CONTRIBUTION A UN MODELE INTERACTIONNISTE DU SENSAmorce d'une compétence interprétative pour les machines

Pierre Beust

THÈSE
présentée
à l'U.F.R. de Sciences pour l'obtention du grade de
DOCTEUR DE L'UNIVERSITÉ DE CAEN
SPÉCIALITÉ INFORMATIQUE

par

Pierre BEUST

CONTRIBUTION À UN MODÈLE INTERACTIONNISTE DU SENS

Amorce d'une compétence interprétative pour les machines

Soutenue le
22 décembre 1998

Jury :

Anne NICOLLE	GREYC CNRS, Université de Caen	Co-directeur de thèse
Laurent GOSSELIN	DYALANG CNRS, Université de Rouen	Co-directeur de thèse
Bernard LEVRAT	LERIA, Université d'Angers	Rapporteur
François RASTIER	INaLF CNRS, École normale supérieure, Paris	Rapporteur
Jean SALLANTIN	LIRMM, Université de Montpellier	Rapporteur
Jacques COURSIL	GIL, Université des Antilles et de la Guyane	Examinateur
Jean-Marie PIERREL	LORIA CNRS INRIA, Université de Nancy I	Examinateur
Bernard VICTORRI	ELSAP CNRS, École normale supérieure, Paris	Examinateur

TABLE DES MATIÈRES

AVANT PROPOS

"Mesdames et messieurs, si vous voulez bien me prêter une oreille attentive ..."
Quelle phrase!
Voulez-vous me prêter l'oreille?
Il paraît que quand on prête l'oreille,
on entend mieux.
C'est faux!
Il m'est arrivé de prêter l'oreille à un sourd,
il n'entendait pas mieux !
Il y a des phrases comme ça ...

R. Devos, *Prêter l'oreille*, Matière à rire, Plon

Dans ce monologue, Devos se livre à un jeu humoristique sur la signification de l'expression *prêter l'oreille*. Il met ainsi l'accent sur une spécificité du langage naturel consistant en une variabilité des significations. Cette spécificité ne se limite pas à l'humour et on peut en trouver des manifestations dans d'autres formes de discours ; c'est par exemple le cas, comme nous le verrons au cours de cette thèse, dans les dialogues techniques. Dans le cadre d'une modélisation en informatique des langues naturelles, considérer ce genre de particularités comme a priori étrangères aux préoccupations d'un système artificiel serait très limitatif. La faculté de langage ainsi modélisée serait une approximation très imparfaite de ce qu'est une langue naturelle. La prise en compte de ces spécificités est donc incontournable et elle fait partie de ce que les usagers attendent de l'utilisation d'un système informatique doué d'une compétence langagière.

À l'heure du développement des réseaux de communication (Internet par exemple) et des technologies multimédia, la question d'une compétence langagière artificielle la plus naturelle possible apparaît comme un enjeu pour l'informatique. En effet, les problématiques de la traduction assistée par ordinateur, du résumé automatique de textes, de la recherche

documentaire dans de grandes bases de données, ou encore de la conception d'interfaces homme-machine plus conviviales, supposent que les machines réalisent des opérations sur des formes mais également sur du sens. Dans cette thèse, nous aborderons le problème du sens pour contribuer à en donner un modèle calculatoire dans une perspective informatique. Nous nous placerons principalement dans une problématique de conception de dialogue homme-machine en langue naturelle. Sans pour autant vouloir construire une machine sensible à l'humour de Devos, nous nous intéresserons à la façon dont des agents logiciels dialoguant pourraient construire et utiliser le sens tel qu'il existe en langue naturelle. Pour mettre en place un modèle pertinent, nous procéderons à une observation de dialogues réels car, comme nous le verrons, dans un dialogue, il y a co-construction du sens par les interlocuteurs et apprentissage de la compétence dialogique et langagière.

1. Environnement de travail

Cette thèse a été réalisée au sein du laboratoire GREYC (Groupe de Recherche en Informatique, Image et Instrumentation de Caen) et du pôle ModeSCo (Modélisation en Sciences cognitives) de la MRSH (Maison de la Recherche en Sciences Humaines) de l'Université de Caen. Elle s'inscrit dans les problématiques du groupe de recherche CODISIMA[1] (Coopération et Dialogue dans les Systèmes Multi-Agents). Elle est co-dirigée par Anne Nicolle, pour l'aspect informatique et par Laurent Gosselin, pour l'aspect linguistique. Elle s'insère dans un environnement de recherche sur les interactions langagières et le dialogue homme-machine que nous allons présenter succinctement.

Cet environnement s'est mis en place avec le projet Compèrobot qui est né d'un rapprochement pluridisciplinaire entre l'intelligence artificielle, la linguistique et la psychologie cognitive. Ce projet, débuté en 1991, est co-dirigé par Anne Nicolle (GREYC - Université de Caen) pour la dimension informatique, et par Jean Vivier (LPCP - Université de Caen) pour la dimension de psychologie développementale. La co-référenciation, comme processus de mise en place de références communes dans l'interaction, est l'objet d'étude central du projet Compèrobot. Elle y est envisagée dans le contexte particulier d'un dialogue enfant-machine. Des expériences réalisées avec la méthode dite du magicien d'Oz ont permis d'examiner les singularités dans ce type d'interaction. Elles ont intéressé la psychologie développementale pour l'étude du comportement de l'enfant dans l'acquisition du langage à travers des situations d'explicitation de consignes ([Vivier 93], [Jacquet 95]) et l'intelligence artificielle pour la spécification de la modélisation d'une compétence dialogique artéfactuelle ([Bricon-Souf 94], [Andrès 95], [Jullien 95]).

[1] Groupe de recherche du GREYC dirigé par Anne Nicolle.

À la suite de Compèrobot, le projet **PIC**[2] (**P**rocessus d'**I**nteraction en **C**onception) est une étude expérimentale et une modélisation des processus cognitifs et sociaux de conception distribuée à travers les traces qu'ils laissent dans les dialogues. Plus précisément, d'un point de vue théorique, il s'agit de montrer l'intérêt d'une perspective interactionniste pour l'étude de processus cognitifs, et d'un point de vue applicatif, il s'agit d'observer la conception de la documentation utilisateur d'un logiciel dans le paradigme de la "conception située et distribuée" ([Brassac & al. 96]) afin de montrer l'intérêt de la co-présence des rédacteurs techniques, développeurs et utilisateurs dès l'étape de la conception du document.

Dans la situation expérimentale mise en place, trois partenaires sont invités à se rencontrer pour une séance de travail dont le but est de commencer à concevoir la *documentation utilisateur* d'un logiciel de gestion des notes (*IUTNotes*) pour un département de l'Institut Universitaire de Technologie de l'Université de Caen. Pendant cette séance de travail de deux heures, le logiciel est installé sur une machine qui est mise à la disposition des trois intervenants.

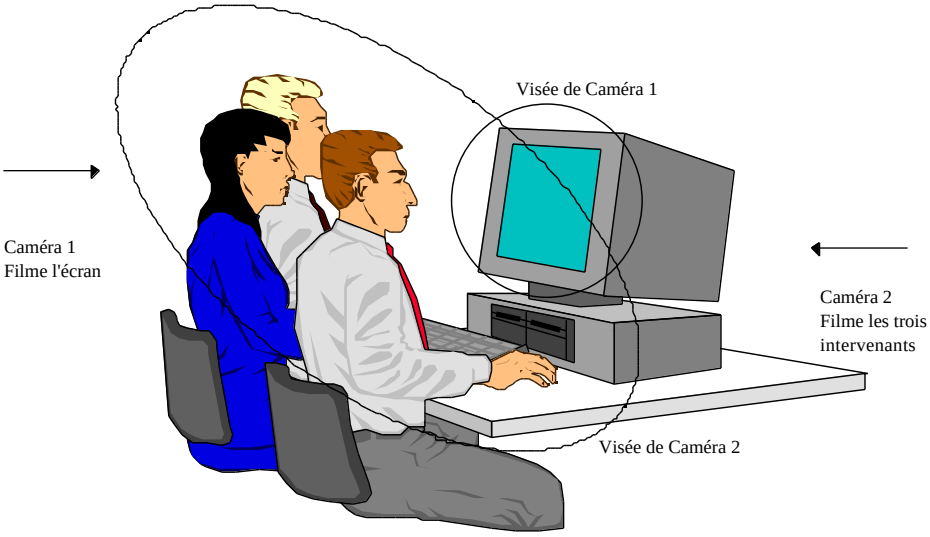


Figure 1 : La situation d'expérimentation du projet PIC

[2] Le projet PIC est un projet interdisciplinaire (soutenu par le GIS sciences de la cognition) commun aux laboratoires GREYC UPRESA CNRS 6072 (Groupe de Recherches en Informatique, Image, Instrumentation de l'Université de Caen), Groupe de Recherche "Modélisation du fonctionnement cognitif langagier" du LPCP EA 1774 (Laboratoire de Psychologie Cognitive et Pathologique - Université de Caen), GRC (Groupe de Recherches sur les Communications - Université de Nancy II), DYALANG (Linguistique - Université de Rouen) et à l'entreprise METAPHORA (Rédaction technique - Caen). Pour de plus amples détails sur le projet PIC, on pourra se reporter au site Web à l'adresse suivante : *http://www.info.unicaen.fr/~fgerard/pic/pic.html*

Les trois personnes présentes sont :

- un des développeurs du logiciel en question,
- une des secrétaires du département de l'IUT où le logiciel doit être mis en exploitation (c'est-à-dire une future utilisatrice),
- un rédacteur technique professionnel de la société Métaphora chargé de rédiger *le manuel de l'utilisateur* du logiciel.

La rencontre est enregistrée en audio et en vidéo (cf. Figure 1). Le corpus que nous utilisons pour expérimenter et valider le modèle proposé dans cette thèse est une retranscription de cette rencontre.

Dans les projets de recherche Compèrobot et PIC, une nouvelle approche pour l'étude et la modélisation des caractéristiques des langues a été proposée : il ne s'agit pas d'essayer de percer les secrets du fonctionnement mental, qui n'est pas directement observable, mais il s'agit d'analyser le déroulement des interactions langagières qui, quant à elles, sont observables. En ce qui concerne l'intelligence artificielle, il résulte de cette démarche que le modèle informatique de la compétence langagière et dialogique visé n'est pas un modèle du fonctionnement cognitif individuel, comme dans une perspective cognitiviste, mais un modèle de l'activité conjointe dans une perspective interactionniste. C'est dans et par les interactions langagières entre des agents humains et des agents logiciels ([Nicolle & al. 98]) que le modèle computationnel est construit. Cette perspective est celle qui a été suivie dans les thèses de Bricon-souf ([Bricon-Souf 94]), Andrès ([Andrès 95]) et Lehuen ([Lehuen 97]).

Cette approche de la langue et de l'interaction trouve un appui théorique dans les travaux linguistiques et informatiques de Coursil et plus précisément dans son modèle du transfert ([Coursil 93b]). Chaque agent étant modélisé comme le reflet de l'interaction dans laquelle il est acteur, il intègre de façon réflexive une image de soi, une image de l'interaction et de ce qui y est partagé, et une image de l'autre (voir des autres). Cette représentation du (ou des) partenaire(s) est construite sur le même modèle ([Delépine & al. 95]) : elle intègre une image de soi, une image de l'interaction et une image du (ou des) autre(s). Ceci implique, dans un mouvement réflexif, que l'agent représente la façon dont il est vu par l'autre. Les écarts entre la représentation de soi et l'image de soi dans la représentation de l'autre peuvent guider un modèle de dialogue dans la gestion de l'enchaînement conversationnel ([Andrès 95]). Cependant, ce que propose avant tout Coursil avec le modèle du transfert n'est pas de rendre compte du sujet parlant, mais c'est de décrire l'activité langagière silencieuse : celle du sujet qui interprète. Le principe de renversement du sujet ainsi posé amène à ne plus simplement considérer l'activité de parole comme un moyen de s'exprimer, mais comme un moyen pour être entendu par l'autre.

Dans ce contexte de recherche pluridisciplinaire sur l'analyse des interactions langagières, nous nous focaliserons sur la modélisation pour l'informatique de la sémantique des productions verbales du point de vue de l'interprétant. Nous contribuerons ainsi à la construction d'un modèle computationnel de la compréhension des énoncés dans l'interaction langagière pour un agent logiciel.

2. Préalable terminologique

Les contextes d'études pluridisciplinaires sont riches car ils apportent différents points de vue sur les mêmes objets et permettent les échanges de concepts et de méthodes d'une discipline à une autre. Le revers de la médaille de ces échanges et des réappropriations de concepts est la possibilité de voir se multiplier des concepts différents sous les mêmes noms. Par exemple, sémantique en logique et sémantique en linguistique sont deux domaines bien distincts qui pourtant portent le même nom. Afin de rendre la lecture de ce document moins déroutante, nous apportons ici quelques précisions de terminologie.

La première de ces précisions concerne les notions de sens et de signification qui seront présentes tout au long de cette thèse. À la manière de [Rastier 96], nous considérons que le sens est une propriété liée aux énoncés, aux séquences d'énoncés ou aux textes et que la signification est une propriété liée aux signes. Cette position va à l'encontre de celle qui consiste à attribuer une signification à la phrase et un sens à l'énoncé, en considérant l'énoncé comme une contextualisation de la phrase (i.e. une énonciation) et la phrase comme le matériau linguistique de l'énoncé. De notre point de vue, l'énoncé est la chaîne linguistique produite par un acte de langage. Le sens de l'énoncé est une prise de position pragmatique par rapport à la valeur locutoire et illocutoire de l'acte, qui donne à cet acte une valeur perlocutoire. En nous plaçant dans le cadre de la théorie des actes de langage de Austin, nous chercherons à rendre compte de l'acte locutoire, et plus précisément de l'acte rhétique. L'acte rhétique est linguistiquement appréhendable car issu de l'énoncé en tant que chaîne linguistique. Il est la contribution au sens de l'énoncé des significations de ses constituants et de leur organisation.

> Nous avons distingué assez schématiquement trois actes : phonétique, phatique et rhétique. L'acte phonétique, c'est la simple production de sons. L'acte phatique, c'est la production de vocables ou mots, c'est-à-dire de sons d'un certain type appartenant à un vocabulaire (et en tant précisément qu'ils lui appartiennent), et se conformant à une grammaire (en tant précisément qu'on s'y conforme). L'acte rhétique, enfin, consiste à employer ces vocables dans un sens et avec une référence plus ou moins déterminés. ([Austin 70, Huitième conférence, p. 109-110]).

Une deuxième précision concerne la notion d'effet de sens que nous utiliserons massivement. Nous entendons par effet de sens une co-adaptation des significations de

plusieurs constituants d'un énoncé précisant un même aspect de la valeur locutoire rhétique. Par exemple, la co-présence dans le même énoncé de *acheter*, *tomates* et *avocats* dans *j'ai acheté des tomates et des avocats au marché* conduit à considérer qu'il est ici question de légumes (et notamment pas d'homme de loi comme dans *j'ai rencontré mon avocat en ville*). C'est cette incidence de la co-présence textuelle sur le sens de l'énoncé que nous appelons effets de sens.

Enfin pour ce qui est de la notion de mot, qui formellement reste difficile à définir, nous la rapporterons (de même que [Pottier 74]) à celle de lexème en tant que signe lexicalisé en langue et reconnu dans l'interprétation d'un énoncé (ainsi *pomme de terre* est un seul mot). Ceci rejoint également le point de vue de Mel'cuk qui considère le lexème comme un mot pris dans une seule acception ([Mel'cuk 86, p. 4, Tome 1]). Nous ne nous intéresserons donc pas aux marques morphologiques qui font qu'un mot peut être vu comme un usage d'un lexème (par exemple *tableau* et *tableaux* sont deux mots issus d'un même lexème avec deux marques morphologiques de nombre différentes), mais nous nous focaliserons sur la sémantique des lexèmes en tant qu'elle permet d'attribuer des significations aux mots.

3. Plan de la thèse

Cette thèse comporte six chapitres :

• Le premier chapitre a pour but de mettre en place une problématique interactionniste de modélisation de la sémantique des langues naturelles en intelligence artificielle, focalisée sur la sémantique du français contemporain. Après avoir situé notre approche parmi les principaux modèles développés en Traitement Automatique des Langues (T.A.L.), nous préciserons notre objectif : la modélisation opératoire du sens pour la gestion des interactions langagières. Nous présenterons notre démarche d'observation de dialogues réels et nous introduirons les principes de base retenus, tels que, par exemple, la différence entre l'interprétation et la compréhension et le rapport entre le sens et la référence.

