venons de donner, on peut remarquer que l'une des conditions se prête effectivement à une telle évaluation : on peut utiliser le verbe *contourner* dès que l'intersection du lieu et de l'enveloppe convexe du chemin n'est pas vide, mais le verbe sera d'autant plus approprié que cette intersection sera grande, plus précisément qu'elle représentera une plus grande proportion de la surface du lieu. Mathet a donc défini pour ce verbe, ainsi

2. Une forme est dite convexe quand elle contient tous les segments dont les deux extrémités appartiennent à la forme. Par exemple, un camembert entier est convexe, mais il perd cette propriété dès qu'on l'entame (il redevient convexe quand on en a enlevé plus de la moitié !).

que pour d'autres présentant des caractéristiques quantifiables du même genre (*longer*, *traverser*, *parcourir*, *envahir*, etc.) un *degré d'adéquation* du verbe avec une situation, et c'est ce degré qui est confronté au jugement des locuteurs, qui est dans ce cas lui aussi gradué.

Ainsi le modèle de Mathet a l'avantage de pouvoir être évalué de manière précise et systématique sur une grande variété de situations et d'expressions linguistiques. Une expérimentation de grande ampleur est d'ailleurs en cours, qui utilise un corpus de descriptions textuelles de déplacement collectées sur le Web.

3. Le palier de la phrase

3.1. Intégration sémantique et énonciation

Deux faits caractérisent la phrase[3] du point de vue sémantique :

– À partir de l'ensemble des « ingrédients » que sont les mots d'une phrase et le « contenu » qui leur est associé, est constitué un « tout » cohérent, structuré, possédant une sorte d'unité de sens, relativement autonome du reste du texte. Il faut aussi noter que ce calcul implique des connaissances générales sur le domaine de l'énoncé. C'est ce que nous appellerons ici *l'intégration sémantique* (expression librement empruntée à J. Caron [CAR 89], p. 158).

– La phrase fait l'objet d'une *énonciation*, par un locuteur ou scripteur. Cette énonciation n'est pas « désincarnée », mais répond à une certaine *visée*, dans un certain *contexte*. Elle réfère à une certaine « réalité extra-linguistique » (conçue, non « objective » ou « ontologique ») avec laquelle l'énonciateur entretient une certaine relation de « croyance » ou plus généralement « d'adhésion », et se situe dans une intention et un rapport communicationnels.

C'est donc tout un ensemble d'*opérations de construction du sens*, de nature syntaxico-sémantique, qui interviennent au niveau de la phrase et qui contribuent à cette intégration sémantique, que nous allons détailler quelque peu. Prenons l'exemple suivant :

– *Je venais juste d'être servi à la terrasse d'un café quand un individu surgi de nulle part s'est effrontément emparé de mon verre.*

Nous distinguerons trois grands types d'opérations :

– Opérations d'instanciation : un certain nombre d'entités sont introduites par le texte, répondant à un certain « typage ». Sur notre exemple : des entités de type 'humain'(le locuteur et l'autre individu), 'lieu'(la terrasse du café), 'objet physique'(le verre).

– Construction des relations actancielles : d*es liens* sont établis entre ces entités, pour former une *structure* sémantique porteuse d'une *information complexe*. Par exemple, l'entité désignée par *un individu* est acteur des actions *surgir* et *s'emparer*, et *mon verre* désigne l'objet qui subit l'action *s'emparer.*

– Repérage spatial, aspectuo-temporel, modal et énonciatif : la situation évoquée par le texte est toujours située dans le temps, souvent dans l'espace, et fait l'objet d'une prise de position du locuteur/scripteur. Dans l'exemple : l'histoire est située dans le passé, avec une relation de consécution immédiate entre l'action de *servir* et celle de *servir*. Le repérage spatial est aussi présent. Et l'adverbe *effrontément* exprime un jugement du locuteur.

En TAL, ces opérations seront réalisées en intégrant *modèle linguistique* et modèles de *représentation et de traitement des connaissances* (cf. [KAY 97]), utilisant le plus souvent les *Graphes Conceptuels* [SOW 84] [MUG 96], ou la *Logique* [DEL 01]. Des méthodes générales, permettant de décrire l'enchaînement de ces traitements sémantiques ont été élaborées, que nous évoquerons au § 3.3.

3. En fait ces deux caractéristiques commencent à se manifester au niveau du syntagme et se poursuivent au niveau textuel.

3.2. Les opérations de construction du sens de l'énoncé

3.2.1. Instanciation

Nous aurons d'abord ce qui peut être caractérisé comme des opérations *d'instanciation*, bien connues des informaticiens et des logiciens. Alors que les éléments lexicaux évoquent des catégories (ou des types, en langage plus informatique), les syntagmes complets qui constituent les phrases évoquent généralement des exemplaires de ces catégories (des instances). Ainsi les noms *homme* et *orateur* désignent des catégories de personnes et le verbe *interpeller* une catégorie d'actions, alors que dans la phrase :

– *Un jeune homme dégingandé a violemment interpellé l'orateur.*

les syntagmes *un jeune homme dégingandé* et *l'orateur* désignent des personnes précises, et le syntagme *a violemment interpellé* un événement particulier[4]. Ces opérations d'instanciation sont généralement le résultat de la combinaison d'unités linguistiques de deux sortes :

– *un ensemble d'unités lexicales* (*jeune* + *homme* + *dégingandé*, *interpeller* + *violemment*) qui contribuent à préciser et à qualifier la classe sémantique dont un exemplaire va être extrait.

– *une unité grammaticale* (notamment déterminant ou marque de temps verbal) qui est le support de l'opération proprement dite.

Ainsi, *jeune homme dégingandé* évoque une catégorie tout comme *homme*, et c'est l'ajout de *un* qui crée une instance singulière de ce type. De même c'est la marque du passé composé qui fait passer de l'abstraction *interpeller violemment* à l'événement concret évoqué par le texte : *a violemment interpellé*.

À vrai dire, les choses ne sont pas toujours aussi simples. Ainsi, les déterminants peuvent prendre des valeurs dites génériques, comme dans *L'homme est un roseau pensant* (cf. chapitre 6). D'autre part la question se pose de savoir par quelle entité particulière effectuer l'instanciation. S'agit-il d'une entité nouvelle introduite par le texte ? Ce sera en général la fonction des descriptions indéfinies (*un X*, *des X*…). Ou bien d'une nouvelle évocation, d'une reprise (par un pronom ou une description définie *le X*, *ce X*… : on parlera d'anaphore pronominale et nominale respectivement) d'une entité déjà présente dans le « contexte du discours » ? Ainsi, dans l'exemple suivant :

– *Un jeune homme dégingandé a violemment interpellé l'orateur. Celui-ci n'a pas su quoi répondre à son contradicteur.*

– La description indéfinie *Un jeune homme dégingandé* introduit un nouveau personnage. *celui-ci* est une anaphore pronominale (renvoyant à l'orateur), et *son contradicteur* une anaphore nominale (renvoyant au jeune homme dégingandé). Cette question est celle du calcul de la *coréférence* : sa portée se situe clairement au niveau du texte et nous y reviendrons en section 3.4.

3.2.2. Relations argumentales et compositionnalité

C'est une des fonctions sémantiques essentielles de la syntaxe de la phrase que de distribuer les rôles des différents acteurs de la « saynète » qui est décrite par la phrase, pour reprendre une métaphore qui remonte au moins à Tesnière. Ce sont d'abord et avant tout les constructions verbales qui sont le support de ces opérations, en permettant de répondre à des questions du type : qui fait quoi, à qui, avec quoi, etc. La détermination de ce que l'on appelle les *relations actancielles* constitue donc une part importante de la sémantique de la phrase. Elle repose sur toute une série de marques grammaticales et syntaxiques : prépositions, ordre des mots, diathèse (voix active, passive, réflexive,…), mais aussi sur les éléments lexicaux. Il n'y a en effet pas de correspondance biunivoque entre une fonction grammaticale (sujet, objet,…) et une relation actancielle donnée, même à voix constante. Par exemple, dans les trois phrases suivantes :

– *Je n'avais pas fermé la porte de ma voiture*

– *Ce dispositif électronique ferme toutes les portes de la voiture*

---

4. Cette opération d'instanciation pose donc, du point de vue du TAL, des problèmes spécifiques de représentation et de calcul qui ne se posaient pas vraiment au niveau précédent, celui du mot. C'est notamment elle qui a conduit des premiers réseaux sémantiques au formalisme des graphes conceptuels.

– *La porte avant droite fermait mal*

on trouve en position de sujet trois rôles actanciels différents : agent, instrument et siège du procès. Il s'agit pourtant dans les trois cas du même verbe à la voix active : c'est, de fait, l'analyse sémantique du groupe nominal sujet et la présence ou non d'un groupe objet qui permettent de déterminer le rôle actanciel en question.

Des relations argumentales du même genre peuvent aussi porter sur d'autres unités que les verbes : des noms (*Le transport des voitures par camion, la bicyclette de Pierre*, etc.), des adjectifs (*rouge de colère, un bidon plein d'essence*, etc.). Là encore, la correspondance entre construction syntaxique et rôle sémantique est loin d'être simple. Ainsi dans l'agression de Pierre, Pierre peut être aussi bien l'agresseur que l'agressé : seul le contexte, et parfois un contexte très éloigné, peut permettre de lever l'ambiguïté.

Ces relations argumentales jouent un rôle clé dans la construction du sens de la phrase. Sous sa forme la plus simple, ce « sens » consiste en une « configuration » dont le centre est une instance d'événement, auquel sont reliées des entités par des relations actancielles de type agent, patient, instrument, etc. Par exemple pour la proposition *Un camion a percuté une barrière métallique*, nous aurons une structure comprenant :

– un *événement* E instance de la classe 'percuter';
– une *entité* C instance de 'camion', dans la relation 'agent' avec E ;
– une *entité* B, instance de 'barrière', dans la relation 'patient' avec E ;
– l'attribution de *propriétés* 'être métallique' à V.

Le calcul opère essentiellement par *compositionnalité* : à partir des relations actancielles et syntaxiques d'une part, des représentations associées aux mots de l'autre, une représentation est calculée pour des syntagmes de plus en plus complexes : à partir de 'voiture', 'blanche' et 'petite', on aura 'petite voiture blanche', concept type, instancié dans 'la petite voiture blanche' ; puis, à partir de 'un camion' (instance de 'camion') et 'a percuté' (instance de 'percuter'), les relations actancielles conduisent à la structure ci-dessus. Nous verrons au § 3.2 comment formaliser ce calcul.

S'il s'agit bien du mécanisme « central » au niveau de la phrase, les limites de la compositionnalité doivent cependant — ici comme à propos du lexique — être relevées (voir [NAZ 98] pour différentes études sur ce thème, sous l'angle du TAL). Par exemple dans une variante de l'exemple précédent : *J'ai percuté une barrière métallique*, nous aurions l'introduction d'une entité implicite 'voiture', provenant des connaissances générales sur « le monde de la route ». On trouve dans un constat d'accident la phrase *J'ai glissé à cause d'un brutal verglas* : c'est clairement la perception du rédacteur, qui est brutale, pas le verglas. Une méthode pour résoudre ce type de problème a été bien systématisée par Pustejowski [PUS 95] : on décèle un conflit de type entre deux unités reliées entre elles (ici : brutal et verglas), et on le résout en cherchant dans la « situation » évoquée un « attribut » susceptible de recevoir la qualification problématique.

3.2.3. Repérage, cadrage et opérations énonciatives

Il s'agit d'un ensemble d'opérations qui ont en commun de commencer au niveau de la phrase, mais d'être généralement d'une portée plus grande que la phrase, et que l'on retrouvera donc en section 4 (le niveau du texte). Les unes servent à situer la « saynète » évoquée par la phrase, que ce soit dans l'espace et dans le temps, s'il s'agit d'une scène décrivant un événement, ou dans un espace notionnel particulier, dans le cas d'une proposition plus abstraite. Les autres servent à préciser le point de vue de l'énonciateur par rapport à son dire : comment présente-t-il la scène en question, quels éléments met-il en relief, dans quelle mesure prend-il en charge sa véracité, etc. Chez Culioli ([CUL 90], p. 127-130) les deux types sont appelés des opérations de *repérage*, avec un repère spatio-temporel pour les premières, et un repère intersubjectif centré sur l'énonciateur pour les secondes. Toutes ces opérations sont portées principalement par des tournures syntaxiques et des morphèmes grammaticaux : marqueurs aspectuo-temporels, marqueurs modaux, marqueurs de thématisation et de focalisation, etc.

En ce qui concerne le repérage temporel, les marques de temps verbal jouent un rôle essentiel, mais se combinent avec d'autres indices (constructions syntaxiques, et type de procès notamment, comme nous le

verrons en section 4. Par contre, pour le repérage spatial, il n'y a pas de marque grammaticale obligatoire dans chaque phrase[5]. On peut faire les mêmes remarques pour les opérations de repérage intersubjectif, qu'il s'agisse des modalités ou des relations de thématisation et de focalisation. Ces opérations se caractérisent par le fait que des marques de toute nature peuvent contribuer à les spécifier. Par exemple, une modalité épistémique comme la plausibilité peut s'exprimer de manières très diverses :
– *la voiture a sans doute dérapé* : adverbe dit « de phrase ».
– *il est probable que la voiture ait dérapé* : construction propositionnelle.
– *la voiture a dû déraper* : verbe modal.

De même pour la thématisation :
– *quant à Jean, il n'a heureusement pas été blessé* : syntagme prépositionnel
– *en ce qui concerne Jean, il n'a heureusement pas été blessé* : subordonnée
– *Jean, lui, n'a heureusement pas été blessé* : pronom en apposition

C'est d'ailleurs souvent la combinaison de plusieurs marques, réparties un peu partout dans la phrase, qui produit la propriété globale en question. Ainsi, c'est la combinaison de l'imparfait, d'un verbe exprimant un événement « ponctuel », et un complément de temps de type prospectif qui fait que la phrase *Une minute plus tard, le train déraillait* peut exprimer le mode 'irréel' : Nous verrons tout cela en détail en ce qui concerne l'un de ces domaines : l'aspectuo-temporalité dont nous présenterons un modèle « opératoire » au § 5.

3.3. Calcul du sens de la phrase

Le calcul du sens *de la phrase* a longtemps été le point de focalisation, presque exclusif, des tentatives de formalisation et de calcul informatique. Trois courants majeurs inspirent en général les réalisations : *la sémantique formelle* [CHA 98], d'origine montagovienne [MON 74] qui fixe souvent le cadre théorique du calcul[6] ; *l'Intelligence Artificielle* qui offre diverses méthodes pratiques de représentation est de calcul du sens [KAY 97] ; et les modèles *d'analyse syntaxique*, qui permettent dans une certaine mesure un calcul préalable de la structure syntaxique de l'énoncé » (cf. chapitre 2 du présent volume).

Pour clore cette section, nous allons donner un aperçu du calcul du sens d'une phrase en logique, illustrant notamment la démarche compositionnelle et divers principes précédemment évoqués. Dans un premier temps, nous spécifierons le mode de représentation du lexique, puis le calcul lui-même. Nous évoquerons enfin brièvement la manière dont des représentations « phrase par phrase » peuvent ensuite être intégrées dans une représentation du sens d'un texte. Nous nous appuyons sur l'exemple de la phrase : *J'ai heurté une barrière métallique.*

3.3.1. Représentation du lexique

Noms et adjectifs

À chaque nom N on associe un prédicat `N (x)` caractérisant toutes les entités x « tombant » sous la dénomination N. De même à un adjectif A est associé `A (x)` signifiant que x possède la qualification A[7]. Ainsi dans le monde de la route, nous aurons notamment les prédicats :

– `voiture (x), route (x), barrière (x), métallique (x),` etc.
– les propriétés et relations définitoires, constitutives de la sémantique dans le paradigme logique, seront exprimées comme connaissances du domaine.

---

5. Nous limitons ici nos observations au français. La situation n'est pas la même dans toutes les langues : s'il semble qu'aucune langue n'impose un repérage spatial, beaucoup de langues, comme le chinois, n'ont pas non plus de marques verbales temporelles. La seule contrainte universelle porte sur l'aspect [COM 76], qui semble devoir toujours accompagner l'opération d'instanciation d'un événement.

6. Dont le représentant le plus « populaire » est sans doute la *Théorie de la Représentation du Discours* (*Discourse Representation Theory* ou *DRT*) de Hans Kamp [KAM 93] cf chapitre 6.

7. Nous ne considérerons pas ici des noms et adjectifs « relationnels » comme *enfant (de)* ou *voisin (de)*. Nous laissons cette variante à l'imagination du lecteur. D'autre part, nous simplifions ici la notation : pour des raisons techniques, on utilisera souvent des λ-notations : `λx voiture(x), λx route(x)…` (cf [CHA 05] et le chapitre 2).

Nous avons également une constante `loc` de type 'humain'pour désigner le locuteur. Constante à interpréter concrètement dans la seconde phase, selon le contexte d'énonciation :.

Verbes

Considérons le cas de *heurter*. Une première idée serait de le décrire comme exprimant une relation entre deux entités, dont l'une est mobile, représentée par un prédicat binaire `heurter (x, y)`. On pourrait de la même manière définir un prédicat `donner (x, y, z)` pour « x donne y à z », etc. Cette solution reçoit plusieurs critiques. En premier lieu, on voudrait pouvoir indiquer le « rôle actanciel » des différents arguments. Par exemple dans `donner (x, y, z)`, x est l'acteur, y l'objet ou patient, z le destinataire. La seconde critique est que l'on voudrait pouvoir référer à un événement, le désigner nominalement. Par exemple, on trouve dans un constat :*... le véhicule B m'a heurté... Le choc m'a projeté...* dans lequel *Le choc* reprend clairement l'événement exprimé par la première proposition. On est alors amené à associer à un verbe un prédicat unaire portant sur *l'événement*[8] exprimé par le verbe. Par ailleurs, on aura des prédicats caractérisant les différents rôles :

– `acteur (E, x)` : « x est l'acteur de E » ;

– `patient (E, x)` : « x est le patient de E » ;

– `destinataire (E, x)` : « x est le destinataire de E », etc.

Concernant notre exemple, nous associerons à *heurter* un prédicat `heurter (E)` (avec un typage de `e` comme événement) les relations actancielles devant être introduite à partir de la construction syntaxique de la phrase.

Mais les verbes se présentent sous forme conjuguée (*ai heurté*, au passé composé dans notre exemple). Autrement dit nous devons traiter le problème du temps morphologique. Nous opérerons pour notre exemple une approximation assez grossière qui ne retient que le temps (passé, présent, futur), et associerons ainsi à *heurtais* la formule :

– `heurter (E)` ∧ `passé (E)` [9].

Une alternative serait de noter le temps verbal comme opérateur sur l'événement associé au verbe. On écrirait alors : `heurter (E) ∧ Passé_Composé (E)`, renvoyant à une phase ultérieure les traitements aspectuo-temporels (cf. § 5).