• Le deuxième chapitre présente l'étude linguistique des rapports sémantiques que peuvent entretenir les mots d'un même énoncé, ou d'une même séquence d'énoncés. Ces mots, en tant que lexèmes, seront les marqueurs sur lesquels s'appuie notre modèle opératoire. Leurs rapports mettent en jeu deux dimensions : celle de leur signification les uns par rapport aux autres, qui renverra au chapitre n°3, et celle des effets mutuels des significations dans la chaîne linguistique, qui renverra au chapitre n°4.

• Le troisième chapitre présente une modélisation informatique de la signification en langue des lexèmes à travers leurs rapports sur l'axe paradigmatique. Nous présenterons

les principes de catégorisation et de différenciation retenus pour cette modélisation, et nous détaillerons le modèle de représentation, appelé *Anadia*, qui en est issu. Nous expliquerons qu'Anadia met en place, dans une boucle interactive entre l'homme et la machine, un système de représentation de la valeur selon Saussure.

• Le quatrième chapitre traite des manifestations dans la chaîne linguistique (énoncé ou suite d'énoncés) des significations représentées au niveau paradigmatique. Il s'agira donc de présenter un modèle sémantique opératoire de l'axe syntagmatique. Ce modèle aura pour but de rendre compte des observations linguistiques détaillées au deuxième chapitre. L'accent sera mis sur la notion d'opérations interprétatives modélisées par une recherche de récurrences dans la chaîne.

• Le cinquième chapitre présente un outil logiciel réalisé par une articulation des modèles opératoires paradigmatique et syntagmatique. Sous la tutelle d'un assistant humain et par l'interaction homme-machine, cet outil a pour but de conduire un processus d'analyse sémantique d'énoncés par lequel il lui est possible d'acquérir des connaissances linguistiques qui seront utilisées dans les analyses ultérieures. L'amorce d'une compétence interprétative artificielle sera ainsi décrite.

• Le sixième et dernier chapitre apporte des éléments de validation du modèle et de l'outil logiciel proposé. Nous y détaillerons un état de la base de connaissance produite dans l'interaction homme - machine pour rendre compte d'une séquence de dialogue du corpus PIC. Nous ferons aussi un bilan de l'apport du modèle proposé dans cette thèse pour la problématique du dialogue homme - machine ainsi que pour d'autres problématiques du T.A.L. En cela, nous montrerons la généricité du modèle.

Le problème de la compétence sémantique pour un agent logiciel s'apparente aux problèmes de prise de décision, pour lesquels [Labat 98] a montré la nécessité d'une interaction entre l'utilisateur et le système afin de définir conjointement le problème, indiquer des préférences et des contraintes, spécifier les solutions et mettre au point des méthodes de résolution efficaces. À travers ces six chapitres, nous allons chercher à montrer que la mise au point d'une compétence interprétative artéfactuelle impose effectivement le recours à une interaction entre l'homme et la machine. Il y a deux raisons à cela : premièrement, le sens dans les dialogues constitue une prise de position conjointe ; deuxièmement, les connaissances linguistiques évoluent sans cesse et ne peuvent donc être données intégralement à un agent logiciel au préalable de sa mise en activité. L'apprentissage s'impose alors comme mode de

conception des agents logiciels interactifs, mais cet apprentissage n'est pas une phase préalable à l'utilisation de l'agent, il est intrinsèque à son activité d'attribution de sens aux énoncés.

CHAPITRE 1 :

ÉTUDE DE LA SÉMANTIQUE DES LANGUES NATURELLES EN VUE DE SON UTILISATION PAR DES MACHINES

Cette thèse a pour but de contribuer à un modèle sémantique de l'interprétation des énoncés langagiers qui soit utilisable par des machines.

La première partie de ce chapitre dresse un état de l'art des modèles de la sémantique des langues naturelles en intelligence artificielle, et permet de situer notre approche. La deuxième partie présente une étude d'interactions langagières réelles, et plus précisément des processus de mise en place de références communes dans un dialogue technique, pour mettre en place les fondements du modèle opératoire visé. Cette étude amène à proposer une conception par amorce de la compétence interprétative pour un agent artificiel, c'est-à-dire un processus où l'on commence par développer quelques capacités de base de cette compétence pour l'expérimenter au fur et à mesure que des connaissances sont acquises par l'agent. La troisième partie présente les principes retenus pour un modèle de l'interprétation. L'approche interprétative retenue fait la différence entre l'interprétation et la compréhension. Nous faisons l'hypothèse qu'au niveau sémantique, le sens d'un énoncé est modélisable en termes de contraintes prescrites uniquement par le matériau linguistique qu'on appellera son interprétation. Le sens en contexte d'un énoncé provient alors des prises de position pragmatiques de l'interprétant "à l'intérieur" de ces contraintes, en particulier en ce qui concerne la référence, ce qu'on appellera la compréhension,

qui est négociable dans l'interaction langagière et qui construit le terrain commun. Le sens se déploie alors dans l'espace formé par le plan sémantique et le plan pragmatique et l'établissement d'un terrain commun par le dialogue apparaît comme le moteur de l'interaction langagière. Notre objectif est de contribuer à faire un modèle computationnel de l'interprétation.

1. Intelligence artificielle et sémantique des langues

La sémantique s'est instituée comme science des significations au sein de la linguistique avec les travaux de Bréal ([Bréal 1897]), qui recherchait les principes généraux gouvernant les changements de sens diachroniques des mots. Dans les travaux qui ont suivi, la sémantique s'est également attachée à rendre compte des significations, et plus largement du sens, dans la dimension synchronique de l'étude du langage. L'enjeu linguistique qui est alors posé est de caractériser ce qui distingue deux séquences langagières (phrase ou énoncé) bien formées quand seulement l'une des deux fait sens. C'est donc mettre en forme, par l'étude des structures de la langue, des règles qui conditionnent la formation du sens pour rendre compte, par ces règles, de l'interprétation humaine sans forcément chercher à la simuler. Une démarche analytique qui conduit à la formation de ces règles consiste à mettre en évidence des marqueurs sémantiques dans le matériau linguistique et à organiser les dépendances réciproques entre ces marqueurs. C'est ainsi, par exemple, que Fuchs étudie la paraphrase en examinant comment plusieurs combinaisons de marqueurs manifestent le même sens ([Fuchs 94]), que Gosselin étudie la temporalité dans le rapport entre des situations relatives d'intervalles de temps et l'organisation de certains marqueurs ([Gosselin 96a]), que Tyvaert étudie la continuité référentielle dans le rapport entre la prédication et l'anaphore ([Tyvaert 95]), ou encore que Rastier étudie l'interprétation des textes par le rapport entre la combinatoire de traits sémantiques et les pratiques sociales où ces textes s'inscrivent ([Rastier 87]).

Les premières approches computationnelles de la sémantique ont été fortement influencées et guidées par les avancées de la syntaxe formelle ([Chomsky 69]). Il en résulte que la sémantique a d'abord été abordée comme une réinterprétation de l'arbre syntaxique[3], dans le but d'obtenir des expressions d'un langage formel ([Fillmore 68], [Jackendoff 72]). Ces approches ont donné lieu à des modèles en couche ([Van Dijk 77]) où le déroulement d'une analyse consiste en un enchaînement de plusieurs étapes dont chacune fournit une donnée pour l'étape suivante (comme dans [Morris 39]). Ainsi, à une analyse lexicale succède une analyse syntaxique, elle-même suivie d'une analyse sémantique donnant finalement lieu à une analyse pragmatique.

[3] C'est ainsi qu'est envisagé le calcul sémantique dans les langages formels. À chaque noeud de l'arbre syntaxique est attaché un opérateur dont les arguments résultent de l'évaluation des sous-arbres issus de ce noeud (ses fils). Le calcul du sens d'un énoncé formel consiste alors à évaluer l'opérateur attaché à la racine de l'arbre syntaxique.

Le développement de la théorie de la compilation et des langages formels[4] fondés sur la logique a conforté cette position qui consiste, entre autre, à donner à la sémantique un statut second par rapport à la syntaxe[5]. Ceci a renforcé l'idée d'une similarité entre la sémantique des langages formels et la sémantique des langues naturelles. L'exemple le plus frappant est la sémantique formelle de Montague ([Montague 70]). Si les différences entre langages formels et langues naturelles[6] ont poussé à rejeter le strict isomorphisme entre les deux sémantiques, il n'en reste pas moins que l'on trouve souvent en intelligence artificielle l'idée de formuler le sens d'un énoncé en langue naturelle par un jeu d'expressions logiques. C'est notamment le cas dans la DRT (*Discourse Representation Theory*) de Kamp ([Kamp & al. 93]) ainsi que dans les travaux qui la prolongent (comme la SDRT : *Segmented Discourse Representation Theory* de [Asher & al. 94]). L'analyse sémantique se donne alors comme but d'établir une représentation propositionnelle canonique du sens qu'une analyse du discours complète ensuite en procédant au calcul d'expressions référentielles, de relations temporelles et de résolution d'ellipses. Le résultat produit est une formule donnée comme une représentation d'un sens que l'on peut qualifier de littéral et qui ne peut rendre compte des divers sens possibles d'un même énoncé. En réponse aux difficultés inhérentes à ce type d'approche, [Groenendijk & al. 96] proposent une sémantique formelle dynamique pour la modélisation du sens en langue naturelle. Cette sémantique dynamique formule le sens comme un potentiel de changement d'information en opposition aux sémantiques formelles statiques où le sens est ramené à des conditions de vérité[7]. Ce potentiel de changement d'information rend compte de l'évolution d'un système de référents possibles au cours du discours. L'approche de [Groenendijk & al. 96] apporte une vision du sens plus "souple" que l'approche purement propositionnelle. Toutefois la sémantique formelle, qu'elle soit statique ou dynamique, présente une différence fondamentale avec la sémantique des langues naturelles : elle est uniquement référentielle (ou dénotationnelle), c'est-à-dire qu'elle présuppose les objets du monde (c'est donc, comme le signale [Sallantin 97, p. 29], que ce monde doit être connu entièrement), ce qui n'est pas le cas des langues naturelles qui permettent, comme on le verra plus loin, de construire le monde en même temps qu'elles l'évoquent.

---

[4] Pour un exposé précis de l'historique du TAL à partir de processus manipulant des symboles, on pourra se référer à [Sabah 90, p. 23-28].

[5] Cette position entre en contradiction avec les travaux de Piaget ([Piaget 23]) en psychologie développementale qui établit chez l'enfant une apparition première d'une compétence pragmatique (l'enfant reconnaissant dans le langage un moyen d'action) et une apparition tardive d'une compétence syntaxique.

[6] Pour plus de détails sur ces différences, on pourra se référer à [Raccah 97] qui montre qu'une première différence fondamentale entre les logiques et les langues naturelles concerne la notion de vérité. En logique, elle est à opposer à la fausseté, tandis qu'en langue naturelle la vérité est à opposer au mensonge.

[7] Dans les langages formels, la sémantique résulte d'une mise en correspondance d'expressions du langage avec un domaine particulier où ces expressions sont vraies ou fausses. Dans la terminologie logique, cette correspondance constitue une *interprétation.*

Pour présupposer les objets du monde indispensables aux logiques sous-jacentes aux sémantiques formelles, deux types d'approches ont été élaborées pour la modélisation en informatique de la sémantique des langues : des approches centrées sur une tâche et des approches centrées sur des corpus homogènes. Le système EDWARD ([Huls & al. 95]) est un exemple du premier type d'approche, où il s'agit de mettre en place un gestionnaire de fichiers informatiques. EDWARD met en place un système de commandes sous forme d'énoncés en langue naturelle où la résolution des références est accompagnée d'une désignation déictique à l'aide du pointeur de la souris (par exemple, on établit la référence de l'énoncé *duplique ce fichier là* grâce à l'icône cliquée après la saisie de l'énoncé). Les approches centrées sur des corpus homogènes permettent aussi de présupposer la référence en construisant préalablement à toute analyse un modèle du monde en lien avec le type de corpus considéré. On trouve ce type d'approche (par exemple le système FASTUS détaillé dans [Appelt & al. 93]) dans les conférences MUC[8] (Message Understanding Conferences). Le but est de traiter des dépêches d'agences de presses où il est question de transactions économiques, de vie politique, ou encore d'annonces d'attentats terroristes. On sait donc, par exemple, que dans le cas d'une transaction, on devra entre autre rechercher dans la dépêche des informations sur des sociétés (au moins deux), des noms de responsables de ces sociétés, un montant financier, une date et un lieu, etc. Ces approches apportent de bons résultats par rapport à ce que l'on en attend ([Chinchor & al. 93]) mais le but du processus analytique est changé : il ne s'agit plus de comprendre un texte (dans l'absolu) mais d'en extraire des informations partiellement connues. Le but est de remplir des bases de connaissances dans une perspective de recherche documentaire, et non de modéliser la recherche du (ou d'un) sens d'un texte.

La diversité des formalismes de modélisation de la sémantique des langues naturelles révèle les différences profondes entre les théories sur le but même d'un calcul sémantique. C'est la nature même du sens qui est en jeu. Comme on l'a vu, rechercher le sens d'un énoncé, cela peut être calculer ses conditions de vérité (comme c'est le cas dans les sémantiques formelles), déclencher une commande ([Huls & al. 95]), ou remplir des bases de connaissances ([Chinchor & al. 93]). Dans d'autres disciplines que l'intelligence artificielle, ce but peut être de rendre compte d'une cohésion textuelle, comme c'est notamment le cas dans la théorie de la sémantique interprétative de Rastier ([Rastier 87]), ou encore de rendre compte de la mise en place de représentations mentales lors de la compréhension, comme c'est le cas dans l'approche psycholinguistique suivie par la théorie des modèles mentaux ([Johnson-Laird 83]).

S'il s'agit d'éclairer la modélisation de l'enchaînement conversationnel, que ce soit en actes ou en paroles, le sens n'est plus un but en soi, un objet qu'il convient de représenter, mais

[8] La dernière conférence MUC (MUC-6) s'est tenue en Septembre 95. Pour plus de précisions, se reporter à l'adresse Internet : *http://cs.nyu.edu/cs/faculty/grishman/muc6.html*.

une ressource qu'il faut cibler pour permettre à des actes d'être accomplis. On a signalé dans l'introduction de cette thèse que le sens est un enjeu en informatique pour la traduction assistée par ordinateur, le résumé automatique de textes, la recherche documentaire, ou encore la conception d'interfaces conviviales. Cet enjeu amène à apprendre aux machines comment manipuler correctement le sens et il n'est pas nécessaire pour cela de le représenter. Ainsi, l'utilisation de la sémantique des langues pour les machines n'implique pas obligatoirement qu'il existe une représentation du sens. C'est le point de vue que nous défendrons en cherchant à modéliser de façon opératoire les processus d'interprétation. C'est également le point de vue défendu par Levrat ([Levrat 93]) qui met au point un outil informatique de *diagnostic de paraphrase* en proposant, en alternative à une représentation interne du sens, un jeu de règles de réécriture.

Les différences de point de vue sur la nature du sens sont également liées à des différences sur la façon de voir la signification des signes linguistiques qui composent les énoncés (i.e. le matériau linguistique). Que l'approche retenue soit, ou ne soit pas, compositionnelle (i.e. le sens d'un énoncé est considéré comme une construction à partir des significations des constituants de l'énoncé), une théorie sémantique qui se donne pour objectif le calcul du sens doit rendre compte de la contribution du matériau linguistique au sens de l'énoncé. Selon un point de vue componentiel, la signification d'un signe peut être représentée par un ensemble de composants élémentaires que l'on appelle des traits sémantiques (ou *sèmes*), tandis que selon un point de vue non componentiel, la signification d'un signe peut être considérée comme une sorte d'atome qu'on ne pourrait subdiviser (c'est le cas d'un prédicat logique par exemple), ou encore ne pas être représenté du tout en tant que tel mais produite par un jeu de relations. C'est notamment le cas en ce qui concerne les systèmes connexionnistes ([Victorri 98]). C'est aussi le cas dans les formalismes de réseaux sémantiques ([Quillian 68] ou plus récemment les graphes conceptuels [Sowa 84]). Dans les réseaux sémantiques, la signification d'un signe est exprimée par un ensemble de relations fixées entre différents concepts (la plus célèbre est la relation *IS-A* exprimant en général un rapport hiérarchique entre un concept et un sous-concept).

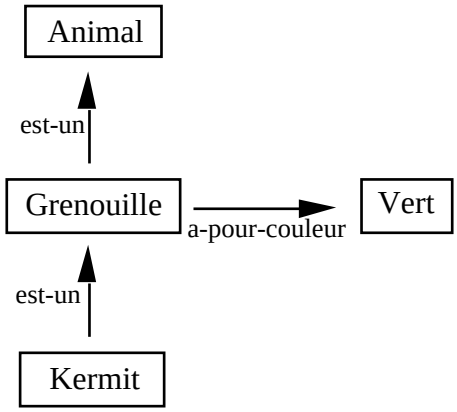


Figure 1.1 : Un exemple de réseau sémantique ([Baader 97, p. 5])

L'utilisation des réseaux sémantiques pour la compréhension du langage naturel pose certains problèmes. Il n'y a pas de claire distinction dans les réseaux sémantiques entre un objet, en tant qu'individu, et un concept en tant que type, car un individu est représenté comme un domaine conceptuel réduit à un singleton. Les relations entre les noeuds du réseau sont non homogènes, et notamment la relation *IS-A* (est-un) bien que représentant principalement un rapport hiérarchique peut aussi être utilisée pour une relation d'instanciation. De ce fait, les réseaux sémantiques ne font pas la différence entre l'appartenance de concepts (comme dans *Une grenouille est un animal*) et le rapport de l'instance à sa classe (comme dans *Kermit est une grenouille*). La langue naturelle (et plus particulièrement le français contemporain) utilise également la copule *être* pour exprimer ces deux types de relations, mais en langue la relation *être un* n'est pas ambigue parce qu'elle est précisée par d'autres indices dans la forme des énoncés. Ainsi, l'appartenance exprimée dans

Une grenouille est un animal.

n'est pas uniquement marquée par le prédicat *est un*, mais elle provient également de l'utilisation du déterminant *Une*. Le déterminant est ici nécessaire dans la forme de l'énoncé, il participe au sens de la relation prédicative, c'est pourquoi la phrase suivante ne peut exprimer une appartenance conceptuelle (c'est linguistiquement un cas barré, c'est-à-dire une non-phrase) :

***** *Grenouille est un animal.*

Alors que, s'il s'agit d'exprimer le rapport d'une instance à sa classe, c'est l'absence de déterminant qui précise le sens de la relation *être un.* Ainsi,

Kermit est une grenouille.

exprime une instanciation alors que les deux phrases suivantes sont des cas barrés :

***** *Une Kermit est une grenouille.*

***** *La Kermit est une grenouille.*

À la différence de ce qui est représenté dans les réseaux sémantiques, la signification du prédicat *être un* en langue n'est donc pas autonome. C'est-à-dire que *être un* en soi n'est pas suffisant pour préciser hors-contexte sa signification, étant donné que cette signification va varier en fonction de la présence ou de l'absence d'autres marqueurs linguistiques. C'est parce que les significations en langue sont le plus souvent non-automomes et donc *a priori* multiples (ce que l'on appelle la *polysémie*) qu'une approche purement compositionnelle du calcul du sens (de la même façon que la compilation dans les langages formels) n'est pas adaptée à la sémantique des langues (nous y reviendrons au chapitre 2).

Les problèmes d'hétérogénéité de relation des réseaux sémantiques ont amené à mettre en place d'autres modèles de représentation de connaissances où les relations sont réifiées (ou au moins étiquetées) comme, par exemple, les *Structured inheritance networks* de Brachman[9] ([Brachman 78]) ou encore les *Frames* de Minsky ([Minsky 75]). Tant qu'il s'agit de représenter en machine des connaissances ontologiques, ces modèles sont tout à fait adaptés. Leurs évolutions ont apporté beaucoup à l'informatique et l'apport des frames, par exemple, à la programmation par objets est indéniable.

Cependant, comme candidats à la modélisation de la sémantique des langues, ces modèles (réseaux sémantiques, frames ou encore graphes conceptuels) présentent à nos yeux un inconvénient majeur : considérer la signification comme une représentation ontologique. On peut alors leur faire la même critique qu'aux modèles basés sur des logiques formelles, à savoir de considérer uniquement la sémantique d'un point de vue référentiel. Non plus parce que la référence est présupposée avant l'analyse, mais parce que la signification des signes est modélisée comme des concepts liés à une représentation du monde[10].

Considérons par exemple la signification du mot *apéritif*[11] en tant qu'un concept faisant référence à une boisson alcoolisée que l'on consomme avant un repas. L'interprétation de l'énoncé

---

[9] Dans le système KL-ONE développé par Brachman, trois axes dans les représentations sémantiques sont distingués : le premier selon lequel s'organise la hiérarchie des concepts en classes et sous-classes, le deuxième qui permet d'expliciter d'autres rapports entre concepts et le troisième qui précise les rapports entre les éléments de description (à l'aide d'un nombre fini de liens primitifs).

[10] Dans les travaux de philosophie du langage, on peut citer les travaux de Gottlob Frege qui s'opposent aux sémantiques référentielles comme celle de Tarski. Frege explique pourquoi l'assertion *"a est identique à b"* est plus informative que l'assertion *"a est identique à a"* quand selon un point de vue référentiel, ces deux assertions sont équivalentes (i.e. quand elles expriment l'identité d'un objet avec lui même). Selon Frege, un énoncé en langue (à la différence d'une expression logique) fait intervenir, en plus des informations relatives à son contenu, des informations concernant la façon d'atteindre ce contenu. Pour cette raison l'identité en langue est posée différemment qu'en logique (d'où la différence de sens entre *"une femme est une femme"* et *"une femme est un être humain de sexe féminin"*).

[11] Cet exemple est extrait de la conférence de Gérard Sabah du 14/12/96 à l'ATALA intitulée "Le point sur le sens": *http://www.limsi.fr/Individu/gs/textes/ATALA-14.12.96/LePointSurLeSens2.html*

Je prendrai un apéritif sans alcool

conduit à un problème car il n'existe aucune boisson alcoolisée sans alcool. Cet énoncé n'est donc pas interprétable dans un modèle où la signification est une référence directe au monde. Cela signifie que *sans alcool* est un sélecteur opérant non pas sur le référent, mais sur le signifié d'*apéritif* et donc que référence et signification sont deux choses distinctes. De même, considérons que la signification du mot *chat* soit une référence à un animal mammifère à quatre pattes qui miaule et qui appartient à la classe des félins. Comment alors interpréter l'énoncé suivant ?

Il y a chats et chats.

En considérant une signification ontologique pour *chat,* cet énoncé apparaît comme un tautologie ou encore comme un non-sens. Pourtant, il est tout à fait envisageable dans le cadre d'une conversation où quelqu'un expliquerait la différence entre un chat de gouttière et un chat siamois. Ce n'est alors pas une tautologie mais une forme spécifique en langue naturelle pour exprimer une différence (en terme linguistique, cet énoncé convoque une opération interprétative appelée *dissimilation* ; cf. chapitre 4). De même, une signification référentielle du mot *chat* conduit l'interprétation de l'énoncé suivant à un non-sens.

J'ai un chat dans la gorge.

Cet énoncé en langue a un sens exprimant que le locuteur se plaint d'avoir la voix enrouée. Il est clair dans le cas présent qu'il n'est pas question d'un animal à quatre pattes. Le sens de cet énoncé n'est donc pas lié à une représentation ontologique de ce qu'est un *chat.*

De l'assimilation entre signification et référence résulte l'isolement du signe linguistique dans un rapport terme à terme avec un objet, étant donné que le signe est défini de façon indépendante des autres signes. Il apparaît alors difficile de constituer la sémantique en système autonome. Une sémantique référentielle ne peut, au plus, que constituer des inventaires. Elle réduit, de ce fait, la linguistique à une grammaire et une ontologie ([Rastier & al. 94, p. 19]). C'est parce que la langue n'est pas un modèle du monde, mais un phénomène social, que les modèles basés sur la référence ne peuvent en rendre compte avec fidélité (ils peuvent rendre compte des modèles de la langue, mais pas de la langue elle-même).

Pour rendre compte des propriétés spécifiques de la sémantique des langues naturelles (par exemple, la dissimilation dans *Il y a chats et chats* ou encore la polysémie, comme nous le verrons dans le chapitre suivant), notre choix est de chercher ailleurs que dans la logique ou dans les modèles de représentation de connaissances les théories qui en permettent une modélisation opératoire pour l'intelligence artificielle. En cela, nous nous situons dans une perspective qui cherche à opérationnaliser une approche linguistique de la sémantique, c'est-à-dire une approche focalisée sur l'analyse de marqueurs sémantiques où le sens n'est pas de

nature référentielle (i.e. extralinguistique) mais concerne l'activation du système de la langue (c'est ainsi d'ailleurs que selon [Bachimont 96, p. 130], on peut mettre en place une théorie sémantique autonome). On retrouve notamment en informatique ce même type d'approche dans les travaux de Tanguy ([Tanguy 97]) qui propose une assistance à l'interprétation des textes basée sur la sémantique interprétative de Rastier ([Rastier 87]), ou encore chez Ferrari ([Ferrari 97]) qui cherche, à partir de marqueurs linguistiques, à localiser automatiquement des métaphores dans des documents écrits.

En nous référant aux théories sémantiques élaborées par les structuralistes, les notions de valeur et d'isotopie vont être utilisées dans une perspective calculatoire. La notion de valeur saussurienne (cf. Chapitre 3) en tant que justification différentielle de la signification permet de se placer d'entrée dans le cadre d'une sémantique non référentielle qui fonde la langue en tant que système. La notion d'isotopie (cf. Chapitre 4), que l'on doit à Greimas ([Greimas 66]), offre les bases pour une approche interprétative de la sémantique des langues naturelles.

2. L'interaction langagière comme objet d'étude

Un bon nombre de travaux en T.A.L. ([Ferrari 97] ou [Déjean 98] par exemple) ont maintenant une démarche expérimentale d'analyse de corpus, et principalement d'analyse de textes, pour mieux concevoir leurs modèles informatiques. Dans le même esprit, nous allons mener une étude d'analyse de corpus de dialogues réels.

Lorsque la recherche porte sur des activités langagières monologales (textes écrits ou discours oraux), elle considère le sujet seul face à la langue. Considérer en premier lieu le sujet (parlant ou interprétant) dans l'étude du langage est une approche qui relève de la psycholinguistique, étude expérimentale des processus psychologiques par lesquels un sujet humain acquiert et met en oeuvre le système d'une langue naturelle (définition extraite de [Caron 89]). On peut voir ici un déplacement du centre d'intérêt du système au sujet. C'est le sujet et ses capacités cognitives individuelles qui sont visés, et l'interaction langagière, en tant que telle, ne constitue pas l'objet d'étude fondamental. Les notions de valeurs (signifiés), sens ou référence restent rapportées aux capacités cognitives du sujet interprétant (comme c'est aussi le cas dans la *Grammaire Cognitive* de [Langacker 87]).

En analysant le sens dans les interactions langagières nous cherchons à faire la synthèse entre deux points de vue. Le premier point de vue est la contribution au sens de la langue en tant que système. Ainsi, dans les dialogues, les énoncés produits par les interlocuteurs fournissent autant de marqueurs sémantiques qu'un texte pour guider un processus opératoire. On y

retrouve notamment des effets de sens produits par l'utilisation conjointe de mots, ainsi que des reprises anaphoriques. Ces marqueurs participent à la cohésion des énoncés d'où provient leur intelligibilité, tout comme on peut parler de cohésion textuelle dans l'analyse de textes (nous en évoquerons l'impact sur le sens des énoncés dans le chapitre suivant). Le second point de vue est la contribution au sens de l'interaction. L'intérêt d'une étude de dialogues réels est de montrer la langue en usage dans une interaction. Quand on cherche avant tout à rendre compte de ce qui se passe lors de dialogues, les notions de valeurs, de sens et de référence s'en trouvent changées dans la mesure où elles ne proviennent plus uniquement des capacités individuelles des interlocuteurs mais où elle sont construites et négociées dans l'interaction. Ainsi, comme l'ont montré les résultats des travaux de Vivier ([Vivier 93]) en psychologie développementale, le dialogue construit le dialogue. En effet, la valeur d'un mot (d'un signe) n'est pas due à un mécanisme individuel d'inférences sur les perceptions, elle est déterminée socialement par l'usage de la parole et chaque sujet y apporte sa contribution. De même, dans un dialogue, le sens n'est pas calculé par chaque sujet de façon individuelle. Il constitue une prise de position conjointe de la part des acteurs qui n'est ni prioritairement le fait de l'un ou de l'autre mais qui est co-construit dans l'échange, et [Clark & al. 86], [Vivier 92], et plus récemment [Brassac & al. 96], ont montré que cette co-construction nécessite au moins trois tours de parole. De ce point de vue, le sens résulte d'un consensus mettant en jeu, au premier plan, l'activité de mise en place d'une référence commune.

La référence, au sens utilisé dans le courant pragmatique et interactionniste, est le rapport que crée un interlocuteur entre les choses et les signes dans l'instant présent. Ainsi, elle est la relation qui unit une expression de la langue (dite en général *expression référentielle*) en emploi dans un énoncé et l'objet, l'événement ou le processus que cette expression évoque. Elle est ce grâce à quoi la langue permet de dire et d'entendre des choses sur le monde dans toutes ses dimensions : réelle, imaginaire et symbolique. Cette référence n'est pas donnée d'emblée ni par l'énoncé, ni par le contexte, mais c'est l'auditeur qui la construit dans et par la compréhension des énoncés. Ainsi, lors d'une interaction langagière, des tentatives de référence sont construites par celui qui écoute et interprète, mais aussi par le locuteur lorsqu'il entend ses propres paroles. Ces tentatives de référence s'expriment par une interprétation en acte qui est soumises à l'accord ou au refus de l'autre partenaire dans le dialogue (ou des autres partenaires, s'ils sont plus que deux). La référence est alors un processus, et ce processus échappe à la responsabilité de chaque sujet. Elle est le fait d'une co-adaptation des sujets et plutôt que de référence, on parlera dès lors de co-référenciation.

La co-référenciation est souvent ce qui permet d'initier une interaction langagière, car la présomption de différences de points de vue entre les différents interlocuteurs les pousse à argumenter, à défendre leur position, donc à verbaliser et ce jusqu'à ce qu'un accord puisse être

trouvé. Quand un accord n'émerge pas d'un dialogue, on peut souvent remarquer que celui-ci se "ferme" dans le sens où personne n'a plus rien à dire à l'autre étant donné que tout espoir de co-construction est perdu. C'est par exemple, les cas de dialogues qui se terminent sur un argument d'autorité de la part de l'un des acteurs. Le but central du processus de co-référenciation est donc de créer un consensus entre les différents acteurs sur la nature ou la désignation des "objets"[12] du monde qu'ils évoquent. Les cas de malentendus sur une référence supposée commune de la part de chaque sujet ne sont en aucun cas des marques d'échec de ce processus de co-référenciation. Ils montrent simplement que le consensus n'est pas encore atteint et lorsque un hiatus dû à ce malentendu se produit explicitement, c'est le signe d'une communication réussie ([Vivier & al. 94], [Brassac & al. 96]).

Le corpus PIC (que nous utilisons pour construire notre modèle informatique et le valider) est particulièrement riche pour l'étude de la co-référenciation. D'une part, on a à faire, non pas à une situation classique de dialogue, mais à un dialogue à trois. D'autre part, l'expérimentation met en place, par la présence du logiciel en fonctionnement, une situation de cognition située où le contexte apporte d'autant plus d'objets pouvant donner lieu à d'éventuelles références communes. Examinons, par exemple, un dialogue extrait du corpus du projet PIC mettant en jeu le rédacteur (R) et le développeur (D):

R1 : *... et les listes servent simplement en consultation*
D1 : *les listes heu*
R2 : *heu si on pouvait revenir à c't écran là*
D2 : *mhh*
R3 : *ou même un des autres*
D3 : *oui les grilles oui les grilles*
R4 : *voilà pardon les les*
. . .
Di : *... et la grille c'est juste une liste en fait ...*

Le rédacteur veut ici attirer l'attention du développeur sur un ensemble de composants de l'interface du logiciel qu'il évoque en R1 par le syntagme nominal *les listes.* En D1, le développeur signale au rédacteur par la reprise de sa dernière dénomination suivie d'une marque d'hésitation, qu'il n'y a pas de référence commune entre eux deux sur le syntagme *les listes* et ceci pousse le rédacteur à tenter d'expliciter ce dont il veut parler. En D3, le développeur fait signe au rédacteur qu'il vient de prendre conscience de l'ensemble d'objets dont il était question mais que ces objets, de son point de vue, correspondent à des grilles et non à des listes. En R4, le rédacteur signale qu'il est disposé à revenir sur sa dénomination initiale. Quelques secondes

[12] "objets" est ici utilisé dans un sens constructiviste comme dans [Mugur-Schächter 89] et désigne le résultat d'une opération de découpage et de catégorisation.

plus tard dans le dialogue, le développeur signale que grille et liste peuvent effectivement, dans le contexte présent, être compris comme des signes qui dénotent un même ensemble d'objets.