Déterminants

Nous sommes confrontés ici au problème de la référence. C'est une question qui ne peut être traitée qu'au niveau du texte. Nous n'avons pas en général les moyens de calculer, au cours de l'analyse de la phrase, la référence de « un X », « le X », « ce X », etc. Nous allons donc retarder ce calcul. Considérons ici simplement les articles indéfinis *un*, *une*, et définis *le*, *la*. Nous leur associerons également des opérateurs, respectivement `UN` et `LE`, de telle sorte que :

– « un/une N » sera codé comme `UN x. N (x)`

– « le/la N » sera codé comme `LE x. N (x)`

La sémantique de `UN` sera, en première approximation, d'être un introducteur d'entité satisfaisant la condition de son argument. Et de même `LE` sera une instruction de repérage d'une telle entité (cf. § 4). Ces opérateurs peuvent être vus comme des sortes de « quantificateurs » — et la notation utilisée repose clairement sur cette analogie.

3.3.2. Exemple de calcul compositionnel du sens d'une phrase

Voyons maintenant comment peut fonctionner la compositionnalité au niveau de la phrase exemple. L'idée est de supposer que l'on dispose d'une analyse syntaxique de la phrase, produisant un arbre comme illustré sur notre exemple par la figure 4.

---

8. Ou, plus généralement, pour reprendre la terminologie du § 5: *le procès*.

9. Rappelons que le symbole ∧ désigne le « et » logique.

Aux feuilles de cet arbre sont associées des représentations lexicales selon les principes qui viennent d'être discutés. Le but est alors en quelque sorte de propager ces valeurs logico-sémantiques jusqu'à la racine de l'arbre. Observons les règles appliquées à chaque nœud, en partant du bas.

– *pour un nom éventuellement qualifié par un ou plusieurs adjectifs (nœud n-adj)* : on opère la conjonction des prédicats associés au nom et aux adjectifs qui le qualifient. Ici donc : `barrière (x)∧métallique (x)`

– *pour un groupe nominal (nœud gn)* : on applique le « quantificateur » (ici UN) associé au déterminant au prédicat associé au nom ou (comme ici) au groupe adjectival régi.

– *pour un groupe verbal (nœud gv)* : le fils *gn* ayant la fonction syntaxique de COD, on attribue à sa représentation sémantique le rôle de `patient`.

– *finalement, pour la phrase (noeud ph)*, nous interprétons la relation syntaxique SUJET par le rôle sémantique `agent`[10]. De plus, nous devons tenir compte du fait que la phrase fait l'objet d'une *énonciation*, donc d'une *instanciation* de l'événement asserté. Nous traduisons naturellement cette opération par l'introduction d'une constante E1 substituée à la variable e.

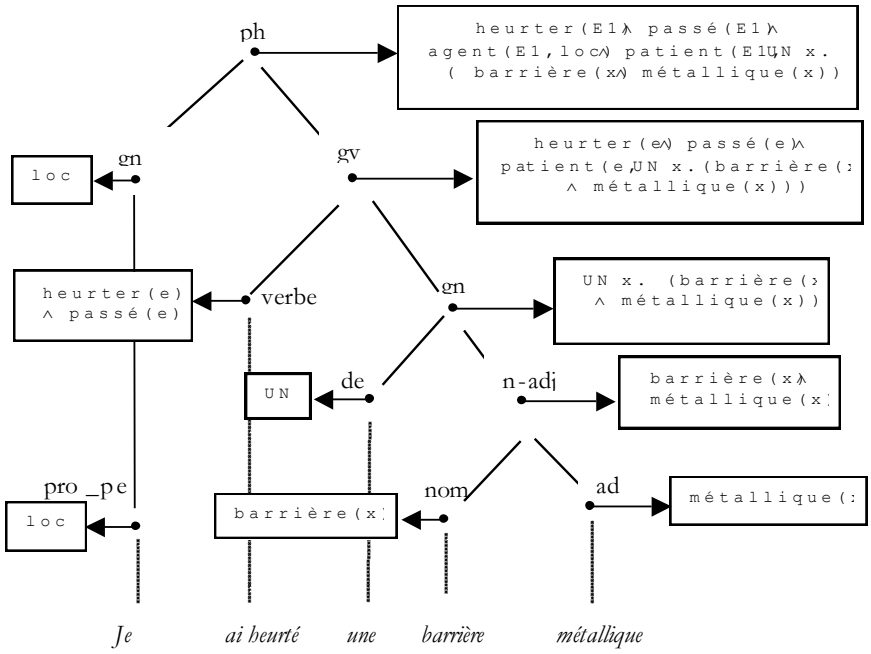


Figure 4 : Sémantique compositionnelle de *j'ai heurté une barrière métallique*

Au final, nous avons la formule logique :

```
heurter (E1) ∧ passé (E1) ∧ agent (E1, loc) ∧
patient (E1, UN x. (barrière (x) ∧ métallique (x))
```

qui paraît effectivement une représentation logique relativement fidèle de l'énoncé.

3.3.3. Discussion

Comme mentionné plus haut, cette formule demeure incomplète : il faut interpréter encore les pseudo-quantificateurs associés aux déterminants (tels que UN) et donner un sens à l'expression `passé (E1)` (ou, dans l'alternative évoquée `passé_composé (E1)`). Pour cette raison, ce type de formules est souvent qualifié de *forme quasi logique*. Seule finalement la représentation obtenue au niveau du texte est véritablement complète. Néanmoins, on accordera que le travail réalisé au niveau de la phrase est effectivement calculatoire et peut constituer une étape cruciale de l'interprétation du texte.

Une autre remarque concerne l'hypothèse de l'existence d'une analyse syntaxique préalable totalement autonome. Cette hypothèse peut se discuter d'un point de vue linguistique ou linguistico-cognitif général.

10. Attendu qu'il s'agit d'une phrase à la voie active. Le lecteur est invité à adapter les règles proposées ici pour la voie passive.

Du point de vue du TAL, on sait qu'elle se heurte aux limites des analyseurs syntaxiques actuels [ABE 00]. De fait, ces analyseurs ont fortement tendance à intégrer des éléments sémantiques, tels qu'un typage d'arguments verbaux en « grandes catégories sémantiques » par exemple. Cette position permet de sauver l'essentiel de l'architecture en 2 temps : syntaxe (légèrement sémantisée) — sémantique compositionnelle. On se heurte également aux limites de la compositionnalité déjà évoquées à plusieurs reprises, qu'il convient en pratique de prendre en compte.

D'autres importantes questions techniques ont été éludées. Pour plus de détails, on pourra consulter [CHA 05] ou, de manière plus extensive, divers ouvrages consacrés à la compréhension automatiques tels que [ALL 95] [DEL 01]. Nous avons néanmoins tenté de convaincre le lecteur de la faisabilité de ce type de calcul — dans les limites et avec les approximations d'usage en TAL.

4. Le palier du texte

4.1 Les différents éléments de cohésion du texte

Pas plus que les mots n'existent de manière isolée, mais (presque) toujours au sein d'unités structurées que sont les syntagmes — et singulièrement de cette unité primordiale que constitue la phrase — les phrases n'apparaissent seules et solitaires, mais au sein d'unités plus vastes que sont les textes[11]. Et les textes, loin d'être de simples successions de phrases, possèdent leurs propres formes d'organisation. C'est, avant même tout discours théorique, une évidence empirique : un « bon » texte possède une « structure » interne qui permet au lecteur d'assembler un ensemble d'idées ou d'impressions en un tout cohérent. Cette structure est multiforme et, à la vérité, particulièrement riche[12]. On parlera souvent de *cohésion textuelle*[13] et de ses divers *facteurs*. Certains apparaissent déjà au niveau syntagmatique, mais souvent se déploient véritablement au sein du texte. Il faut voir également qu'au-delà d'une simple « impression de cohérence » c'est la question proprement sémantique de l'organisation de « l'information » portée par le texte, et de son appréhension par le lecteur, qui est ici en jeu. Et il faut se souvenir que le texte est *l'unité fonctionnelle* par excellence : le type d'organisation dont il relève dépendra donc fortement du genre (narratif, présentatif, poétique, avec toutes sortes de ramifications) et contribuera de manière décisive à la fonction spécifique de ce genre. Nous envisageons maintenant quelques-unes de ces structures.

4.1.1. Coréférence

C'est un des facteurs de structuration les plus évidents. Nous avons vu que les expressions linguistiques — en particulier mais non exclusivement nominales — évoquent, réfèrent à des « objets »[14] (entités ou événements, concrets ou abstraits…) extralinguistiques, relevant d'un certain domaine de discours. On dira alors que deux expressions *coréfèrent* si elles renvoient au même « objet ». La figure 5 illustre ce mécanisme. On voit en particulier qu'une même entité peut être reprise plusieurs fois, sous diverses dénominations : se forme ainsi une *chaîne de coréférence*. L'usage est de marquer par un même indice (i, j, k…) les éléments d'une chaîne. On notera les deux modes principaux de reprise : par un pronom ou un groupe nominal. On parlera respectivement *d'anaphore pronominale* ou *nominale*. Du point de vue textuel, la présence de telle chaîne assure une forme de suivi du texte, à travers « les objets

11. Ou « discours », donc dans une acception qui intègre l'oral. Avec une nuance : on désigne plutôt par le premier les productions et par le second la manière dont elles sont produites.

12. On emploie aussi le terme de « texture » pour signifier cet entrelacement de relations très variées et relativement souples, moins « logiques » que la structure de la phrase.

13. On utilise également le terme de *cohérence textuelle*, la *cohésion* renvoyant plus aux diverses marques linguistiques qui, jointes à divers processus inférentiels, conduisent à une *représentation cohérente* du texte dans l'esprit du lecteur. Pour approfondir ces questions, voir notamment les travaux de Michel Charolles [CHA 95].

14. Objets d'un monde conçu, et non directement du monde « sensible », peuplé de Père Noël, de beauté convulsive et de tiers-mondisme autant que de tables et de chaises.

dont on parle », chaque nouvelle mention d'une entité apportant de nouvelles informations : cumulatives en quelque sorte, dans le cas d'un texte informatif (comme une dépêche d'agence), ou temporellement organisées dans une narration. À ce titre c'est un élément essentiel dans le processus de compréhension.