Cet extrait de dialogue montre que le consensus que constitue la co-référenciation est double (nous y reviendrons un peu plus loin). Il y a bien consensus sur l'objet à considérer, étant donné qu'ils sont tous deux en accord sur le même objet à la fin de la séquence alors que ce n'était pas le cas au début, mais il y a de plus un consensus sur une ou plusieurs façons d'évoquer cet objet. C'est de cette manière que la parole construit le monde. L'objet évoqué prend alors une place particulière dans l'intersubjectivité des interlocuteurs, place qu'il n'avait pas auparavant bien qu'étant présent dans le contexte. Il entre dans le terrain commun de l'interaction langagière.

La notion de terrain commun a été introduite par Clark ([Clark & al. 86]) pour rendre compte de l'identification des référents dans les interactions langagières. Selon Clark, deux interlocuteurs, qu'on appellera A et B, doivent partager un ensemble de connaissances (ou de croyances) communes pour qu'une inter-compréhension soit possible mais, de plus, il est nécessaire que A pense que B possède ces connaissances, et pense en outre que B pense que lui-même les possède; et réciproquement. On retrouve ici la notion d'empathie et les principes de base du modèle du transfert de Coursil ([Coursil 93], [Delépine & al. 95]). Le terrain commun (*common ground* chez Clark) regroupe donc les connaissances, croyances et suppositions mutuelles des interlocuteurs au moment de l'interaction. Les références communes prennent forme de connaissances dans ce terrain commun, et c'est sur la base de ce terrain commun qu'un interprétant va comprendre (ou pas) les énoncés d'un locuteur et que le locuteur va choisir des dénominations pour de nouveaux objets afin que son auditeur puisse les identifier le plus facilement possible pour ainsi créer de nouvelles références communes. Dans une perspective de réalisation d'un agent logiciel dialoguant en langue naturelle, le terrain commun et son évolution au cours de l'interaction apparaissent donc comme des éléments centraux à modéliser. Nous proposerons dans cette thèse avec le modèle *Anadia* (cf. chapitre 3) une méthode interactive de construction d'un terrain commun entre une machine et un partenaire humain.

Le consensus au centre de la co-référenciation est rendu possible par une fonction essentielle du langage : la **monstration**. D'un point de vue développemental, en analysant des interactions entre jeunes enfants et adultes, Bruner ([Bruner 83]) souligne le rapport de dépendance du processus de dénomination à l'égard de celui de désignation (il lui donne le nom de monstration) dans et par les interactions non verbales dans un premier temps, puis verbales dans un second temps. Pour notre part, nous nous limiterons à une définition qui fait de la monstration cet aspect spécifique des énoncés langagiers qui, avant d'apporter des informations, permet de rendre présents des objets, des situations, des procès, qu'ils soient réels ou

imaginaires. La monstration est ce qu'évoque [Milner 89] dans l'idée que le langage a pour fonction principale une "fonction désignative". C'est d'ailleurs une différence fondamentale entre la langue et la logique qui n'a qu'une effectivité informative et prédicative et donc pour laquelle la référence aux objets et aux procès évoqués est un préalable[13]. Du point de vue de la sémantique de la temporalité en français contemporain, [Gosselin 97] met en évidence que cette monstration marque la frontière entre le nécessaire et le possible, car le propre de la représentation langagière est de conférer à un moment quelconque, pris comme moment de référence (objet de la monstration), la propriété modale essentielle du présent, à savoir faire la séparation entre le passé (nécessaire car irréversible) et le futur (possible). Pour l'analyse sémantique des processus de co-référenciation, la monstration marque également un moment séparateur au sein de l'interaction. Lors de la mise en place d'une référence commune, il est possible de dégager deux processus, l'un préalable à la monstration, l'autre lui faisant suite. Le premier est une négociation sur la nature de l'objet qui permet de l'appréhender comme l'élément d'une catégorie. C'est donc principalement une négociation sur son identité catégorielle ([Beust & al. 97a]). Le second processus est aussi une négociation dont le but est, cette fois, de se mettre d'accord sur une dénomination de l'objet qui vient d'être montré. C'est à l'issue de ce processus qu'est réellement mise en place une référence commune. La monstration est donc bien distincte de la référence, car elle la précède dans certains cas, ou coïncide avec elle dans les énoncés déictiques et les rappels au terrain commun de l'interaction. L'exemple n°1[14] décrit un cas de négociation pré-monstrative. Les actants, pour mettre en place une référence commune sur la notion de *centre*, négocient les critères qui permettront une monstration partagée. Cette négociation passe ici par une reformulation du concept visé, notamment on passe de *centre* à *centre d'intérêt*. C'est en précisant ensemble les propriétés de la catégorie de l'objet visé que les actants "isolent" l'objet et permettent ainsi la monstration. On assiste alors à une co-précision de l'identité catégorielle de l'objet.

[13] Considérons par exemple l'énoncé *La chemise de Paul est bleue*. Cet énoncé est formalisé en logique par l'expression suivante: *soit c tel que chemise(c), Paul(c) -> bleue(c).* La partie *soit c tel que chemise(c)* est necessaire à la cohérence logique de l'expression entière et son but est bien de référencer un objet et de le considérer comme un présupposé avant de pouvoir le prédiquer. La mise en forme logique d'énoncés langagiers implique effectivement un ou plusieurs présupposés explicites qui ne sont pas nécessaires dans la forme linguistique.

[14] Dans cet exemple, ainsi que dans l'exemple n°2 suivant, les extraits de corpus sont présentés à la façon d'une partition de musique où les trois lignes D, U et R représentent respectivement les dires du développeur, de l'utilisateur et du rédacteur. Le chevauchement vertical des lignes indique que les acteurs parlent en même temps.

D :	qui appartiennent aux centres ensuite y a des enseignants qui enseignent donc y a enseignant heu
U :	
R :	vous avez un exemple ss de centre par (...) parceque

D :	mhh ben j'sais pas si y en là dedans on va regarder j'sais pas si y a les informations qui
U :	oui (...)
R :	je département c'est heu GEA

D :	correspondent hin voila y a pas de centre d'intérêt mais on peut en créer un si vous voulez (...) vas y
U :	ah oui mais y a cinq centres en fait
R :	

D :	mais je crois que ça va pas marcher ouais
U :	
R :	là là typiquement cette notion de centre c'est une notion qui est commune heu à tout l'IUT

Exemple n°1 : Négociation pré-monstrative

À l'inverse, l'exemple n°2 constitue une négociation post-monstrative dont l'issue donne lieu à une référence commune. Ici, les actants ont bien pris acte de l'objet visé par la monstration et aucun d'eux ne semble avoir de doute sur le fait que ce qu'il considère comme objet du processus de co-référenciation est une entité commune. Il ne s'agit plus ici de préciser des critères sortaux mais de trouver une dénomination commune de l'objet montré pour faciliter la suite de l'interaction. Ce processus post-monstratif est plus simple et moins long que le processus pré-monstratif, mais il nécessite au moins trois tours de parole. Au premier tour, l'un des actants propose une dénomination (en l'occurrence *GAE*). Au second tour, un de ses partenaires valide cette dénomination ou en propose une autre (ici *GEA*). Dans le troisième tour, le premier sujet fait signe au second qu'il valide sa solution, ce qui permet à tous de poursuivre l'interaction.

D :	
U :	au département des entreprises et des administrations GEA mhh
R :	GAE GEA

Exemple n°2 : Négociation post-monstrative

Les langues naturelles disposent d'un ensemble de moyens par lesquels un énoncé renvoie, non pas uniquement à un contenu, mais également à l'activité par laquelle il a été produit, l'acte d'énonciation. C'est notamment pour cela qu'il n'y a de pragmatique que pour les langues naturelles et pas pour les langages formels. La deixis (qui provient d'un mot grec signifiant "acte de montrer") fait partie de ces moyens d'évocation de l'énonciation. La fonction d'une marque déictique est de désigner directement la situation d'énonciation, d'interaction, et elle n'est interprétable que dans un rapport à cette situation. C'est une référence à un objet (pris au sens large) présent dans le contexte. Cette présence n'est pas forcément matérielle (plutôt symbolique) car on peut faire, par exemple, référence par un déictique (maintenant, bientôt,

l'année prochaine, ...) à un moment dans le temps qui est déterminé par rapport au moment présent. C'est le cas du mot *Hier* dans l'énoncé suivant :

Hier, je suis allé à la piscine.

On peut aussi évoquer de façon déictique les partenaires de l'interaction par l'usage de pronoms personnels de la première ou de la deuxième personne (c'est le cas de *je* dans l'exemple précédent). À ce repérage déictique des partenaires peuvent s'ajouter des indications sur les statuts mutuels des sujets ; c'est notamment le cas des formules de politesse dont le nombre et la subtilité varient en fonction des langues (c'est par exemple le cas de la distinction tu/vous en français, qu'on ne retrouve pas en anglais ou qui est inversée en arabe). Enfin, on peut faire référence par la deixis au lieu de l'énonciation et dans ce cas, on peut souvent observer que la place du locuteur sert de point de référence dans l'espace. C'est par exemple le cas du déictique *ici* qui, selon [Pierrel & al. 97, p. 222], induit un pavage hiérarchique de l'espace. Dans des situations de dialogues, ces références déictiques sont souvent accompagnées de gestes accomplissant physiquement un acte de monstration. C'est l'exemple du mot *ça* dans :

Passe moi ça, s'il te plaît.

Dans la situation de conception distribuée du projet PIC, la présence du logiciel permet un usage important de déictiques. Elle facilite la conception du Manuel de l'Utilisateur car le rédacteur peut examiner directement les difficultés qu'éprouve l'utilisateur face au logiciel ou questionner immédiatement le développeur sur tel ou tel aspect de l'application. Le logiciel peut donc être considéré comme une terrain d'expérience pour les trois participants et c'est donc bien de cognition située dont il s'agit. Les choses que les sujets évoquent sont souvent présentes physiquement, matérialisées par le logiciel. Ceci a pour conséquence de soulager les interlocuteurs d'un bon nombre d'échanges pré-monstratifs car les objets de référence potentiels ne sont pas à expliquer mais simplement à montrer physiquement dans un espace partagé. On ne retrouve pas cette situation dans des expériences comme celle décrite dans [Clark & al. 86] dans laquelle les auteurs mettent en place une situation où deux sujets doivent classer des figures de Tangram en étant tout deux séparés par un écran. Dans cette situation, les sujets disposant chacun du même jeu de figures sont contraints de verbaliser des catégories et donc d'utiliser exclusivement la monstration langagière, puisque le pointage gestuel est rendu impossible par l'écran. Dans notre situation, le fait de disposer d'un dispositif de pointage, comme la souris, implique que beaucoup de négociations verbales sur l'identité des objets ne sont plus nécessaires du fait d'une monstration déjà réalisée (par l'intermédiaire du pointeur de la souris). La co-référence aux objets de l'écran s'effectue essentiellement de façon déictique comme le montrent les extraits du corpus suivants :

... Je ne sais pas si y'en a une ***là*** *…*
... donc ***là*** *c'est SUGEA ...*

*... donc **là** on peut faire OK ...*
*... donc **là** tu vas sur le menu ...*
*... i'faut cocher cliquer **là** …*
*... si on pouvait revenir sur **c't écran là** …*
*… **là ici** on va sur recherche ...*
*... la p'tite flèche non **là voilà** …*
*... c'est l'option **là vas-y** sélectionne **voilà** …*
*... j'vais aller chercher **la d'dans** …*
*... le p'tit truc **ici** …*
*... clique sur **celui là** puisque c'est **celui là** …*

Cette utilisation importante de marqueurs déictiques est directement liée à la situation de cognition située qui donne un rôle particulier au quatrième partenaire de l'interaction : la machine ([Beust & al. 97c], [Beust & al. 98]).

Quand il est fait usage de déictiques, la co-référenciation est difficilement observable à partir de l'analyse linguistique. Étant donné que la simple prise en compte des dires des sujets ne suffit plus à rendre compte de l'effectivité de l'interaction, il convient alors d'observer les bandes vidéo pour trouver les indices pertinents. Étudier dans une telle situation les procédures d'établissement d'un terrain commun demande une prise en compte de l'ensemble du contexte (un lieu, un temps, des personnes présentes, leur rôles, ...) ainsi qu'une analyse comparative précise de ce qui est dit par rapport à ce qui est affiché à l'écran de l'ordinateur. Les déictiques sont donc les indices d'une situation plus facile en ce qui concerne les interlocuteurs mais plus délicate en ce qui concerne une analyse exhaustive de l'interaction. Cette analyse exhaustive est un objet d'étude pour des recherches pluridisciplinaires couvrant l'ensemble des sciences du langage. À l'inverse, quand c'est la fonction de monstration langagière de l'énoncé qui permet d'attirer l'attention du partenaire sur telle ou telle chose, le processus est alors complexe mais il donne lieu à bon nombre de productions verbales. C'est un avantage pour notre analyse du corpus. En effet, les reformulations fréquentes et les références déjà établies dans le terrain commun fournissent alors un matériau linguistique riche. Des effets de sens y sont linguistiquement observables.

La démarche suivie dans cette thèse consiste à analyser la sémantique des énoncés produits au cours des interactions, et en particulier en ce qui concerne les processus de co-référentiation et de construction du terrain commun. Nous laissons ainsi l'analyse de la deixis à une modélisation pragmatique des conditions d'énonciation. L'objectif de cette démarche analytique est de mieux comprendre comment la formation du sens se fait au travers des négociations entre les interlocuteurs dans l'interaction langagière. Éclairé par cette observation, il s'agira de contribuer à un modèle computationnel du sens des énoncés qui puisse s'inscrire

dans la modélisation de la fonction de monstration des énoncés et dans la réalisation d'une compétence dialogique.

On trouve un exemple de modélisation de la compétence dialogique dans le modèle de dialogue homme-machine proposé dans [Lehuen 97]. Le modèle de Lehuen met l'accent sur la logique interne du dialogue dans un rapport à une tâche, notamment par la gestion des incomplétudes par les dialogues incidents, mais l'analyse sémantique des énoncés y est approximée par un jeu de filtres syntaxico-sémantiques (i.e. des phrases à trous). Cette approximation pose certains problèmes qui affectent la gestion du dialogue et de la tâche. Considérons par exemple, dans le cadre d'une consultation de bibliothèque, un système de dialogue qui dispose pour la compréhension des énoncés du filtre suivant (c'est le cas du système *Coala* de Lehuen) :

Je cherche <TITRE> *de* <AUTEUR>

Le système n'est alors pas en mesure de faire la différence entre les deux énoncés qui suivent :

Je cherche Mort sur le Nil d'Agatha Christie

Je cherche à savoir les modalités d'inscription à la bibliothèque

Ceci le conduit à rechercher, en réponse au deuxième énoncé, un ouvrage dont le titre serait *"à savoir les modalités"* et l'auteur *"inscription à la bibliothèque"*. En disposant d'un modèle de l'interprétation plus fin, indiquant qu'il n'est pas question d'une recherche d'ouvrage dans le deuxième énoncé, ce problème pourrait être évité. L'amélioration de l'interprétation pour résoudre ce genre de problèmes est un objectif que nous nous donnons. Il intègre notre travail dans le cadre du dialogue homme-machine. Après avoir décrit le modèle computationnel de l'interprétation proposé, nous reviendrons sur ce problème en expliquant les apports du modèle pour la réalisation d'agents logiciels dialoguant en langue naturelle (cf. chapitre 6).