Le matin du 16 Avril, le docteur Bernard Rieux $_i$ sortit de son $_i$ cabinet et buta sur un rat mort $_j$, au milieu du palier. Sur le moment, il $_i$ écarta la bête $_j$ sans y prendre garde et descendit l'escalier. Mais, arrivé dans la rue, la pensée lui $_i$ vint que ce rat $_j$ n'était pas à sa place et il $_i$ retourna sur ses $_i$ pas pour avertir le concierge $_k$. Devant la réaction du vieux M. Michel $_k$, il $_i$ sentit mieux ce que sa $_i$ découverte avait d'insolite. La présence de ce rat $_j$ lui $_i$ avait paru seulement bizarre tandis que, pour le concierge $_k$ elle constituait un scandale. La position de ce dernier $_k$ était d'ailleurs catégorique : il n'y avait pas de rats dans la maison.

Premier paragraphe de *La Peste :* chaînes sur B. Rieux (i), le rat (j) et le concierge (k).

Figure 5. *Chaînes de coréférence*

4.1.2. Temporalité et narration

Un autre type de structure aisément repérable, particulièrement dans les textes narratifs (comme celui de la figure 5), est constitué des rapports de succession/simultanéité entre les événements relatés. Ce type de relations commence au niveau de la phrase, mais se poursuit à l'évidence au niveau du texte dans son entièreté. Pour les représenter, nous pouvons associer à chaque *procès* (expression d'un événement ou d'un état stable) une *période* (ou *intervalle*) pendant lequel « il se produit ». Le problème est alors de calculer, à partir de divers indices linguistiques (temps verbaux, structure actancielle, nature des propositions…) et, éventuellement, de connaissances sur « le monde », un ensemble de relations entre ces intervalles, spécifiées par le texte., Une théorie, permettant effectivement de calculer des chronogrammes à partir du texte même, sera présentée au § 5.

Les relations temporelles se superposent souvent avec d'autres relations entre procès, par exemple de type rhétorique (cf *infra*) : causalité, explication, élaboration… Par exemple *A 8h dans Marie eut faim. Elle mangea une pomme.* nous avons à la fois succession temporelle et causalité. Dans la narration, la structure temporelle constitue une sorte d'ossature « objective » d'une structure plus abstraite dans laquelle interviennent des schémas culturels « de haut niveau », fortement « codés ». Un exemple classique et emblématique est fourni par l'analyse des contes populaires (cf. [PRO 28]) qui repère dans ce genre spécifique des schémas assez stéréotypés associant des *personnages,* ou *figures* (le bon, le méchant, le sauveur, la princesse, le roi ou détenteur du pouvoir…) et une *intrigue* répondant par exemple à une succession du type : situation initiale paisible — épreuve (matérielle, morale…) – lutte – résolution (succès) – récompense – retour une situation paisible. Mais cette « macro organisation narrative » n'est pas spécifique des contes et apparaît, de manière éventuellement un peu moins figée, dans d'autres productions littéraires romanesques (Voir par exemple [ECO 85]). On la retrouve jusque dans des textes aussi simples et courts que des constats d'accidents ! Malgré la difficulté évidente, liée au caractère éminemment culturel de telles structures, il existe diverses tentatives de formalisation et d'opérationnalisation informatique, dont [SAB 00] se fait l'écho.

4.1.3. Structuration thématique et indexation intra-textuelle

Un texte est *à propos* de quelque chose, il possède un *sujet*, ou *thème* (la terminologie varie, reflétant des approches diverses dans un domaine encore mal stabilisé). Déterminer le thème d'un document est notamment la grande affaire des moteurs de recherche qui le décrivent en général (dans la technologie courante) par un ensemble de mots-clés. Mais à y regarder de plus près, on s'aperçoit que ce thème général se décompose ou se décline en un ensemble de « sous-thèmes » abordés dans divers *segments textuels*. La notion de *structuration thématique* renvoie à ce découpage en segments « thématiquement homogènes » et, bien sûr, à leur étiquetage par une forme explicitant leur thème. L'organisation en

chapitres, sections et sous-sections, chacune avec un intitulé censé en évoquer le contenu, est, lorsqu'elle est présente, une indication forte de la structure thématique. De même, des marques dispositionnelles telles que les énumérations. Mais d'une part cette organisation n'existe pas toujours, et d'autre part il existe d'autres unités thématiques, moins fortement indiquées, voire privées de marques explicites : les paragraphes ont en général une certaine homogénéité, mais souvent des regroupements de paragraphes sont pertinents, ou au contraire des assemblages de quelques phrases immergés totalement dans le texte.

– On peut considérer que cette structure conduit à une forme *d'indexation intra-documentaire*, c'est-à-dire une indexation par un « thème », plus ou moins restrictif, de segments textuel. Mais d'autres formes de segmentation-indexation peuvent être relevées, notamment dans la notion de *cadre de discours* de Michel Charolles [CHA 97]. Un cadre de discours est un segment textuel initié par une expression adverbiale détachée en initiale de phrase, appelée alors *introducteur de cadre*, qui exprime une contrainte d'interprétation portant sur l'ensemble des propositions qui composent le cadre. On aura différents types de cadres, selon la nature de l'introducteur : thématiques, introduits par une expression telle que *en ce qui concerne, quant à*… ; spatiaux et temporels, introduits respectivement par des groupes prépositionnels de ce type ; praxéologiques, comme dans *dans le domaine de la linguistique textuelle*… etc. La figure 6 fournit un exemple de cadre temporel, l'ensemble du segment devant être interprété (« indexé ») relativement à la période 1965-1985.[15]

> **De 1965 à1985**, le nombre de collégiens et de lycéens a augmenté de 70 %, mais selon des rythmes et avec des intensités différents selon les académies et les départements. Faible dans le Sud-Ouest et le Massif Central, modérée en Bretagne et à Paris, l'augmentation a été considérable dans le Centre Ouest, en Alsace, dans la région Rhône-Alpes et dans les départements de la grande banlieue parisienne où les effectifs ont souvent plus que doublé […]

Figure 6. *Cadre de discours temporel*

L'étude de la segmentation, en particulier thématique, des textes constitue un domaine de recherche actuel en linguistique en en TAL (voir par exemple des éléments de synthèse dans [PER 05] [ENJ 05b] [HER 04]). Du point de vue théorique, la notion de thème textuel entretient des rapports complexes, sans réduction de l'une à l'autre, avec la notion de thème phrastique évoquée dans la section précédente [BIL 05]. Son repérage automatique a des applications potentielles importantes en ingénierie documentaire, pour améliorer les moteurs de recherche dans le sens d'une meilleure qualité des descripteurs, pour réaliser une indexation intra-documentaire, c'est-à-dire permettant d'accéder directement aux *passages* pertinents du document, ou encore pour le résumé automatique.

Structure rhétorique

Il y a plusieurs « définitions » de la rhétorique. Une première l'attache à l'art de convaincre et à la notion d'argumentation — par exemple dans un article scientifique ou un texte/discours politique. C'est là ce que l'on peut appeler une rhétorique « ancienne », initiée par Aristote, et toujours présente dans un courant logico-philosophique (voir par exemple [PER 77]). Une autre, bien illustrée dans le chapitre 7, concerne ce que l'on peut appeler les « effets de style ». Une autre conception, partie prenante de tendances très actuelles en linguistique du texte, se pose la question générale de l'ensemble des relations qui existent entre les énoncés d'un texte, et qui en fondent la *cohérence*. On peut voir cette structure comme complémentaire de la structure thématique : ce dont on parle (éléments thématiques) *vs* l'articulation, l'enchaînement, entre ces différents éléments (structure rhétorique). L'argumentation devient alors un cas particulier d'articulation entre énoncés, avec il est vrai une dimension logique forte et caractéristique[16].

15. On peut alors considérer que le thème de ce segment est un objet complexe : le retard scolaire dans cette période particulière. On débouche sur la notion de *thème composite* proposée dans [BIL 05], qui généralise la notion de sujet ou thème traditionnel en associant plusieurs « dimensions » d'un espace notionnel (ici : scolarité + temps).

16. Ainsi, l'argumentation est souvent définie comme une démonstration (enchaînement d'inférences) « souples », n'ayant pas le caractère « absolu » du raisonnement mathématique et avec les ambiguïtés dues à son expression en langue naturelle.

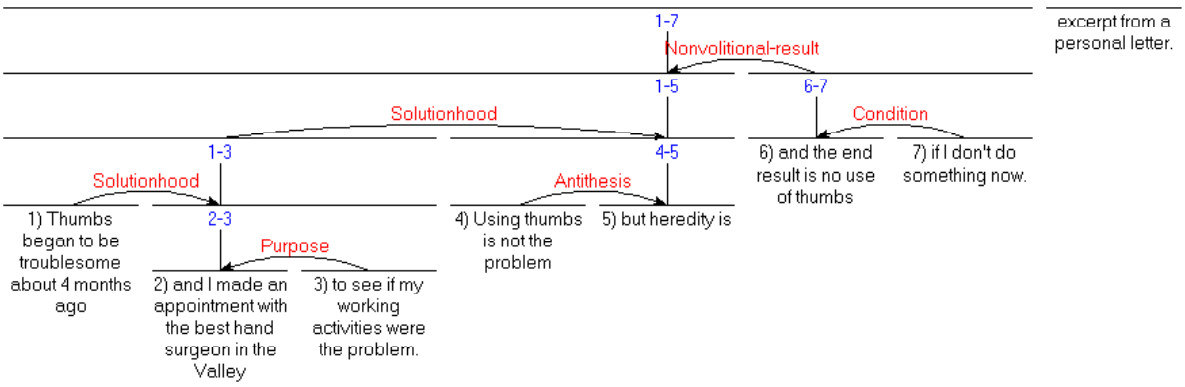


Figure 7 : *Analyse d'un texte selon la RST*

Des travaux relativement récents tentent de décrire et formaliser cette structure, particulièrement dans le cadre de deux théories à visée formalisatrice, la RST (*Rhetorical Structure Theory*) de Mann et Thompson [MAN 87] (voir également [SAB 00] pour une présentation rapide en français) et la SDRT (*Segmented Discourse Representation Theory*) de Nicholas Asher [ASH 93] continuatrice de la DRT de H. Kamp. Nous illustrons ce type d'analyse dans la figure 7 par un exemple emprunté à la première.