3. Principes du modèle proposé

Cette partie décrit les principes fondamentaux retenus pour proposer un modèle de la sémantique des énoncés produits au cours d'un dialogue. Dans un premier temps, nous nous placerons dans le cadre d'une approche interprétative, en faisant la distinction entre l'interprétation et la compréhension. Dans un second temps, nous proposerons un modèle du sens des énoncés langagiers qui prend en compte cette distinction. Enfin, étant donné que notre démarche analytique est centrée sur la co-référenciation dans les dialogues, nous préciserons la place de la référence dans ce modèle.

3.1. Une approche interprétative

Une approche interprétative consiste à se focaliser sur l'allocutaire dans son activité d'interprétation des paroles des autres, mais aussi des siennes lorsqu'il les entend. À l'instar de [Coursil 92], on s'intéresse donc avant tout à l'activité langagière silencieuse. De ce point de vue, l'activité de production langagière est seconde par rapport à l'interprétation, contrairement aux théories génératives de Chomsky ([Chomsky 69]).

Cette approche est à l'opposé des théories intentionnelles du sens illustrées notamment par la théorie de la communication de Shannon et Weaver ([Shannon & al. 49]) et par la théorie de la cognition de Fodor ([Fodor 86]), où le sens d'un énoncé est vu comme le décodage d'une intention du locuteur, c'est-à-dire qu'il suppose l'existence d'une représentation non langagière du sens qui détermine la production verbale[15]. Ici, nous ne construirons pas le sens d'un énoncé à partir des éventuelles intentions de son locuteur. Le sens sera considéré comme étant le fait de l'interprétant, et construit à partir des observations *in situ* de ses réactions à l'énoncé. L'activité de l'interprétant est observable dans le dialogue par l'enchaînement conversationnel, à la différence de l'activité du locuteur qui ne l'est pas étant donné qu'on ne peut observer ses intentions (on peut juste observer, par les énoncés produits, les résultats de cette activité mais pas l'activité en elle-même).

L'approche interprétative amène à distinguer deux niveaux dans la compétence linguistique : la *compréhension* et l'*interprétation* (constituant alors l'objet d'étude de l'approche interprétative). La compréhension est une activité langagière typiquement humaine qui produit la conscience d'un résultat ou d'un effet pragmatique d'une séquence linguistique, et ce, sans forcément savoir comment on arrive à ce résultat ou cet effet ([Rastier & al. 94, p. 11]). Elle prend en compte l'empathie d'un sujet pour un autre sujet. C'est-à-dire que la compréhension suppose avoir saisi un sens mais aussi le pourquoi de ce sens dans la situation présente. C'est par exemple le cas d'une personne qui ferme une fenêtre en réponse à une autre qui vient de lui dire *il fait froid* (on dira bien ici que l'énoncé a été compris). La compréhension est principalement l'objet d'étude de la pragmatique et de la psychologie (plus précisément de la psycholinguistique) qui cherche expérimentalement, en analysant les conditions de production de tels résultats ou effets, à mettre au jour les processus qu'ils nécessitent. L'interprétation, quant à elle, n'est pas exclusivement une activité humaine. Bien sûr, on dira d'un pianiste qu'il interprète une partition lorsqu'il la joue mais l'on considère également, à juste titre, qu'une machine qui réalise une séquence d'actions interprète le code (la chaîne) attaché à cette séquence (on parle notamment de langages interprétés pour décrire les langages de programmation qui ne

[15] Cela suppose également que la langue est un code pour exprimer le sens. [Levrat 93, p. 15] rappelle les dangers de la métaphore LA LANGUE EST UN CODE qui n'est utile qu'en tant que moyen d'explication, mais qui ne peut servir de justification pour la construction d'un modèle scientifique.

compilent pas le code avant de l'exécuter). Dans notre problématique, il s'agira d'étudier l'interprétation des chaînes linguistiques. On conçoit alors l'interprétation comme un premier plan, non contextuel (donc sémantique), de l'activité de compréhension. Elle se limite à la mise en forme de la contribution du matériau linguistique de l'énoncé à son sens et ne vise pas à rendre compte des inférences pragmatiques possibles liées à la contextualisation de ce sens.

L'extraction de sens dans un modèle de l'interprétation en langue naturelle pose la question de la compositionnalité, c'est-à-dire du rapport entre le sens de l'énoncé et les significations de ces composants. La position qui consiste à construire le sens à partir d'un assemblage compositionnel des significations des constituants revient, une fois de plus, à calquer sur les langues naturelles un modèle atomiste des significations ainsi qu'un modèle de l'interprétation identique au processus de compilation dans les langages formels. Une approche interprétative fait le choix de la position inverse qui considère que le sens de l'énoncé est premier par rapport aux significations qu'il met en jeu. Ainsi, ce n'est que lorsque l'on a saisi le sens que l'on peut procéder à l'explication des significations des constituants de l'énoncé. Ceci est particulièrement évident en ce qui concerne les métaphores, et Lakoff et Johnson ([Lakoff & al. 85]), en montrant leur omniprésence dans les langues naturelles, ont conduit à voir autrement la question du sens. Par exemple, dans l'énoncé

L'inflation dévore tous nos profits. ([Lakoff & al. 85, p. 42])

on ne retient pas une signification possible du verbe *dévorer* que l'on pourrait paraphraser par *manger férocement* car c'est le sens de l'énoncé qui guide l'interprétant à retenir pour *dévorer,* un usage métaphorique. Pourtant on ne peut pas soutenir que les significations des mots et des structures syntaxiques ne participent pas à la construction du sens de l'énoncé. Pour s'en convaincre, il suffit de remplacer l'un des mots de l'énoncé précédent pour voir apparaître un sens nouveau (qui cette fois, n'est plus métaphorique en ce qui concerne *dévore* mais métonymique en ce qui concerne *profits*) :

Le lion dévore tous nos profits.

On se trouve alors, en apparence, devant un problème comparable à celui de *l'oeuf et de la poule,* à savoir qui, du sens ou des significations, détermine qui ? C'est ce problème qui nous amène, dans cette thèse, à proposer un modèle computationnel du sens guidé par une recherche des dépendances sémantiques (cf. chapitre 2) entre les lexèmes d'un énoncé.

3.2. Le sens comme un potentiel

Les analyses de conversations menées par le GRC (Groupe de Recherche sur les Conversation, Université de Nancy II) en psychologie sociale montrent que le sens est façonné

par les différents interlocuteurs au cours de l'interaction. Examinons par l'extrait de dialogue suivant (emprunté à Dessalles, école d'été de l'ARC, Bonas 1995) :

E1 : *oh, c'est marrant ! j'ai exactement ce tableau chez moi, il a la même taille, il représente la même chose*

L1 : *on te l'a peut être volé*

E2 : *non, le mien, il est plus sombre, il est plus beau*

Dans cet exemple de dialogue, [Brassac & al 96] analysent l'énoncé E1 comme étant porteur d'un **potentiel de sens**. On ne peut dire *a priori* s'il s'agit d'une affirmation, de l'expression d'un regret, d'une inquiétude ou encore d'une question. À l'issue d'une analyse de E1, on peut simplement dire que l'énoncé est susceptible de porter tout ceci. Dans les tours de parole suivants, l'énoncé L1 constitue une proposition d'actualisation de ce potentiel qui relance l'interaction, et le rapport de E2 à l'enchaînement (E1 -> L1) traduit alors une prise de position à deux par rapport à la proposition d'actualisation. Cette position reste négociable dans la suite de l'interaction et détermine l'enchaînement conversationnel. Il convient donc d'étudier le sens d'un énoncé dans le rapport de cet énoncé à ceux qui le précèdent et le suivent dans l'interaction, et non de façon isolée.

À travers cet exemple, les auteurs montrent aussi que, dans les conversations, le locuteur n'est pas "propriétaire" du sens de ses énoncés étant donné que le sens est négocié dans la suite de la conversation (il en est de même dans un texte car le sens qu'actualise le lecteur échappe à la volonté de l'auteur). C'est un argument supplémentaire pour une approche interprétative et non intentionnelle du sens, car selon les auteurs la seule intention que nous pourrions éventuellement attribuer au locuteur est celle d'avoir voulu communiquer quelque chose.

En nous inspirant de ces observations et de ces analyses, nous allons également concevoir le sens à la manière d'un potentiel au centre de l'interaction que les interlocuteurs actualisent conjointement. Cependant, à la différence des psychologues, nous allons proposer un modèle dans lequel le potentiel de sens est exprimable et calculable d'un point de vue informatique.

Considérer ainsi le sens change la façon traditionnelle de le voir en traitement automatique. Il n'est plus l'objectif opératoire de l'analyse d'un énoncé, mais d'une séquence d'énoncés. On ne cherche donc pas à extraire et à représenter en soi le sens d'un énoncé, mais à calculer son potentiel de sens, c'est-à-dire sa participation au sens de la séquence dont il fait partie. Cette participation est ce que l'énoncé prescrit pour l'enchaînement conversationnel. C'est un jeu de contraintes sémantiques. Le sens d'un énoncé en situation est l'actualisation pragmatique de ce jeu de contraintes ; les choix d'actualisation de ce potentiel participent à la compréhension de l'énoncé et ne sont analysables qu'à un niveau pragmatique. Les différents sens possibles d'un

énoncé sont envisagés comme des actualisations dans différents contextes de son potentiel de sens, et aucune de ces actualisations n'est plus canonique qu'une autre.

Pour chaque énoncé, la mise en place de son potentiel de sens par une recherche des contraintes sémantiques qu'il prescrit, constitue son interprétation. Nous verrons dans le chapitre suivant une approche similaire, celle de Gosselin ([Gosselin 96a]), où le sens est également conditionné à un jeu de contraintes[16]. Avec le modèle que nous proposons dans cette thèse, nous allons monter comment une machine peut mettre en forme un jeu de contraintes sémantiques à partir des énoncés d'un dialogue.

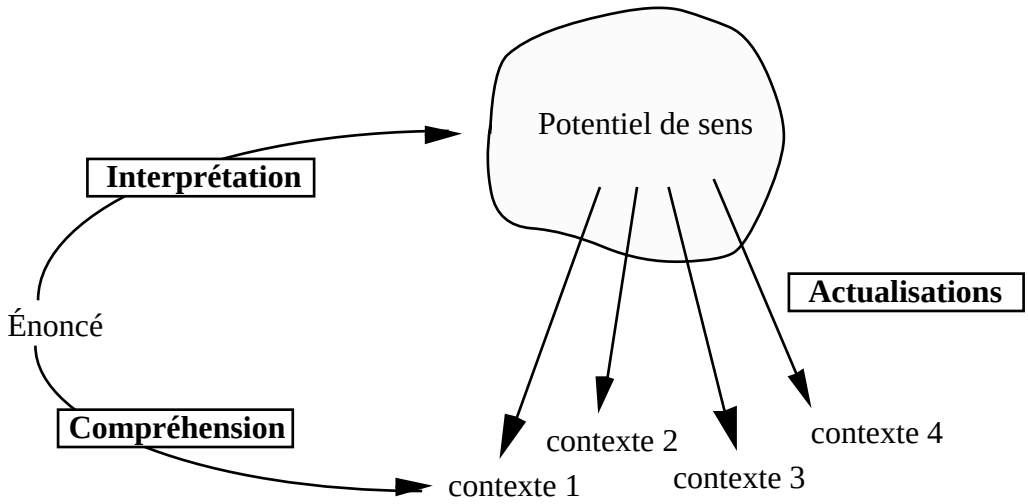


Figure 1.2 : Le sens comme un potentiel.

Considérons, par exemple, un énoncé extrait du corpus PIC. Cet énoncé est produit, au début de la séance de travail, au moment où la future utilisatrice du logiciel se présente au rédacteur technique :

à l'IUT alors je suis secrétaire (Corpus PIC, p. 2)

Cet énoncé met en place des contraintes sur le sens que l'on peut y attribuer. Ainsi, la personne qui l'a produit parle d'elle car le discours est direct et qu'il y est fait usage du déictique *je.* De même, elle exprime un état (par le verbe *être*) et elle signale à ses partenaires qu'elle a, par cet état, la qualité dénotée par le qualificatif *secrétaire* qui correspond à une fonction. Elle exprime par l'usage de la préposition *à*, que cette fonction est en rapport à une structure ou un lieu appelé *IUT*. Il est donc vraisemblablement question de sa profession, marquée par les emplois dans le

[16] Gosselin propose un modèle de la temporalité en français contemporain pour exprimer, à partir de textes, des contraintes sémantiques sur les positions relatives d'intervalles temporels auxquels le texte fait référence (ce modèle linguistique a fait l'objet d'une implémentation informatique sous la forme d'un système expert).

même énoncé du syntagme *l'IUT* et du qualificatif *secrétaire.* Enfin, elle signale par le connecteur *alors* une orientation argumentative, probablement une relation d'implication (à moins qu'une utilisation fréquente de *alors* soit une habitude phatique de l'utilisatrice, ce qui n'est pas flagrant dans le corpus). C'est finalement tout ce que l'énoncé véhicule, et pourtant on pourrait le comprendre de diverses façons, comme le montrent les paraphrases suivantes :

(1) *à l'IUT, elle est, entre autre, secrétaire, mais elle y a aussi d'autres fonctions*
(2) *à l'IUT, elle est secrétaire, mais il n'y a bien qu'à l'IUT, car ailleurs, elle a d'autres fonctions*
(3) *à l'IUT aussi, elle est secrétaire*
(4) *à l'IUT, elle est secrétaire pour l'instant mais son statut va changer*
…

Ces paraphrases sont principalement construites en déclinant les orientations argumentatives dues au connecteur *alors* qui n'est pas sémantiquement précisé dans l'énoncé (il n'est pas aussi contraint sémantiquement que les autres composantes de l'énoncé, d'où un certain degré de liberté quant à sa signification présente). Ainsi, en interprétant *alors* dans une argumentation sur le lieu de travail on obtient la paraphrase (2) et en l'interprétant dans une argumentation sur le temps, on obtient la paraphrase (4). Cette liste de paraphrases possibles n'est évidemment pas exhaustive et l'on pourrait encore trouver un bon nombre de sens possibles à l'énoncé initial. Ces sens possibles sont révélateurs d'une compréhension de l'énoncé dans une situation d'énonciation (dans son contexte) et en accord avec les contraintes sémantiques issues de son interprétation.

Les contraintes sémantiques peuvent être de plusieurs types :

• Elles peuvent être des contraintes de monstration quand il est question de critères définitoires d'un ou plusieurs possibles objets, événements ou processus. Considérons par exemple l'énoncé *J'ai acheté une table en bois.* Dans le cadre d'une analyse sémantique, l'énoncé indique une contrainte qui signale qu'il est très vraisemblablement question d'un objet matériel. Nous verrons au chapitre suivant que l'on déduit cette contrainte d'un effet de sens du à la co-présence des mots *table* et *bois* dans le même groupe nominal qui renforcent chacun l'idée d'un objet matériel[17]. Cette contrainte (purement sémantique) ne constitue pas une référence de l'énoncé au monde. Simplement, elle indique des critères définitoires de possibles objets du monde sur lesquels portent les contenus sémantiques des constituants de l'énoncé. En ce sens, c'est une contrainte de

[17] Si à la place de *table en bois*, on avait *chèque en bois*, l'effet de sens serait différent. On ne conclurait pas à un rapport à la matérialité mais à une expression métaphorique.

monstration (au même titre que le déictique *Je* dans un discours direct donne lieu à une contrainte de monstration indiquant que le locuteur parle de lui même).

• Elles peuvent être des contraintes de prédication quand elles rendent compte d'un état ou d'un procès évoqué par l'énoncé ou d'une relation entre plusieurs actants. Par exemple, le verbe *donner* induit une contrainte de prédication qui met en relation trois entités (il organise une scène selon [Victorri & al. 96]) : un sujet (celui qui donne), un objet (ce qui est donné), et un troisième actant (celui à qui le sujet donne).

• Elles peuvent être également des contraintes de modalité quand l'énoncé fait intervenir des verbes modaux (tels que *vouloir, pouvoir, souhaiter, croire*) ou des adverbes tels que *sincèrement, franchement,* ou des tournures insistant sur ou atténuant ce qui est dit.

• Elles peuvent être enfin des contraintes d'argumentation quand elles expriment le ou les schémas argumentatifs associés à un énoncé ou liant cet énoncé par rapport aux énoncés précédents dans l'interaction. Par exemple, le connecteur *donc* entre deux propositions permet de former une contrainte indiquant une relation de cause à effet entre les deux propositions en question. De même, un énoncé commençant par *mais* indique a priori une position inverse ou complémentaire dans le schéma argumentatif initié par les énoncés précédents.

En déterminant les contraintes sémantiques attachées à un énoncé, nous nous focaliserons plus particulièrement sur les contraintes de monstration car ce sont principalement celles qui sont pertinentes dans une analyse sémantique des processus de co-référentiation. De plus, elles conditionnent les autres types de contraintes. En effet, pour rendre compte d'une prédication, il faut mettre en évidence les actants de cette prédication, et ceci relève de la monstration des actants[18]. De même, mettre au jour une argumentation ou une modalité, implique de rendre compte des contraintes de prédication. En quelque sorte, en faisant un parallèle avec les catégories de la phénoménologie de Peirce (Priméité, Secondéité, Tiercéité), la monstration est un premier, la prédication est un second, et l'argumentation ainsi que la modalité sont des troisièmes.

Les contraintes sémantiques d'un énoncé déterminent la valeur locutoire rhétique de l'acte de langage dont l'énoncé est la trace linguistique. Elles contribuent également à la force illocutoire de l'acte. La force illocutoire dépend de la direction d'ajustement induite par l'acte de langage ([Searle 72]). Partant des travaux de Searle et d'un modèle sémiotique du monde,

[18] Ceci ne veut pas dire pour autant que la prédication n'intervienne pas dans l'établissement de la monstration étant donné que, comme nous le verrons par l'étude des relations actancielles dans le chapitre 2, la relation entre un verbe et ses actants peut permettre de lever des ambiguïtés sémantiques.

[Nicolle & al. 98] définissent seulement quatre types de force illocutoire : la force assertive (les mots doivent s'ajuster au monde), la force déclarative (du fait de l'énonciation, les mots et le monde s'ajustent), la force directive (le monde doit s'ajuster aux mots par l'action de l'allocutaire), et la force commissive (le monde doit s'ajuster aux mots par l'action du locuteur). [Nicolle & al. 98] montrent, avec le modèle de contraintes sémantiques proposé ici, la contribution des personnes verbales, des modes et temps des verbes à la détermination de la force illocutoire.

La métaphore du sens en tant que potentiel permet de préciser le "cahier des charges" de notre modélisation informatique dans l'approche interprétative :

• Le sens d'un énoncé ne se réduit pas à une de ses actualisations. Il ne se réduit pas non plus à l'ensemble de ses actualisations car celles-ci ne sont ni énumérables, ni toutes connues. Le sens n'est pas non plus mesurable[19], c'est un champ de jugement ([Coursil 92]) dans lequel une actualisation est une prise de position de l'interprétant.

• L'analyse sémantique computationnelle doit exprimer, à partir de l'énoncé, l'ensemble des contraintes qu'il prescrit. Dans les chapitres suivants, nous mettrons en évidence certaines de ces contraintes par la recherche des *effets de sens co-textuels* portés par l'énoncé. Ces effets de sens co-textuels sont le fait de la co-présence des mots de l'énoncé (comme par exemple la co-présence de *table* et *bois* dans *une table en bois*) qui précisent ensemble leur signification en contexte.

• Les contraintes sémantiques n'entretiennent pas de rapport avec le contexte d'énonciation de l'énoncé, qui n'est pertinent que dans le choix de l'actualisation du sens. De ce fait, le potentiel de sens (i.e. les contraintes sémantiques) n'est pas référentiel, il est déterminé en langue (i.e. il n'est pas extralinguistique). Sa recherche opératoire ne nécessite pas une modélisation ontologique du monde. C'est pour cette raison que l'on peut en faire un modèle informatique par amorce sans préalablement avoir fourni à la machine une base de connaissances exhaustive, ce qui n'est pas réalisable comme l'a montré l'échec du projet CYC de Lenat ([Lenat & al. 90]).

3.3. Sens et référence

On ne peut soutenir que le sens et la référence n'entretiennent pas de relations fortes car, à l'évidence, la langue permet d'évoquer le monde et les objets qu'il contient. Ainsi, faire la ou les bonnes références prescrites dans un énoncé (ou une phrase) est une preuve de la bonne

[19] Pour prolonger la métaphore, un potentiel n'est jamais mesurable. Ce que l'on peut mesurer, c'est une différence de potentiel (par exemple, l'électricité).

compréhension de celui-ci. En considérant le sens comme un potentiel, nous sommes amenés à préciser ses rapports avec la référence.

Comme on le constate dans l'analyse des processus de co-référentiation dans les dialogues, la référence n'est pas un simple rapport entre la langue et le monde, mais plutôt une sorte de moteur de l'interaction langagière ; ce que Victorri et Fuchs évoquent par les mots suivants :

> Disons, simplement pour l'instant que, pour nous, la référence ainsi conçue n'est pas une simple mise en correspondance de la parole avec les objets et les événements "du monde", mais une véritable opération de construction d'un monde, créé dans et par le discours, et dont les relations avec la réalité peuvent être plus ou moins complexes. ([Victorri & al. 96, p. 26])

Cette phrase met en garde le lecteur contre une simplicité apparente de la référence (comme fonction entre des choses pré-existantes) et insiste sur sa nature procédurale. Elle signale également un aspect fondamental du langage, qu'il convient de rappeler lorsqu'on s'intéresse aux problèmes de référence : l'activité langagière ne met pas en place, via la référence, un rapport de la langue à un réel préexistant mais un rapport de la langue au monde, qu'elle construit en le décrivant ou en permettant de l'inventer.

La référence est une relation très générale, qui, dans l'effectivité du langage, se réalise sous plusieurs aspects : on peut avec des mots faire référence à des objets présents ou non, à des situations, à des événements passés, présents ou futurs, à des idées, à des états mentaux et même à des expressions langagières. Dans les exemples suivants, la phrase fait référence à elle même (elle est dite auto-référencielle) :

Cette phrase contient cinq mots.

Cette phrase pas de verbe. ([Hofstadter 88, p. 6])

Mais en utilisant la langue, on peut aussi faire référence à des purs produits de l'imaginaire, c'est-à-dire à des choses qui n'ont rien à voir avec le réel mais qui n'en sont pas moins des objets du monde dans la mesure où ils sont partagés par plusieurs individus. L'exemple de la licorne est prototypique, car tout le monde sait qu'aucune licorne n'a jamais réellement existé et pourtant chacun peut s'en faire une représentation parce que c'est un objet de notre culture, c'est-à-dire de notre monde. De même, pour reprendre un exemple de Ducrot ([Ducrot & al. 95, p. 302]), l'île au trésor (objet imaginaire) est un objet de référence possible, autant que la gare de Lyon (objet réel). On pointe ici la propriété de la langue qui est de pouvoir évoquer des choses qui ne sont pas (ou même qui ne peuvent pas être), c'est-à-dire le fait que la langue puisse servir à raconter des histoires sans imposer un rapport direct à la réalité.

Pour reprendre l'exemple de la licorne, ce n'est pas parce que la référence n'est pas réelle, qu'il n'y a pas de référence du tout. Cela ne signifie pas non plus que, lorsque la référence n'est pas réelle, elle se confonde avec la signification du signe. Il y a une distinction entre le domaine des signifiés, qui construisent les contraintes sémantiques, et le domaine des référents des signes, construits par ceux qui écoutent. Signifié et référents sont deux notions qui ne se recouvrent pas (cf. [Ducrot & al. 95]). Ceci nous place au coeur d'une distinction saussurienne : le signe unit non une chose (référent) et un nom (signifiant), mais un concept (le signifié) et une image acoustique (le signifiant). C'est-à-dire que le signifié de *cheval* n'est pas défini par les propriétés ontologiques des chevaux en général mais représente tout ce que "cheval" n'est pas. Le signifié, contrairement à la notion de concept ontologique, n'entretient aucun lien avec une classification des espèces animales. Il est défini de façon purement différentielle, c'est-à-dire, non pas positivement par son contenu mais négativement par ses rapports avec les autres signes du système. Ainsi, les signes prennent la forme de valeurs pures :

> "Leur plus exacte caractéristique est d'être ce que les autres ne sont pas." ([Saussure 78, p. 162])

Le signifié d'un signe ne regroupe donc en aucun cas les traits qui proviennent d'une description, même partielle, des objets qu'il peut désigner .

Pour préciser les rapports entre le sens et la référence, [Victorri & al. 96] introduisent une distinction sur la nature même de l'énoncé. Il s'agit de la différence entre *énoncé-type* et *énoncé-occurrence*. Ainsi, l'énoncé-type (i.e. la phrase) est la suite ordonnée des marques linguistiques de l'énoncé-occurence. L'enoncé-occurrence est, quant à lui, un événement situé dans une interaction entre sujets dont les effets sont observables en termes psychologiques ou pragmatico-référentiels. L'exemple donné est celui de l'énoncé *Il fait froid* qui représente le même énoncé-type mais plusieurs énoncés-occurrence en fonction des conditions de son énonciation. Ainsi *Il fait froid* en réponse à *Tu m'accompagnes au marché ?* évoque la température extérieure (peut-être de l'ordre de 0°) alors qu'énoncé au cours d'une soirée lorsqu'il est demandé de passer à table évoque la température de la pièce (probablement d'environ 15°) pour signaler l'intention de garder sa veste. Dans un autre contexte, cet énoncé pourrait être utilisé pour inviter l'interlocuteur à fermer une fenêtre ouverte. On a donc ici plusieurs sens bien différents, c'est-à-dire plusieurs énoncés-occurences du même énoncé-type. Quant au sens de l'énoncé-type, il est la contribution constante du matériau linguistique dont il est constitué au sens de toute occurrence de cet énoncé ([Victorri & al. 96, p. 27]). Rendre compte du sens de l'énoncé-type est donc posé comme un enjeu de la sémantique, alors qu'observer et rendre compte du sens d'un énoncé-occurence est posé comme un enjeu de la

pragmatique. La référence en tant que relation observable entre les mots et les choses est un effet de l'énoncé-occurrence et constitue de fait un objet d'étude pragmatique[20].

Au lieu d'envisager la pragmatique comme le dernier niveau linguistique[21], nous considérons, à la suite des travaux de Bricon-Souf ([Bricon-Souf 94]), la linguistique et la pragmatique comme deux plans d'analyse orthogonaux. S'il est vrai qu'un énoncé isolé de son contexte perd, d'une certaine façon, son sens, il n'en reste pas moins que les signes qui le composent ainsi que sa structure sont, du point de vue de l'interprétant, porteurs d'indications pour la mise en contexte du sens. De notre point de vue, la référence constitue une actualisation de ce potentiel. Rendre compte des références d'un énoncé ne relève donc pas de son interprétation. Ce n'est pas pour autant un problème uniquement pragmatique qu'une modélisation exhaustive du contexte d'énonciation suffirait à résoudre. La référence est à rechercher dans le rapport d'un monde à une monstration langagière (issue d'un interprétation) ou déictique.

Considérons, par exemple, le dialogue fictif suivant (extrait de [St Dizier De Almeida &. al. 98]).

A1 : *Bernard est dans le bureau de l'association*
B1 : *C'est pas possible, je l'ai vu ici il y a 5 minutes*
A2 : *Non, il est le trésorier*
B2 : *Ah, il appartient au bureau de l'association*

L'énoncé B1 traduit une actualisation du potentiel de sens de l'énoncé A1. Dans cette actualisation, la personne B signale un problème d'ordre spatial quant à la localisation d'un certain *Bernard,* c'est-à-dire que B a considéré dans son actualisation, *le bureau de l'association*

[20] En ce qui concerne la référence dans une modélisation informatique, Andrès insiste sur la dualité du processus de co-référentiation ([Andrès 95, p. 79]). D'une part, il s'agit pour chaque expression référentielle de rechercher les référents possibles dans le monde ; c'est la vision pragmatique de la référence. D'autre part, il s'agit de participer à une construction commune. Du point de vue interactionniste, le résultat de cette construction est un consensus ; c'est la référence commune issue du processus de co-référenciation.

[21] La pragmatique (en tant que branche de la linguistique) est une discipline très jeune si on la compare à l'ensemble de la linguistique. On peut en estimer le point départ dans des travaux de philosophie du langage de ce siècle et de façon plus précise à partir des conférences données en 1955 par Austin ([Austin 70]) et en 1967 par Grice ([Grice 67]). Par rapport aux linguistes de l'époque, ce que Austin et Grice apportent de nouveau sur le langage, c'est de considérer de façon centrale le rapport du langage à la communication. Austin introduit la notion d'acte de langage pour souligner que le langage a, avant tout, une fonction actionnelle plus que descriptive et Grice montre que le langage naturel n'est pas imparfait, comme le pensent à l'époque les logiciens et les philosophes analytiques (issus du cercle de Vienne), mais respecte des règles fondées sur une conception relationnelle de la communication. Dès lors, les rapports entre la structure des langues et l'usage de la langue se posent car l'usage avait été laissé de coté dans la tradition structuraliste (comme le témoigne la distinction entre langue et parole de Saussure qui fonde la linguistique sur l'étude de la langue). On distingue donc classiquement la pragmatique et la linguistique en tant que la pragmatique a pour objet l'étude de l'usage de la langue par opposition à l'étude du système de la langue qui concerne la linguistique *stricto sensu.*

de manière locative. Dans la suite de l'interaction, A communique à B une autre actualisation et rectifie celle de B. Dans l'actualisation de A, *le bureau de l'association* ne renvoie plus à un lieu mais à un concept qui représente un groupe de personnes. Les différences entre les actualisations de A et de B portent sur ce que dénote *le bureau de l'association* . C'est-à-dire que le rapport de ce groupe nominal au monde n'est pas imposé par les contraintes qui forment le potentiel de sens de l'énoncé A1. Il dépend de l'actualisation du potentiel. La référence de *le bureau de l'association* se situe dans le rapport d'une monstration langagière négociée dans l'interaction à un monde institué comme terrain commun de l'interaction. Elle relève d'une compréhension en contexte de l'énoncé et elle n'est analysable que dans l'après-coup de l'énonciation, en fonction de l'enchaînement conversationnel.

La référence ne provient donc pas seulement de l'interprétation mais de l'actualisation des contraintes dans le contexte ou dans le terrain commun. Les maximes de Grice (maximes de quantité, de qualité, de pertinence et de manière[22]) expliquent pourquoi, en général, les contraintes issues de l'interprétation du matériau linguistique donnent lieu à des références non ambiguës.

4. Objectifs

En dégageant une problématique interactionniste de modélisation computationnelle du sens des énoncés langagiers, nous mettons en avant la dimension sémantique de l'analyse des interactions langagières. Il ne s'agira donc pas de rendre compte des aspects actionnels et référentiels de la compréhension mais nous chercherons un modèle opératoire de l'interprétation. Celle-ci consiste, comme nous l'avons vu, à mettre en évidence le sens d'un énoncé sous forme de contraintes qui délimitent ses actualisations pragmatiques possibles. Nous construirons, dans la suite, ces contraintes à partir des effets de sens co-textuels qu'impose l'énoncé à son interprétant. Ces effets de sens proviennent de la co-adapation des significations des constituants de l'énoncé. Par exemple, par la co-présence de *IUT* et de *secrétaire* dans l'énoncé *à l'IUT alors je suis secrétaire*, une contrainte d'interprétation indique qu'il est ici question d'une activité professionnelle.

L'analyse des interactions langagières montre, plus encore que l'analyse de textes ou de discours, la langue en usage. C'est évidemment un atout pour une approche pragmatique, mais également pour une approche sémantique comme la nôtre. En effet, pour chaque énoncé à analyser, ceux qui le suivent dans le cours du dialogue permettent d'observer ses interprétations et le sens co-construit qui en est issu. En plus de considérer que le sens d'un énoncé dépend de

[22] [Grice 67].

ses constituants, nous considérons également que l'entour linguistique d'un énoncé, formé d'autres énoncés, amène des indices pour guider un processus ayant pour objectif d'en extraire un sens. L'observation du corpus est donc une source de richesses pour l'apprentissage visé d'une compétence interprétative artéfactuelle. Le but est de mettre la machine, avec un modèle interactionniste de l'interprétation, dans une situation d'apprentissage dans et par l'interaction. Il s'agira pour la machine, d'une part, d'analyser les interactions pour apprendre, et d'autre part, d'apprendre (en même temps) à analyser les interactions. Cette approche constitue une suite aux travaux de Coursil ([Coursil 92]) qui propose de *mettre la machine dans la langue* (plus précisément dans l'activité signifiante de la langue) plus que *de mettre la langue dans la machine*. Cette position, qui ramène essentiellement l'apprentissage des langues naturelles à leur pratique, forme une alternative aux approches consistant à extraire des compétences langagières humaines un ensemble de règles qui puissent diriger le fonctionnement de la machine (c'est par exemple le cas dans les modèles d'analyse syntaxique).