4.2. Le calcul des éléments de cohésion textuelle

Contrairement au niveau de la phrase, il n'existe pas de modèles de traitement intégré des différentes composantes sémantiques qui interviennent au niveau du texte. D'une manière générale d'ailleurs, les recherches sur les procédés de calcul sémantique sont beaucoup moins avancées à ce niveau. C'est donc une liste assez hétérogène de méthodes que nous allons présenter ici.

On peut tout de même classer les différentes techniques en deux grandes catégories, suivant le type de résultat qui est visé. Certains travaux cherchent d'abord et avant tout à *segmenter* un texte et à caractériser sémantiquement les segments obtenus : il s'agit de trouver de quoi on parle dans les différentes parties du texte, en détectant les ruptures dans le fil du discours. Il peut s'agir d'un changement de thème de discours, mais aussi d'un changement dans le cadre spatio-temporel, ou encore d'un changement de point de vue sur le thème traité.

D'autres au contraire cherchent à repérer des éléments qui se renvoient les uns aux autres tout au long du texte : il s'agit alors de mettre en évidence des *chaînes*, plus ou moins longues, d'expressions qui partagent une partie de leur contenu sémantique. Là encore, ces chaînes peuvent être de divers types : on peut distinguer d'une part des chaînes de coréférence d'entités, de lieux ou encore d'intervalles temporels, et d'autre part des récurrences lexicales et isotopies.

À vrai dire, ces deux types de calcul ne s'opposent pas entièrement, mais sont plutôt dans un rapport complémentaire. En effet, les ruptures thématiques s'accompagnent souvent de ruptures dans les chaînes, puisque les liens sémantiques sont plus étroits à l'intérieur d'un même segment qu'entre segments. Bien entendu, toutes les chaînes ne s'arrêtent pas aux frontières des segments : la cohésion du texte dans son ensemble provient en grande partie de ces chaînes qui relient les segments entre eux. Mais celles qui s'arrêtent présentent l'intérêt de caractériser le segment dans lequel elles se trouvent : notamment, les entités les plus importantes du point de vue thématique sont généralement évoquées par des chaînes de coréférence qui comportent le plus de pronoms clitiques (*il*, *elle*, *le*, etc.). Le calcul des chaînes peut donc aider à la segmentation et à la caractérisation des segments. Inversement, la connaissance des thèmes principaux des segments peut aider au calcul des chaînes : comme on va le voir, pour résoudre une anaphore pronominale, il faut connaître les entités les plus saillantes à un instant donné : celles qui sont déjà thématisées.

4.2.1. Segmentation et caractérisation de segments

Il convient d'abord de noter que la notion de segment textuel, possédant une certaine unité qui le caractérise au sein d'un document, n'est pas une notion simple : différents points de vue, différentes intentions de lecture, peuvent conduire à des découpages différents, et en général les limites des segments sont relativement floues. D'autre part, comme souvent en sémantique du texte, les critères pragmatiques (connaissances du domaine, codes de communication…), fort complexes à prendre en compte, jouent un rôle important dans la délimitation de tels segments. Néanmoins, nous évoquerons ici un ensemble de travaux, concernant la structuration thématique et l'encadrement du discours, qui commencent à approcher de manière convaincante certains mécanismes, en incluant des mises en œuvre informatiques.

S'agissant de la première problématique, l'approche la plus courante se fonde sur la notion de *cohésion lexicale* [HAL 76]. Dans la lignée de [YOU 91] et [HEA 94], elle vise à produire une segmentation linéaire du texte (dite *text-tiling*) — c'est-à-dire un découpage en segments adjacents thématiquement homogènes — en exploitant la répétition des mots comme indicateur d'homogénéité thématique. Chaque segment minimal (par exemple le paragraphe) est caractérisé par un vecteur associant à chaque descripteur (mots pleins ou lemmes par exemple) une valeur numérique représentative de sa fréquence dans ce segment, généralement obtenue par *tf. idf*. Une mesure de distance vectorielle permet alors d'évaluer la cohésion thématique de chaque couple de segments. Ces derniers pourront alors être regroupés, par seuillage sur cette distance, en unités homogènes. Du point de vue de la caractérisation des segments, on utilisera les descripteurs les mieux « notés » de chaque vecteur[17].

En se basant sur la distribution des formes de surface, ces méthodes ne requièrent aucune ressource externe, mais simplifient à l'extrême la dimension sémantique du phénomène de cohésion lexicale. Certains auteurs tentent de résoudre ce problème en faisant appel à une lemmatisation préalable et/ou à des réseaux lexicaux [KOZ 93], constitués éventuellement de manière automatique par examen de cooccurrences dans un corpus de référence [FER 02]. En complément de ces méthodes numériques, on peut exploiter des marqueurs, ou indices, linguistiques : typiquement, des annonces thématiques telles que *en ce qui concerne X, à propos de X, considérant X*… [POR 01] [FER 01].

Du point de vue des mécanismes sémantiques généraux, ces modèles exploitent donc deux principes : un principe *d'émergence* par répétition (récurrence lexicale) et un principe de *rupture* (soit implicitement par variation dans la dominante lexicale, soit explicitement par des marques discursives). Par contraste, le modèle de l'encadrement du discours repose centralement sur la présence de *marqueurs explicites* à caractère grammatical, dispositionnel et sémantique. L'ouverture d'un cadre est fort peu ambiguë, puisqu'elle est indiquée par une expression adverbiale détachée en position initiale de phrase. Notons à ce propos que l'introducteur a donc non seulement une fonction d'indexation sémantique du segment sur lequel il va porter (fonction *idéationnelle* selon la terminologie de Halliday), mais se comporte également comme une véritable *instruction discursive* d'ouverture de ce segment. Le problème de la *portée*, c'est-à-dire de la délimitation de la borne finale du cadre, est plus complexe (et sujet à variations selon les lecteurs). Plusieurs indices interviennent : dispositionnels, en particulier le changement de paragraphe ; morpho-syntaxiques tels que cohésion des temps des verbes ; sémantiques : par exemple une expression sémantiquement incompatible avec l'introducteur arrête en général le cadre. Cette question fait l'objet de travaux en linguistique et psycholinguistique [LED 01] [CHA 03]. Les réalisations informatiques les plus avancés sur le problème de la portée concernent les cadres temporels, avec des résultats très encourageants (près de 80 % en rappel, et 90 % en précision) [BIL 03].

4.2.2. Calcul de chaînes

Comme nous l'avons vu, il existe deux types de chaînes : les chaînes de co-référence et les isotopies. Les chaînes de co-référence, et plus précisément les chaînes de coréférence d'entités, ont fait l'objet du

17. On reconnaîtra là des méthodes bien connues en Recherche Documentaire traditionnelle, adaptées pour cette nouvelle tâche. Rappelons que le facteur tf.idf est une mesure qui pondère la fréquence d'un terme dans le segment considéré (*tf*) par l'inverse de sa fréquence dans le document dans son ensemble (*idf*).

plus grand nombre de travaux de modélisations, le problème le plus pressant étant, pour le TAL, la résolution des anaphores pronominales (nous verrons au § 5 un traitement des anaphores temporelles dans le cadre d'un modèle général du temps et de l'aspect, pour construire la structure temporelle d'un texte). Pour s'en tenir à l'essentiel, on peut citer ici deux courants principaux (pour une revue plus complète, voir [VIC 05]) : les approches axées sur la notion de *focus* [SID 83], qui, sous leur forme la plus aboutie, ont donné la *théorie du centrage* ([GRO 95], [WAL 98]), et les approches axées sur la notion de *saillance* ([ALS 87], [LAP 94], [DUP 03]). Les premières cherchent à définir, phrase après phrase, l'élément central sur lequel est focalisé le discours, et donc le meilleur candidat pour la résolution d'une anaphore dans la phrase suivante. L'idée essentielle du modèle est de caractériser les divers types de transition, permettant des changements de focus, plus ou moins préparés et attendus.

Les approches fondées sur la saillance sont plus globales : il s'agit de maintenir une liste des entités introduites dans le texte et de les ordonner par degré de saillance, celle-ci étant d'autant plus forte que l'entité a été évoquée plus fréquemment, plus récemment et de manière plus « thématique » (en position de sujet, par exemple). Cette liste est mise à jour après chaque phrase et elle sert à résoudre les anaphores de la phrase suivante. Dans le modèle de Michel Dupont [DUP 03], qui est sans doute le plus complet, la liste des entités est organisée dans un *modèle des attentes*. Ce modèle des attentes s'inscrit dans une théorie cognitive de l'interprétation qui confère une place essentielle à l'anticipation en fonction du contexte. L'idée générale, c'est que le lecteur ne reçoit pas passivement un message, il le fait entrer dans une boucle interprétative : avant la réception d'un énoncé, il anticipe ce qui va être écrit ; le message est donc confronté à ses attentes qu'il valide ou infirme. Et c'est en tenant compte des attentes du lecteur que le scripteur formule son propos. Cette boucle interprétative commence avant même la lecture de la première phrase du texte : l'environnement du texte alimente déjà les attentes du lecteur. Elles seront très différentes suivant qu'il entame la lecture d'un roman ou qu'il s'apprête à déchiffrer la notice d'utilisation de sa nouvelle machine à laver. Ensuite, tout au long du texte, le lecteur « met à jour » ses attentes en fonction de ce qu'il vient de lire. Dupont a montré que cette théorie constituait un cadre adéquat à la formalisation de la notion de saillance. Le modèle des attentes comporte donc une liste d'entités auxquelles est attribué un degré de saillance, mis à jour au fur et à mesure de l'avancée du texte. Trois plages principales de saillance sont définies :

– Les entités de *forte saillance* : ce sont des entités qui viennent d'être mentionnées dans le texte (dans les deux ou trois phrases précédentes). Elles occupent en quelque sorte le devant de la scène.