Dès lors, l'apprentissage de la compétence sémantique suit une méthode de conception par amorce ([Nicolle 96]) où le système, à l'état initial, est un noyau de compétences qui contient très peu de connaissances mais où elles seront acquises au fur et à mesure de son activité interprétative. Une telle méthode donne une place centrale à l'interaction homme - machine qui est considérée comme le moteur de l'apprentissage. Dans cette thèse, l'interaction est donc présente à deux niveaux :

1) En tant qu'objet d'étude, étant donné que l'évolution de la compétence interprétative de la machine est expérimentée sur un corpus qui retranscrit une situation réelle d'interaction langagière (celle mise en place dans le projet PIC).

2) En tant que processus mis en place par le modèle de représentation des significations (le modèle *Anadia* que nous détaillerons au chapitre 3), étant donné que la description d'une signification nécessite, dans ce modèle, un point de vue partagé, et donc une boucle interactive entre la machine et l'utilisateur humain. La machine occupe alors un rôle d'apprenant et c'est la raison pour laquelle le modèle que nous proposons est *interactionniste*.

CHAPITRE 2 :

LES EFFETS DU CO-TEXTE SUR LA SIGNIFICATION DES MOTS

En présentant notre problématique de modélisation du sens en langue naturelle, nous avons considéré le sens comme un potentiel qui s'actualise dans l'interaction. L'étude de cette actualisation pose la question de la référenciation et trouve donc sa place dans le champ de la pragmatique où, comme on l'a vu, la prise en compte du contexte est nécessaire. À l'inverse, la construction du sens en tant que potentiel, ne met en jeu que le rapport entre les productions langagières et le système de la langue (tant qu'il s'agit d'IA ou de linguistique car une étude en psychologie ou en psycholinguistique ne pourrait se passer d'une prise en compte du sujet). Cette modélisation peut être envisagée prioritairement selon les quatre dimensions de la sémantique des énoncés : la monstration, la prédication, l'argumentation, et la modalité. Nous nous focaliserons essentiellement ici sur la fonction de monstration, même si dans certains cas elle est difficilement dissociable de la fonction de prédication.

L'objet de ce chapitre est une étude linguistique préalable à la construction du modèle sémantique computationnel que nous proposons dans cette thèse. Le but de cette étude est de mettre en évidence l'autonomie du niveau sémantique, et, plus précisément, de montrer qu'une étude du co-texte suffit pour mettre au jour des phénomènes de co-adaptation des significations des différents lexèmes de la chaîne syntagmatique. Les effets de sens évoqués dans ce chapitre portent donc sur les occurrences des mots dans les énoncés. La modélisation du mot hors contexte fera, quant à elle, l'objet du chapitre suivant.

Dans une première partie, nous rapporterons l'intérêt de l'étude des effets de sens du co-texte à un aspect fondamental des langues : la polysémie. Dans une deuxième partie, nous rappellerons le statut du co-texte par rapport au contexte pour mieux situer cette étude dans le champ de la sémantique (le contexte étant une donnée d'une analyse pragmatique). Par la suite, nous évoquerons un exemple d'analyse du co-texte dans le but de considérer l'effet de multiples marqueurs sémantiques. Ceci nous permettra de préciser les marqueurs que nous retenons pour guider le processus opératoire. En examinant les effets de sens dûs à la co-présence de ces marqueurs dans le co-texte, il sera mis en évidence des phénomènes d'influence et de dépendance sémantiques. Enfin, nous évoquerons l'omniprésence de ces phénomènes.

1. La polysémie des langues naturelles

Dans l'activité langagière de chaque instant, le problème d'un "calcul" du sens ne se pose que très rarement car le sens d'un énoncé est quelque chose qui nous apparaît "immédiatement", hormis dans des cas assez spéciaux volontairement construits pour poser problème (c'est le cas des énigmes par exemple). Ainsi, on peut avoir l'impression que le sens est donné, qu'il suffirait de le "voir", sans charge cognitive liée à l'interprétation ([Rastier & al. 94, p. 68] emploient à ce sujet l'expression de *perception sémantique*). Cette impression d'immédiateté vient de ce qu'on est déjà plongé dans la langue et que l'on ne peut s'en détacher[23]. Les difficultés liées à l'interprétation se posent lorsque l'on apprend une langue étrangère ou lorsqu'il s'agit de calculer le sens d'un énoncé (que ce soit à des fins linguistiques ou à des fins d'Intelligence Artificielle), de façon objective, en le détachant de compétences préalables acquises socialement. Dès lors, les choses les plus évidentes apparaissent comme des difficultés majeures, et essayer, par exemple, de déterminer la signification de l'adverbe "encore" est un problème qui reste ouvert malgré de nombreuses études très fines ([Fuchs 95], [Victorri & al. 96]). C'est à ce problème de pluralité des significations que sont liées les notions d'homonymie et de polysémie.

La pluralité du sens des mots est d'abord apparue dans la constitution des dictionnaires (dés le XVI^{ème} siècle[24]), mais elle n'est abordée pour la première fois en linguistique que dans les travaux de Bréal (en 1897) qui définit alors la polysémie comme la capacité d'un mot à "prendre un nouveau sens" tout en conservant le ou les anciens sens. Bréal la définit dans les termes suivants :

---

[23] Pour faire une métaphore de cet impossible détachement du sujet et de la langue, on peut citer les mots de Wittgenstein, *"la langue est une cage dont on ne peut sortir"* ([Wittgenstein 61]).

[24] D'après le *Larousse Multimédia Encyclopédique*, le premier grand dictionnaire concernant la langue française est le *dictionnaire français - latin* de Robert Estienne en 1539.

> "Le sens nouveau, quel qu'il soit, ne met pas fin à l'ancien. Il existent tous deux l'un à coté de l'autre. Le même terme peut s'employer tour à tour au sens métaphorique, au sens restreint ou au sens étendu, au sens abstrait ou au sens concret... À mesure qu'une signification nouvelle est donnée au mot, il a l'air de se multiplier et de produire des exemplaires nouveaux, semblables de forme mais différents de valeur. Nous appelons ce phénomène de multiplication la *polysémie*." [Bréal 1897, pp. 154-155].

D'un point de vue sémiotique, on aura dans le cas de la polysémie un seul signe dont le signifié est multiple, alors que dans les cas d'homonymie on a deux signes dont les signifiants sont identiques. À titre d'exemple, considérons le mot *bureau.* À ce mot sont associés plusieurs sens, que ce soit la table sur laquelle on écrit, la pièce ou l'immeuble où l'on travaille, les gens avec qui on travaille ou même le fond de l'écran d'un ordinateur. En fait, le premier sens de *bureau* qui maintenant est désuet faisait référence à une étoffe de laine servant à recouvrir les tables. À partir de ce sens, diverses métonymies ou métaphores successives ont permis de dériver les sens que l'on connaît aujourd'hui, et leurs rapports de l'un à l'autre apparaissent clairement (par exemple, on voit bien le rapport métonymique entre les significations de *bureau* comme pièce et comme meuble). Du fait de cette justification, *bureau* est un mot qui a la propriété d'être polysémique. À l'inverse, si on considère le mot *avocat*, on ne peut trouver aucun lien entre sa signification en terme de fruit et sa signification en référence à un homme de loi. *Avocat* est donc le signifiant de deux mots homonymes.

Quand un mot hors contexte évoque plusieurs sens, la question qui se pose alors est de savoir s'il ne s'agit encore que d'un seul mot ou finalement de plusieurs. C'est ce critère de l'unicité du mot qui permet de faire la différence entre les notions de polysémie et d'homonymie. À la différence de la polysémie qui est la propriété d'un seul mot, l'homonymie est une relation entre deux mots. On dira alors qu'un mot est l'homonyme d'un autre si tous deux ont des sens radicalement différents mais "accidentellement" le même signifiant. La question de savoir si deux sens portés par un même signifiant relèvent d'une relation homonymique ou d'une réelle polysémie relève d'un critère diachronique. C'est-à-dire que si un sens est une dérivation d'un premier sens au cours d'une évolution de la langue, on conviendra que l'on à affaire à un unique mot (ayant de fait la propriété d'être polysémique). À l'inverse, s'il est clair qu'il n'y a aucun rapport entre les sens, on aura alors deux mots en relation homonymique. Au cours des évolutions diachroniques de la langue, un mot polysémique peut devenir homonymique si le rapport entre ses différents sens ne se justifie plus. C'est notamment le cas pour le mot *grève*, aujourd'hui homonymique car l'on ne définit plus la signification de *grève* comme *arrêt du travail* par rapport à sa signification comme *bord de rivière*, et pourtant les deux significations sont historiquement liées par le rapport à la *place de grève*.

Comme on vient de le voir, la polysémie concerne des mots possédant plusieurs significations. C'est bien sûr, du point de vue de la langue, une source de richesse considérable qui offre une grande liberté dans les interprétations, mais du point de vue du traitement automatique des langues, c'est une difficulté de taille. D'autant plus qu'on ne peut envisager de traiter les mots polysémiques de manière *ad-hoc* comme des cas particuliers du fait de l'étendue du phénomène qu'est la polysémie. À titre d'exemple du critère massif de la polysémie dans la langue, Victorri et Fuchs ([Victorri & al. 96]) se livrent à une petite expérience : considérant que si l'article de dictionnaire concernant un mot comporte au moins deux subdivisions alors le mot en question est polysémique, si l'on entreprend un comptage de ce type d'article dans le dictionnaire *Le petit Robert*, on arrive à un résultat de 40% des entrées du dictionnaire (alors que l'homonymie ne représente quant à elle que 5%). Si de plus on examine les résultats de plus près, on s'aperçoit que les mots monosémiques sont généralement distribués dans les registres des mots techniques, des nomenclatures ou des mots vieillis alors que les mots polysémiques, au contraire, font le plus souvent partie des quelques milliers de mots qui constituent le vocabulaire de base. Au vu de ces résultats, [Victorri & al. 96] donnent à la polysémie une place centrale dans la langue et un statut de règle plus que d'exception.

On touche ici à une différence fondamentale entre les langues naturelles et les langages formels. Si on considère comme Raccah ([Raccah 96]) qu'on peut définir un terme comme étant un mot qui n'a qu'une seule signification, c'est-à-dire un mot monosémique, alors les langues naturelles sont majoritairement constituées de mots tandis que les langages formels sont, eux, majoritairement (souvent même exclusivement) constitués de termes (éventuellement polymorphiques[25] mais jamais polysémiques sinon ce ne serait pas des termes). Ceci implique, en ce qui concerne le calcul du sens en langue naturelle, qu'une approche purement compositionnelle est d'entrée inadaptée car elle suppose, pour être efficace, une unicité des significations des mots. On ne peut donc pas calculer le sens d'un énoncé en langue naturelle comme on calcule le "sens" d'un programme à l'aide d'un compilateur. En effet, si on cherche à calculer un sens en composant des significations possibles pour chaque mot le long de l'arbre résultant de l'analyse syntaxique, on est amené à construire un grand nombre de solutions tout en sachant que très peu d'entre elles seront pertinentes. Ceci pose deux problèmes, le premier est celui d'un filtrage des solutions qui est coûteux en temps, en mémoire et donc en efficacité du système, et le second, plus radical encore, est celui d'un risque d'explosion combinatoire directement dû à la multiplicité des significations des éléments de base.

Ces problèmes de combinatoire que rencontrent les modèles classiques, basés avant tout sur la logique ([Morris 39], [Van Dijk 77]), renseignent sur les critères fondamentaux des

[25] Une expression d'un langage formel est dite polymorphique lorqu'elle peut se manifester sous plusieurs formes en gardant la même signification. C'est par exemple le cas du prédicat d'égalité qui peut s'appliquer à des nombres ou à des chaînes de caractères (5 = 5 et "abc" = "abc") mais qui sera évalué au moyen de fonctions différentes.

langues naturelles qu'on ne peut plus considérer comme secondaires (la polysémie et la métaphore en sont de très bons exemples) et nous invitent à approfondir une étude linguistique des phénomènes, étude préalable à une modélisation computationnelle de la langue.

Parmi les notions souvent évoquées dans les études linguistiques, celle de contexte paraît tenir une place centrale dans l'observation sémantique de productions langagières. C'est grâce à l'influence du contexte que la polysémie n'est pas qu'une source d'ambiguïté, mais bien une subtilité du système qu'est la langue. En effet, même un mot extrêmement polysémique comme *bureau* ne pose en général pas de problème en emploi quant à la détermination de sa signification. Par exemple, dans un énoncé comme

Tu as lu le livre qui est posé sur le bureau?

les significations en référence à l'immeuble ou au groupe de personnes sont d'emblée exclues par un contexte qui conduit l'interprétation à ne retenir que la signification du mot *bureau* en rapport au meuble. Dans un tel exemple, il est clair que considérer toutes les significations possibles de *bureau* est inutile et préjudiciable à l'efficacité du système, car c'est ne pas avoir tenu compte des fortes indications qu'exprime le contexte.

Reste maintenant à se donner les moyens de calculer et d'exprimer les indications données par le contexte. C'est là tout le but de notre étude. Ainsi, dans l'exemple donné ci-dessus, on va exprimer l'indication contextuelle sur la signification du mot *bureau* par la relation entre le mot *bureau* et le prédicat *posé sur*. Ce prédicat contraint la signification de son objet et on l'exprimera ici par une relation de **dépendance sémantique** entre *posé sur* et le mot *bureau.*

Au delà de la polysémie lexicale, il existe une polysémie des expressions, que Gosselin qualifie de contextuelle et généralisée, qui ne peut s'exprimer comme une simple combinatoire de la polysémie lexicale. Les métaphores culturelles dont Lakoff et Johnson ont montré qu'elles structurent le langage, la pensée et l'action des humains (par exemple la métaphore LA DISCUSSION, C' EST LA GUERRE[26] [Lakoff & al. 85, p. 14]) donnent lieu à ce type de polysémie. Dans [Gosselin 96b] la "polysémie contextuelle généralisée" désigne le fait que la signification d'un marqueur puisse varier en fonction non seulement des formes, mais aussi des significations des autres marqueurs qui l'entourent dans le co-texte, lesquelles varient aussi de façon semblable. La modélisation de la dépendance sémantique est un mode de traitement de

[26] Lakoff et Johnson définissent la métaphore comme un moyen de faire comprendre quelque chose (et d'en faire l'expérience) en termes de quelques chose d'autre ([Lakoff & al. 85, p. 15]). Ainsi, la métaphore LA DISCUSSION, C'EST LA GUERRE permet de parler de la discussion en utilisant des mots propres à la guerre, comme en témoignent les exemples suivants (empruntés à Lakoff et Johnson) : Vos affirmations sont *indéfendables.* Il a *attaqué chaque point faible* de mon argumentation. Ses critiques visaient *droit au but*. J'ai *démoli* son argumentation. Je n'ai jamais *gagné* sur un point avec lui. Tu n'es pas d'accord ? Alors, *défends*-toi ! Si tu utilises cette *stratégie*, il va t'écraser. Les arguments qu'il m'a opposés ont tous *fait mouche*.

cette polysémie, et est une façon de rejeter une sémantique atomiste qui suppose une invariabilité contextuelle des significations. Nous nous plaçons alors, comme Gosselin, dans le cadre d'une sémantique **holiste** qui reconnaît un principe de compositionnalité en considérant que le sens d'un énoncé résulte bien des significations de ses composants mais qui reconnaît également un principe de contextualité dans la mesure où la signification d'une expression est au moins partiellement déterminée par le contexte dans lequel elle apparaît. Cette approche holiste est également celle opérationnalisée d'une manière connexionniste dans le modèle dynamique de construction du sens de Victorri ([Victorri & al. 96]). L'approche holiste prend sa source dans la sémiotique de Saussure et, plus précisément, dans la valeur syntagmatique des signes :

> "Le tout vaut par ses parties, les parties valent aussi en vertu de leur place dans le tout, et voila pourquoi le rapport syntagmatique de la partie au tout est aussi important que celui des parties entre elles" [Saussure 1915,p. 177]

Tout au long de ce chapitre, nous allons montrer comment les constituants de la chaîne co-déterminent leurs significations grâce à leur caractère polysémique. Les phénomènes de dépendance sémantique correspondent à de telles co-déterminations et leur modélisation opératoire est un moyen d'exprimer les indications du contexte. Précisons donc cette fameuse notion de contexte.