– Les entités de *saillance moyenne* : elles sont aussi présentes, mais en arrière-plan, pour filer notre métaphore scénographique (cf. la notion de *scène verbale* dans [VIC 99]). Entrent dans cette catégorie d'une part des entités dont la saillance a été forte et s'est dégradée parce qu'elles n'ont plus été mentionnées depuis plusieurs phrases, et d'autre part des entités dont on n'a pas encore parlé mais dont on s'attend à ce qu'elles soient évoquées/

– Enfin, une *saillance faible* est attribuée à toutes les autres entités présentes dans le modèle des attentes. En pratique on n'y trouvera que les entités dont la saillance a été moyenne et s'est dégradée, mais en théorie, on doit considérer que sont dans ce cas toutes les entités qui font partie des connaissances partagées entre le scripteur et ses lecteurs présumés (depuis la lune et le soleil jusqu'à Jésus-Christ ou Napoléon).

Dupont se sert de ces plages de saillance pour résoudre les anaphores en utilisant un principe de *concordance* : la forme linguistique utilisée pour l'anaphore (pronom personnel, démonstratif, syntagme nominal défini plus ou moins complet – cf. la théorie de l'accessibilité de Mira Ariel [ARI 90]) doit être en adéquation avec le degré de saillance de l'entité dans le modèle des attentes.

En ce qui concerne le deuxième type de chaînes, les isotopies lexicales, les méthodes de calcul sont nettement moins avancées. Sur la base du modèle développé par Rastier [RAS 95], un certain nombre de chercheurs ont conçu des méthodes d'aide au repérage et à la construction des chaînes isotopiques (voir notamment [TAN 00] et [PER 03]). D'autres chercheurs tentent de caractériser automatiquement un type de texte, en recherchant une forme d'isotopie donnée : ainsi Mathieu Valette [VAL 04] présente un système capable de détecter des sites Web racistes, en parvenant notamment à les distinguer des sites antiracistes, ce qui s'avère très délicat : les discours racistes et antiracistes partagent en effet un vocabulaire commun très important. Pour y arriver, Valette utilise les notions de *forme sémantique* et de

fond sémantique définies par Rastier [RAS 01]. Les discours racistes et antiracistes sur le Web partagent un même fond sémantique (on y trouve le même ensemble d'isotopies), mais les formes sémantiques, c'est-à-dire les groupements locaux saillants qui structurent les sèmes de ce fond sémantique sont radicalement différentes. Ces premiers efforts de modélisation montrent que la sémantique interprétative peut rendre compte de manière opérationnelle de la dimension lexicale de la cohésion et de la cohérence sémantiques des textes.

5. Un modèle de la temporalité en français

Le modèle de la temporalité développé par Laurent Gosselin que nous allons présenter pour clore ce chapitre est exemplaire à plus d'un titre :

– c'est un modèle « transversal » au sens où il traverse les trois paliers que nous avons définis ici : partant de représentations au niveau des unités linguistiques, il permet de calculer des relations temporelles pour chaque énoncé, et de les combiner pour aboutir à des représentations globales, des *chronogrammes*, au niveau du texte.

– c'est un modèle « cognitif » : il s'appuie sur une conception de la temporalité, la *sémantique aspectuelle*, qui considère que l'accès à la référence, en l'occurrence aux *procès* (situations et évènements), est toujours indirect : ils sont montrés, « mis en scène », selon un certain point de vue. Le locuteur ouvre des fenêtres temporelles dans lesquelles se déroulent ces procès, qu'il ne laisse voir qu'à travers ces fenêtres.

– c'est un modèle « calculatoire » : toutes les marques linguistiques qui concourent à l'expression de la temporalité sont décrites systématiquement, et les règles de calcul sont toutes explicitées. Ce modèle permet donc une démarche hypothético-déductive dans laquelle on peut tester la validité de telle ou telle règle, et, au-delà, du modèle tout entier, en comparant les résultats du calcul avec le jugement des locuteurs pour n'importe quel énoncé du français, que celui-ci soit construit spécifiquement pour le test ou collecté dans des corpus.

Nous allons brosser ici à grands traits les différents éléments du modèle, sans autre ambition que d'en donner un aperçu suffisamment général pour justifier notre appréciation. Pour une présentation complète, on pourra se référer à [GOS 96] et [GOS 05a][18]. On trouvera aussi dans [GOS 05b] un résumé synthétique du modèle et des implémentations informatiques qui en ont été faites.

5.1. Représentation du sens : relations entre intervalles temporels

Les représentations reposent dans le modèle sur des relations entre intervalles sur un axe temporel : relations de simultanéité, d'antériorité, de postériorité, de recouvrement, etc. Différents types d'intervalles sont définis :

– intervalle d'*énonciation*, noté ici IE : moment où le locuteur produit l'énoncé ;

– intervalle de *procès* (noté ici IP) : il est associé à un procès donné, ses deux bornes marquant le début et la fin du procès.

– intervalle de *référence* ou de *monstration* (IR) : c'est le moment dont on parle, la fenêtre temporelle par laquelle les procès sont montrés/perçus. Il y en a au moins un par verbe conjugué (deux s'il s'agit d'un temps composé).

– intervalle *circonstanciel* (IC) : associé à un complément circonstanciel de temps, qu'il s'agisse d'un complément de durée (*pendant une heure, depuis deux jours*,…) ou de localisation (*à huit heures, lundi dernier*,…).

Les représentations sont fondées sur des relations de trois types : l'aspect, le temps absolu et le temps relatif.

18. On trouvera notamment dans [GOS 05a] une extension du modèle à la dimension modale, centrée sur la notion de « coupure modale », introduite par la fenêtre temporelle de monstration (par laquelle sont montrés/perçus les procès) : tout ce qui est postérieur à cette fenêtre est modalisé comme possible, non irrévocable, même si c'est situé dans le passé, dans la mesure où cela n'a pas encore été « dévoilé » par le décours temporel du récit.

– l'aspect est défini par une relation entre l'intervalle du procès et un intervalle de référence. Il peut être *aoristique* (IP et IR coïncident ; ex. : *Il but une bière en fin d'après-midi*), *accompli* (IP précède IR ; ex. : *il avait déjà bu une bière au déjeuner*), *prospectif* (IR précède IP ; ex. : *Il allait encore boire une bière au dîner*), ou enfin *inaccompli* (IR est strictement inclus dans IP ; ex. : *A 17h, il buvait donc sa deuxième bière*). On remarquera que dans ce dernier cas, on ne « voit » pas les bornes du procès. Cela se justifie par des exemples tels que : *A 17h, il buvait donc sa deuxième bière depuis dix minutes quand les gendarmes l'ont arrêté*, dans lequel on précise que le procès a débuté dix minutes avant l'intervalle de référence ; quant à la borne finale du procès, elle n'a vraisemblablement pas été atteinte, à moins que les gendarmes n'aient fait preuve d'une mansuétude peu commune.

– le temps absolu : il a l'une des trois valeurs classiques, passé, présent et futur. Mais il est défini comme une relation entre l'intervalle d'énonciation et l'intervalle de référence, et non pas l'intervalle de procès. Cela se justifie cette fois par un exemple comme *Quand je l'ai vu il y a une heure, il croupissait au fond d'une cellule*. Ce qui est « vu » du procès est bien situé dans le passé, mais rien ne dit que le procès ne soit pas encore en cours à l'instant où le locuteur parle (au présent), et peut-être même pour longtemps (dans le futur).

– enfin le temps relatif est une relation entre intervalles de référence, d'antériorité, de simultanéité ou de postériorité. Par exemple, dans *Les gendarmes m'ont appris qu'il n'avait pas payé sa bière ce midi*, il y a relation d'antériorité entre l'intervalle de référence associé à la subordonnée (le moment du délit) et celui associé à la principale (le moment où le locuteur parle avec les gendarmes).

5.2. Calcul du sens : satisfaction de contraintes et résolution de conflits

Calculer le sens consiste à établir toutes les relations aspectuo-temporelles qui sont exprimées dans un énoncé donné. Pour cela, on associe à chaque marqueur de la langue des *instructions* qui sont des contraintes sur les intervalles temporels que nous venons de définir. Les marqueurs les plus importants sont :

– les marques de temps verbal : elles définissent des contraintes sur les relations entre les intervalles IE et IR d'une part, et IP et IR d'autre part. Par exemple, l'imparfait impose que IR précède IE (c'est un temps du passé) et que IR soit strictement inclus dans IP (c'est un temps imperfectif). Les temps composés définissent des relations avec deux intervalles de référence IR_1 et IR_2. Ainsi le passé composé définit un premier IR_1 coïncidant avec IP et antérieur à IE (d'où l'aspect aorisitique de certains passés composés, comme *Il a mangé il y a une heure*), et un deuxième IR_2 coïncidant avec IE (d'où son aspect accompli dans d'autres emplois, comme *Il a mangé depuis une heure*).

– le type de procès lexical : suivant la nature du verbe et de ses compléments, diverses contraintes portent sur l'intervalle IP. Par exemple *atteindre le sommet du rocher* impose que les deux bornes de IP soient infiniment proches (le procès est *ponctuel*), alors que *escalader le rocher* impose au contraire deux bornes disjointes.

– les compléments circonstanciels temporels : par l'intermédiaire de l'intervalle qui leur est associé, ils imposent des contraintes sur l'intervalle du procès IP quand ils sont intégrés au groupe verbal, ou sur l'intervalle de référence IR quand ils sont détachés (adverbes de phrase). Ainsi dans *Hier, j'ai escaladé le rocher en une heure et demi*, le complément détaché *hier* localise IR, tandis que le complément intégré *en une heure et demi* indique la durée de IP.

Bien entendu, il peut arriver que certaines contraintes, imposées par des marqueurs dans un même énoncé, soient incompatibles et conduisent donc à un *conflit*. Contrairement à la plupart des théories qui adoptent ce type d'approche par satisfaction de contraintes, comme les grammaires d'unification ou la théorie de l'optimalité par exemple, les conflits ne provoquent pas l'échec de la construction ou l'abandon de certaines des contraintes inconciliables, mais ils se résolvent par la complexification de la représentation globale. Des procédures régulières, qui se révèlent très puissantes, aboutissent à une représentation dans laquelle toutes les contraintes, sans exception, peuvent être satisfaites. Les deux procédures essentielles de complexification sont le *glissement de sens* et l'*itération*. Par exemple, dans l'énoncé *Hier, j'ai atteint le sommet du rocher en une heure et demi*, le conflit entre le caractère ponctuel du procès *atteindre* et la durée imposée par le circonstanciel se résout par un glissement de sens : on construit un deuxième procès, la phase préparatoire du procès évoqué, auquel s'applique la contrainte de durée, ce que l'on peut paraphraser par *j'ai mis une heure et demi à atteindre le sommet*. L'itération intervient par exemple quand il y a conflit entre un temps imperfectif imposant l'aspect inaccompli, donc IR inclus strictement dans IP, et un circonstanciel de durée impliquant l'aspect aoristique, donc IR égal à

IP. C'est ainsi que l'on rend compte du caractère itératif de l'énoncé *L'an dernier, j'escaladais ce rocher en moins d'une heure*.

5.3 Le niveau textuel : résolution d'anaphores temporelle

L'hypothèse principale du modèle pour traiter le niveau textuel est de considérer que les intervalles de référence sont intrinsèquement anaphoriques. Le lien temporel entre les énoncés est donc assuré par des règles de résolution d'anaphores temporelles qui portent sur ces intervalles de référence. Ces règles prennent en compte à la fois une notion de *saillance* des différents intervalles temporels et une notion de *proximité* relative privilégiant les intervalles les plus récents. L'exemple des trois phrases suivantes, tirées de [GOS 05b], donnera une idée de l'efficacité et de la finesse de ce mécanisme :

– *Mardi, il pleuvait*
– *Mardi, Pierre s'est promené. Il pleuvait*
– *Mardi, Pierre s'est promené. Il est rentré parce qu'il pleuvait.*

Dans le premier énoncé, l'antécédent de l'intervalle de référence IR, associé à *il pleuvait*, est *mardi*. Étant donné le caractère inaccompli de l'énoncé, on en déduit qu'il a plu toute la journée, ou, tout au moins, que la journée a été « pluvieuse ». Rien de tel pour le deuxième énoncé, où l'antécédent de IR est le procès *se promener* : ce qui est exprimé c'est seulement que le moment de la promenade a été pluvieux. La restriction est encore plus forte dans le troisième énoncé, où l'antécédent est le procès *rentrer* : on asserte cette fois qu'il a plu quand Pierre est rentré, ce qui n'implique pas qu'il ait plu pendant toute la promenade, ni *a fortiori* toute la journée.

Ce modèle est donc effectivement « calculatoire », y compris au niveau textuel, par des règles explicites : il a d'ailleurs été implémenté par Person [PER 04] [GOS 05b]. qui a montré que l'on pouvait obtenir automatiquement une représentation temporelle de textes complets sous la forme de *chronogrammes* (représentations géométriques de l'axe du temps sur lesquelles sont indiqués les différents intervalles temporels évoqués en respectant les relations qu'ils entretiennent).

6. Conclusion

Comme nous espérons en avoir convaincu le lecteur par la présentation, même très rapide, de modèles spécifiques, les avancées dans le domaine de la modélisation en sémantique sont assez nombreuses et importantes, y compris au niveau le plus difficile, celui du sens textuel. On peut estimer que cette discipline est arrivée à un certain degré de maturité, en intégrant de mieux en mieux des connaissances et des théories linguistiques. On devient aujourd'hui de plus en plus capable de construire des systèmes informatiques capables d'analyser correctement, fût-ce de manière partielle et sélective, des textes sur des domaines spécifiques de l'expérience humaine, comme l'information géographique ou médicale, par exemple.

Il est intéressant de noter que l'approfondissement de l'apport de la linguistique ne s'est pas accompagné, comme on aurait pu le craindre, d'une complexification des outils computationnels. Par exemple, le dernier modèle que nous avons présenté, le modèle de la temporalité de Gosselin, s'appuie sur des observations très fouillées sur la langue et sur une théorisation linguistique des plus solides. Il n'en reste pas moins qu'il est implémentable avec des outils informatiques relativement légers, qui permettent de traiter des textes « à la volée ». À l'heure où les besoins de l'informatique documentaire ont été démultipliés par Internet et l'accès à de très gros corpus, il est intéressant de constater que l'on peut disposer d'outils à la fois légers et efficaces : on peut penser en particulier que de tels outils seront au cœur de la prochaine génération de moteurs de recherche, qui accordera sans aucun doute une bien plus grande place à la sémantique linguistique.

Bibliographie

[ABE 00] ABEILLE A., BLACHE P., « Grammaires et analyseurs syntaxiques », in J.M. Pierrel (dir), *Ingénierie des langues*, chapitre 2, Hermès, 2000, p. 51-76.

[AGI 01] AGIRRE E., MARTINEZ D., « Knowledge sources for Word Sense Disambiguation », *Lecture Notes in Computer Science*, 2166, 2001, p. 1-10.

[ALL 95] ALLEN J., *Natural Language Understanding*, Benjamin/Cummings Pub. Co., 1995.

[ALS 87] ALSHAWI H., *Memory and context for language interpretation*, Cambridge Unversity Press, 1987.

[ARI 90] ARIEL M., *Accessing Noun Phrases Antecedents*, Londres, Routledge, 1990.

[ASH 93] ASHER, N., *Reference To Abstract Objects in Discourse*, Kluwer Academic Pub.

[BIL 03] BILHAUT F., HO-DAC M., BORILLO A., CHARNOIS T., ENJALBERT P., LE DRAOULEC A., MATHET Y., MIGUET H., PERY-WOODLEY M.-P., SARDA L., « Indexation discursive pour la navigation intradocumentaire : cadres temporels et spatiaux dans l'information géographique », *TALN 03*, Batz-sur-Mer, p. 315-320, 2003.

[BIL 05] BILHAUT F., « Composite Topics in Discourse », *Symposium on the Exploration and Modelling of Meaning, SEM 05*, Biarritz, Nov. 2005.

[CAD 01] CADIOT P., VISETTI Y.-M., *Pour une théorie des formes sémantiques, Motifs, profils, thèmes*, Paris, PUF, 2001.

[CAR 89] CARON, J., *Précis de psycholinguistique*, PUF, coll. Le psychologue, 1989.

[CHA 98] CHAMBREUIL M., *Sémantiques*, Hermes, 1998.

[CHA 05] CHARNOIS T., ENJALBERT P., « Compréhension automatique », in P. Enjalbert (éd.), *Sémantique et traitement automatique du langage naturel*, ch. 7, Hermès, p. 267-308 2005.

[CHA 95] CHAROLLES M., « Cohésion, cohérence et pertinence du discours », *Travaux de linguistique*, 29, p. 125-151, 1995.

[CHA 97] CHAROLLES M., « L'encadrement du dicours — Univers, champs, domaines et espace », *Cahier de recherche linguistique*, 6, p. 1-60, 1997.

[CHA 03] CHAROLLES M., « De la topicalité des adverbiaux détachés en tête de phrase », *Travaux de Linguistique*, 47, 2003, p. 11-51.

[CRO 04] CROFT W., CRUSE D.A., *Cognitive Linguistics*, Cambridge University Press. 2004.

[CRU 00] CRUSE D.A., « Aspects of the microstructure of word meanings », in RAVIN Y., LEACOCK C. (eds), *Polysemy : theoretical and computationl approaches*, Oxford University Press, p. 30-51, 2000.

[CUL 90] CULIOLI A., *Pour une linguistique de l'énonciation*, Ophrys, 1990.

[DER 90] DEERWESTER S., DUMAIS S. T., FURNAS G. W., LANDAUER T. K., HARSHMAN, R., « Indexing By Latent Semantic Analysis », *Journal of the American Society For Information Science*, 41, 1990, p. 391-407.

[DEL 01] DELSARTE PH., THAYSE A., *Logique pour le traitement de la langue naturelle — application à la langue française*, Hermès, 2001.

[DES 85] DESCLES, J.-P., « Représentation des connaissances : archétypes cognitifs, schèmes conceptuels, schèmes grammaticaux », *Actes Sémiotiques*, VII, 1985, p. 69-70.

[DUP 03] DUPONT M., Approche cognitive du calcul de la référence, thèse d'Univ., Caen, 2003.

[DUT 03] DUTOIT D, NUGUES P., DE TORCY P., « The Integral Dictionary : a lexical network based on computational semantics », *ICCSA, International Conference on Computational Science and its Applications*, Calgary, Springer, 2003.

[ECO 85] ECO U., *Lector in fabula*, trad. française, le livre de poche, n° 4098, Grasset, 1985.

[ENJ 05a] ENJALBERT P., VICTORRI B., « Les paliers de la sémantique », Enjalbert P (dir.), *Sémantique et Traitement automatique du langage naturel*, ch. 2, Hermès, 2005, p. 55-98.

[ENJ 05b] ENJALBERT P, BILHAUT F., « L'accès assisté à l'information documentaire », Enjalbert P (dir.), *Sémantique et Traitement automatique du langage naturel*, ch. 9, Hermès, 2005, p. 335-369.

[FAU 84] FAUCONNIER G., *Les espaces mentaux*, Editions de Minuit, 1984.

[FER 01] FERRET O., Grau B., Minel J.-L., Porhiel, S., « Repérage de structures thématiques dans les textes », *TALN 01*, Tours, 2001, p. 163-172.

[FER 02] FERRET O, « Segmenter et structurer thématiquement des textes par l'utilisation conjointe de collocation et de la récurrence lexicale », TALN, Nancy, 2002, p. 24-27.

[GOS 05a] GOSSELIN L., *Temporalité et modalité*, De Boeck-Duculot, 2005.

[GOS 05b] GOSSELIN L., PERSON C. « Temporalité », Enjalbert, P (dir.), *Sémantique et traitement automatique du langage naturel*, ch. 5, Hermès 2005, p. 173-213.

[GOS 96] GOSSELIN, L., *Sémantique de la temporalité en français*, Duculot, 1996.

[GRO 95] GROSZ B, JOSHI A., WEINSTEIN S., « Centering : a framework for modeling the local coherence of discourse », *Computational Linguistics*, 21 (2), 1995, p. 203-225.

[HAL 76] HALLIDAY M, ASAN R., *Cohesion in English*, Longman, 1976.

[HEA 94] HEARST M., « Multi-paragraph segmentation of expository texts », *Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics*, 9:16, 1994.

[HER 04] HERNANDEZ N., Description et Détection Automatique de Structures de Texte, Thèse en Informatique de l'Université de Paris XI, décembre 2004.

[IDE 98] IDE N., VÉRONIS J., « Word sense disambiguation : The state of the art », *Computational Linguistics*, 24 : 1, 1998, p. 1-40.

[JAC 05] JACQUET G., VENANT F., VICTORRI B., « Polysémie lexicale », Enjalbert P (dir.), *Sémantique et traitement automatique du langage naturel*, Hermès, ch. 3, 2005, p. 99-132.

[KAM 93] KAMP H., REYLE U., *From Discourse to Logic*, Kluwer Academic Press, 1993.

[KAY 97] KAYSER D., *La représentation des connaissances*, Hermès, 1997.

[KLE 90] KLEIBER G., *La sémantique du prototype : Catégories et sens lexi*cal, P.U.F., 1990.

[KLE 94] KLEIBER G., *Nominales : essai de sémantique référentielle*, Armand Colin, 1994.

[KLE 99] KLEIBER G. *Problèmes de sémantique : la polysémie en questions*, P.U.F., 1999.

[KOZ 93] KOZIMA H., « Text Segmentation Based on Similarity between Words », *Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics (Student Session)*, Colombus, USA, pp. 286-288, 1993.

[LAK 87] LAKOFF G., *Women, Fire and Dangerous Things*, Univ. of Chicago Press, 1987.

[LAN 87] LANGACKER R. W., *Foundations of Cognitive Grammar*, vol. 1 : *Theoretical Prerequisites*, Stanford University Press, 1987.

[LAP 94] LAPPIN S., LEASS H., « An algorithm for pronominal anaphora resolution », *Computational Linguistics*, 20 (4), p. 535-561, 1994.

[LED 01] LE DRAOULEC, A., PERY-WOODLEY, M-P. « Corpus-based identification of temporal organisation in discourse », *Proc. Corpus Linguistics 2001*, Lancaster, p. 159-166, 2001.

[POR 01] PORHIEL, S., « Linguistic expressions as a tool to extract thematic information », *Proc. Corpus Linguistics 2001*, Lancaster, p. 477-482.

[MAN 87] MANN, W. C., THOMPSON, S. A., « Rhetorical Structure Theory : a Theory of Text Organization », Livia Polany (éd.), *The Structure of Discourse*, Ablex Publishing Corporation, 1987.

[MAT 00] MATHET Y., Etude de l'expression en langue de l'Espace et du Déplacement : analyse linguistique, modélisation cognitive, et leur expérimentation informatique, Thèse de doctorat d'informatique, Université de Caen, 2000.

[MAT 05] MATHET Y., « Sémantique de l'espace et du déplacement », Enjalbert P (dir.), *Sémantique et traitement automatique du langage naturel*, Hermès, ch. 6, 2005, p. 215-259.

[MIL 90] MILLER G. A., BECKWITH R., FELLBAUM C., GROSS D., MILLER K.J., « Introduction to WordNet : An on-line lexical database », *International Journal of Lexicography*, vol 3, n° 4, p. 235-312, 1990. Voir aussi http ://www.cogsci.princeton.edu/~wn.

[MON 74] MONTAGUE R., « English as a formal language », in R. Thomason (ed.), *Formal philosophy, selected papers of Richard Montague*, Yale University Press, 1974.

[MUG 96] MUGNIER M.-C., CHEIN M., « Représenter des connaissances et raisonner avec des graphes », *Revue d'Intelligence Artificielle*, vol. 10, n° 1, p. 7-56, 1996.

[NAZ 98] NAZARENKO A. (éd.), *Compositionnalité*, *Revue T.A.L.*, Vol 39, n° 1, 1998.

[NUN 78] NUNBERG G., *The pragmatics of reference*, Indiana Univ. Linguistics Club, 1978.

[NUN 97] NUNBERG G., ZAENEN A., « La polysémie systématique dans la description lexicale », *Langue Française*, 113, 12-23, 1997.

[NYC 98] NYCKEES, V, *La sémantique*, Belin, 1998.

[PER 77] PERELMAN, C. *L'empire rhétorique. Rhétorique et argumentation*, Librairie philosophique J. Vrin, Première édition 1977, Seconde édition, 2002.

[PER 03] PERLERIN, V., BEUST, P., « Pour une instrumentation informatique du sens », *Variation, construction et instrumentation du sens*, Siksou M. (éd.), Hermès, 2003, p. 197-228.

[PER 04] PERSON, C., Traitement automatique de la temporalité du récit : implémentation du modèle linguistique SdT, Thèse de l'université de Caen, 2004.

[PER 05] PERY-WOODLEY, M.P., « Discours, corpus, traitements automatiques », in A. Condamines (dir.) *Sémantique et Corpus*, Hermès, 2005.

[PLO 98] PLOUX S., VICTORRI B., « Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes », *Traitement automatique des langues*, 39/1, p.161-182, 1998.

[PRO 28] PROPP, V. J., *Mophologie du conte*, Seuil, 1928.

[PUS 95] PUSTEJOVSKY J., *The generative lexicon*, MIT Press, 1995.

[RAS 87] RASTIER F., *Sémantique interprétative*, PUF, 1987.

[RAS 95] RASTIER F.,, « La sémantique des thèmes ou le voyage sentimental, Rastier, F. (éd.) *L'analyse thématique des données textuelles*, Paris : Didier, 1995, p. 223-249.

[RAS 01] RASTIER F., *Arts et sciences du texte*, PUF, 2001.

[ROS 75] ROSCH E., Cognitive representations of semantic categories, *Journal of experimental psychology*, 104, 1975, pp. 193-233.

[SAB 00] SABAH, G., GRAU, B., « Compréhension automatique de textes », in Pierrel J.M. (dir.), *Ingénierie des langues*, Hermès, 2000.

[SCH 98] SCHÜTZE H., « Automatic word sense discrimination », *Computational Linguistics*, 24 (1), 1998, p. 97-124.

[SID 83] SIDNER C., « Focusing in the comprehension of definite anaphora », in B. Grosz *et al.* (eds), *Readings in natural language processing*, Morgan Kaufmann, p. 362-394, 1983.

[SOW 84] SOWA J. F., *Conceptual Structures : Information Processing in Mind and Machine*, Addison-Wesley, 1984.

[TAL 00] TALMY L., *Towards a Cognitive Semantics*, vol. 1, MIT Press, 2000.

[TAN 00] TANGUY L., THLIVITIS T., « Parcours Interprétatifs (inter) textuels dans le cadre d'une assistance informatique », *Cahiers de Praxématique*, 33, p. 185-215, 2000.

[VAL 04] VALETTE M., « Sémantique interprétative appliquée à la détection automatique de documents racistes et xénophobes sur le Net », *Actes du 7e Colloque International sur le Document Electronique*, Patrice Enjalbert et Mauro Gaio (éds.), p. 215-230, 2004.

[VIC 96] VICTORRI B., FUCHS C., *La polysémie, Construction dynamique du sens*, Hermès, 1996.

[VIC 99] VICTORRI B., « Le sens grammatical », *Langages*, 136, 1999, p. 85-105.

[VIC 04] VICTORRI B., « Les grammaires cognitives », in C. Fuchs (éd.), *La linguistique cognitive*, Ophrys, p. 73-98, 2004.

[VIC 05] VICTORRI B., « Le calcul de la référence », in P. Enjalbert (éd.), *Sémantique et traitement automatique du langage naturel*, Hermès, chapitre 4, 2005, p. 133-172.

[WAH 93] WAHLSTER W., « Verbmobil : Translation of face-to-face dialogs », *Proc. 3rd Europ. Conf. On Speech Communication and Technology*, p. 29-38, Berlin, 1993.

[WAL 98] WALKER M., JOSHI A., PRINCE E., *Centering theory in discourse*, Clarendon Press, 1998

[YAR 00] YAROWSKY D., « Hierarchical decision lists for word sense disambiguation », *Computers and the Humanities*, 34 (1-2), 2000.

[YOU 91] YOUMANS, G., « A New Tool for Discourse Analysis : The Vocabulary-Management Profile », *Language*, 67 (4), 1991, p.763-789.