2. Contexte et co-texte

Le contexte non linguistique regroupe tous les éléments qui caractérisent l'environnement d'une production langagière (les interlocuteurs, leur tâche, le lieu, le temps, l'histoire de l'interaction, ...) et qui n'apparaissent pas dans une analyse purement linguistique, qui ne peut faire intervenir comme contexte que le co-texte. L'étude de textes écrits amène à privilégier le co-texte, ou contexte linguistique, alors que dans les dialogues, on s'aperçoit facilement que beaucoup d'éléments non linguistiques, situationnels, entrent en jeu pour rendre compte de l'effectivité de l'interaction. C'est notamment le cas des objets désignés par les déictiques dont nous avons signalé, dans le chapitre précédent, l'importance dans les corpus de situations de cognition située.

Kleiber ([Kleiber 94]) critique les modèles en couche (du type syntaxe - sémantique - pragmatique) où le contexte n'entre en ligne de compte que dans la dernière étape et seulement si l'on en a besoin. Il qualifie cette approche de standard, en opposition à une approche dite cognitive où les mécanismes interprétatifs sont fortement liés à une effectivité de la mémoire. Dans cette approche, que l'on retrouve par exemple dans [Bassi Acuña 95], le contexte est considéré de façon dynamique comme une réalité cognitive. Il inclut en mémoire à court terme la

donnée linguistique[27] et ce qui est mémorisé de la situation, et en mémoire à long terme les connaissances générales. L'approche cognitive considère alors le contexte comme un élément décisif dans le processus d'interprétation de toute expression verbale et lui donne le statut d'une donnée première d'un modèle du sens.

Le problème est que, dans ce rapport entre le sens et le contexte, ce ne sont pas toujours les mêmes données contextuelles qui sont pertinentes. Les différences entre approches peuvent alors être des différences de choix de données contextuelles. Pour reprendre des exemples de Kleiber, le sens de l'énoncé

Dix vols par jour

est d'autant plus contextuel que la polysémie du mot *vol* rend le sens de l'énoncé inséparable de sa situation. Savoir qu'il s'agit ici d'un message publicitaire pour une compagnie aérienne ou d'un rapport de commissariat de police permet d'en trouver le sens. De même, dans le cas d'un dialogue, le statut des personnes est une donnée contextuelle importante. Ainsi, le sens de l'énoncé suivant apparaît lorsque l'on sait qu'il est extrait d'un dialogue entre deux mathématiciens.

L'idéal de l'anneau se plonge facilement dans un corps.

Un tel énoncé reste donc obscur et ambigu tant qu'il est présenté sans les données contextuelles pertinentes. C'est une ambiguïté volontairement entretenue et qui s'apparente à la construction d'énigmes où le but est justement de ne pas donner le contexte pour que la personne devant la résoudre reconstruise ce contexte pour trouver la solution.

La prise en compte de telle ou telle donnée contextuelle est alors fonction de la tâche qu'on se donne ou du corpus sur lequel on travaille. Par exemple, en ce qui concerne l'analyse des corpus du projet PIC, on peut décrire, dans un premier temps, le contexte par les données suivantes :

la tâche au centre de l'interaction,
les personnes présentes,
les statuts et rôles de ces personnes,
le lieu de l'interaction,
les objets présents ou évoqués,
le temps dans lequel se déroule l'interaction.

Dans un second temps, le problème est de savoir si le contexte, tel qu'il est décrit, est envisagé de façon exhaustive et plus précisément de savoir où s'arrête sa description :

[27] La donnée linguistique doit être mémorisée à court terme (et oubliée dés la fin de son interprétation) selon le principe de non consignation de la chaîne parlée, mis en évidence dans [Coursil 92].

> "Si je décris par exemple une consultation médicale hospitalière, le contexte ce sera le *hic et nunc* de la consultation, mais aussi l'hôpital particulier où elle a lieu et son fonctionnement, l'institution hospitalière en général, l'ensemble du système des soins tel qu'il fonctionne en France, et à la limite, la société française dans son entier." ([Kerbrat-Orecchioni 96])

Ce problème, que soulève aussi Kleiber, concerne finalement la possibilité de préciser clairement les composantes du contexte afin de le rendre objectif et utilisable pour toute analyse. Latraverse y voit même un problème de fond pour ce qui est de son utilité théorique :

> "La notion de contexte est d'une telle souplesse et d'un accueil si généreux qu'il est difficile de considérer qu'elle a des frontières suffisamment établies pour jouer un rôle théorique non univoque." ([Latraverse 87, p. 194])

Pour apporter une réponse à ce problème, Kleiber propose de limiter le contexte à une dimension mémorielle issue de la perception de la situation. Ceci ne pose pas de problèmes au regard d'une approche purement linguistique pour laquelle le but est de décrire les phénomènes langagiers, mais dans le cadre d'une modélisation en intelligence artificielle, la nature projective (*vs*. descriptive [Vernant 98, p. 112]) de l'analyse imposerait de faire "entrer" le monde dans la machine avant même toute interprétation car, d'une certaine façon, une machine ne peut percevoir que ce qu'elle connaît déjà. Pour des raisons évidentes liées aux phénomènes d'explosion combinatoire, un tel choix n'aboutirait qu'à un échec.

Faire l'inventaire exhaustif des données contextuelles est effectivement impossible dans l'absolu car le contexte ne peut pas être déterminé *a priori* de façon fiable puisque c'est aussi le sens de l'énoncé qui le détermine. Par exemple, l'énoncé

Pour avoir le focus, cliquez dans la barre supérieure de la fenêtre.

place son interprétation dans le contexte des applications informatiques avec interface graphique, même si ce contexte n'était pas déjà évoqué précédemment (et même si cet énoncé est à comprendre de façon métaphorique dans un contexte qui n'entretient aucun lien avec l'informatique). Au regard d'un tel exemple, il paraît difficile de considérer que le contexte est une donnée première de l'interprétation, étant donné qu'elle le détermine en partie. On ramène donc le contexte à la notion de terrain commun mis en évidence en psychologie ([Clark & al. 86], [Vivier 92]). Le terrain commun permet de borner le contexte, évitant ainsi le recours à une description du monde dans son ensemble. Il est autant une donnée première qu'un but dans l'interaction. L'intérêt d'une prise en compte du contexte n'est donc pas de participer à une construction du sens mais de savoir comment ce sens s'actualise dans l'interaction. C'est pourquoi le sens que nous chercherons à cerner est un sens purement linguistique, non situationnel. C'est la contribution du matériau linguistique à la compréhension.

L'étude de la co-détermination des significations dans la chaîne intéresse la construction du sens (en tant que potentiel), et non son actualisation pragmatique. De ce point de vue, les indices de ces co-déterminations sont à chercher dans le co-texte et non dans le contexte. Cette position ne constitue pas un retour aux modèles en couche car il ne s'agit en aucun cas de déléguer au niveau pragmatique d'éventuelles insuffisances du niveau sémantique. La sémantique et la pragmatique sont deux niveaux d'étude, et non deux étapes d'une analyse. Il s'agit de ne pas confondre leurs modèles (autonomes les uns par rapport aux autres) et leurs mise en forme opératoire dans une implémentation (où chacune détermine les autres, d'où des architectures de tableau noir comme dans [Bricon-Souf 94]). C'est pourquoi les ambiguïtés pragmatiques ne sont pas décelables au niveau sémantique. Par exemple, l'énoncé

Prends ça

peut poser des problèmes de référence pour le déictique *çà*, mais l'ambiguïté éventuelle est alors intégralement pragmatique car, au niveau de la sémantique, on ne peut rien avancer de plus que la présence du déictique. Ce n'est pas le cas dans l'énoncé suivant où la sémantique peut mettre en évidence une relation anaphorique entre *ça* et *les poires*.

Ces poires sont belles, ça devrait bien se vendre.

Les deux niveaux sont donc clairement séparés : grâce au co-texte, la sémantique établit un sens potentiel modélisable en terme de contraintes et la pragmatique actualise ce sens dans un contexte, notamment par l'établissement de relations de référence.

Reprenons l'exemple cité précédemment : *Dix vols par jour* qui, sans contexte, est particulièrement ambigu. Il ne présente pourtant plus aucune ambiguïté si on le considère dans les deux extraits de dialogues fictifs suivants :

A1 : *tu as vu la dernière publicité pour Air Inter au sujet de la ligne Paris-Caen ?*
B1 : *ah oui, qu'est-ce qu'ils disent ?*
A2 : *dix vols par jour*

A1 : *j'ai lu un rapport sur les délits dans les supermarchés*
B1 : *ah oui, qu'est-ce qu'ils disent ?*
A2 : *dix vols par jour*

Dans un dialogue, un énoncé fait partie d'une séquence de productions langagières. Si la séquence suffit à lever des ambiguïtés potentielles (comme c'est le cas dans les deux exemples ci-dessus), ces ambiguïtés ne sont donc pas pragmatiques mais bien sémantiques. Le co-texte rend compte d'utilisations conjointes de mots et ces mots influent les uns sur les autres pour créer du sens en précisant leurs significations. Ainsi, dans le premier exemple de dialogue,

l'utilisation conjointe de *vols* et de *Air Inter* génère l'effet de sens qui rend l'énoncé non ambigu, de même pour *vols* et *délits* dans le deuxième exemple.

Pour chaque mot, l'énoncé qui le contient ainsi que les autres énoncés du dialogue constituent son co-texte (son contexte linguistique). C'est dans ce co-texte que des marqueurs sémantiques avec lesquels un lexème produit des effets de sens peuvent être recherchés. Dans une situation de dialogue, notre hypothèse est de rendre compte des effets de sens co-textuels en faisant apparaître des phénomènes de dépendance entre lexèmes. Cette hypothèse paraît d'autant plus réaliste que, si une ambiguïté se pose réellement au cours de l'interaction langagière, celle-ci va donner lieu à un dialogue incident et donc elle sera verbalisée. De ce fait, grâce à la dynamicité du dialogue, les traces linguistiques permettant une levée de cette ambiguïté apparaîtront dans le co-texte.

Dans l'étude d'une production langagière, on peut mettre au jour, à différents niveaux de la chaîne syntagmatique, divers constituants de cette chaîne. Par exemple, à un niveau phonétique, on peut par exemple signaler des sons et des pauses, ou encore des montées et des descentes de fréquence ou de puissance. De même, à un niveau phonologique, on peut mettre en évidence des phonèmes et des syllabes. À un niveau lexical, on fait notamment apparaître des radicaux, des désinences et des lexèmes. À un niveau syntaxique, on trouve différents groupes dont les groupes nominaux et les groupes verbaux. Au regard d'une analyse d'une chaîne syntagmatique, dès qu'un des constituants de la chaîne est porteur d'une information pertinente, on lui donne le statut de **marqueur**. Ainsi, dans une analyse sémantique, les lexèmes, les morphèmes ou encore les groupes syntaxiques sont des marqueurs, mais les phonèmes, par exemple, n'en sont pas. Notre perspective consiste à envisager la mise en place du sens dans sa dimension de monstration comme une co-adaptation de différents marqueurs du co-texte. Cette approche a pour objectif d'en prolonger une autre, développée par Gosselin ([Gosselin 96a]), plus ciblée sur la dimension aspectuo-temporelle du sens que nous allons présenter.

3. L'effet des marqueurs du co-texte sur le sens. L'exemple de la temporalité

Gosselin met en place un modèle calculatoire et cognitif des relations aspectuo-temporelles que porte un texte en français contemporain. Le but est de pouvoir prédire pour un verbe et en fonction de son co-texte, les déformations éventuelles du procès auquel il fait référence. Nous allons ici rapidement en présenter les principes généraux. Dans ce modèle, l'accent est mis sur une représentation des propriétés aspectuo-temporelles des marqueurs. Le calcul sémantique réside alors dans l'assemblage de ses propriétés. Les représentations aspectuo-temporelles des

marqueurs (aspect, temps absolu, temps relatifs) sont définies par les relations que peuvent entretenir quatre types d'intervalles disposés sur l'axe du temps : l'intervalle de procès [B1,B2], l'intervalle d'énonciation [01,02], l'intervalle de référence (ou intervalle de monstration qui correspond à ce qui est perçu du procès) [I,II], et l'intervalle circonstanciel [ct1,ct2]. Ainsi, à chaque énoncé est associé un unique intervalle d'énonciation, à chaque circonstanciel temporel correspond un intervalle circonstanciel, et à la principale et à chaque subordonnée sont liés un intervalle de référence et un intervalle de procès comme le montre l'exemple suivant (extrait de [Gosselin 96b]) :

Ce jour là, Pierre annonça qu'il était malade.

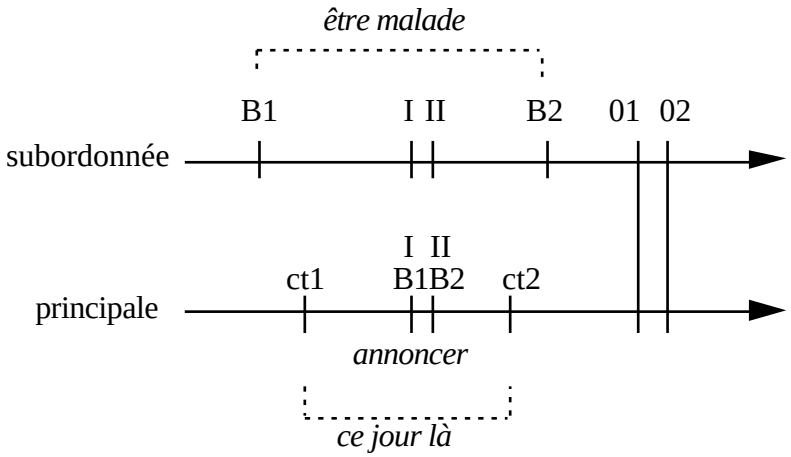


Figure 2.1 : Intervalles temporels

L'aspect grammatical a quatre valeurs (aoristique, inaccompli, accompli et prospectif) qui sont définies par les relations entre l'intervalle de procès et l'intervalle de référence :

aspect aoristique :	[I,II] coïncide avec [B1,B2]
	ex. : *il mangea sa soupe*
aspect inaccompli :	[I,II] inclus dans [B1,B2]
	ex. : *il mangeait sa soupe (depuis 5 min)*
aspect accompli :	[I,II] postérieur à [B1,B2]
	ex. : *il a terminé sa soupe (depuis 5 min)*
aspect prospectif :	[I,II] antérieur à [B1,B2]
	ex. : *il va sortir (car il est habillé)*

De même, les trois valeurs du temps absolu (passé, présent, futur) trouvent une justification dans les rapports entre l'intervalle de référence et l'intervalle d'énonciation :

passé :	[I,II] est antérieur à [01,02]
présent :	[I,II] et [01,02] coïncident (c'est le cas général mais il peut arriver que

les intervalles se recouvrent partiellement)

futur : [I,II] est postérieur à [01,02]

Enfin, le temps relatif résulte de la relation entre l'intervalle de référence de la subordonnée [I',II'] et celui de la principale [I,II] :

antérieur : [I',II'] est antérieur à [I,II]
simultané : [I,II] et [I',II'] coïncident
ultérieur : [I',II'] est postérieur à [I,II]

Avec ces représentations, les informations temporelles portées par le verbe dans l'énoncé trouvent une description en terme d'intervalles. Ces intervalles apportent des précisions sur le procès auquel le verbe fait référence. Ces informations sont représentées par des règles de la forme suivante :

Si le temps morphologique est l'imparfait **alors** l'aspect est inaccompli

Si l'aspect est inaccompli **alors** B1 < I et II < B2

Dans certains cas, les conclusions de ces règles entrent en contradiction. Ce sont ces cas que Gosselin appelle des conflits (il signale d'ailleurs qu'ils sont très fréquents). Ces conflits sont résolus par d'autres règles qui résolvent les contradictions en déformant, par exemple, les bornes des intervalles. Ces modes de résolution de conflits amènent à déformer de façon régulière et prédictible la représentation initiale du procès et donc à changer contextuellement la signification du verbe. C'est donc bien d'un effet de sens qu'il s'agit, effet qui résulte de la co-présence du verbe et de certains marqueurs. L'exemple suivant montre une telle résolution de conflit :

Il nageait pendant des heures.

Cet énoncé est à l'imparfait. Par application des règles, on en déduit que l'aspect est inaccompli et donc que B1 < I et II < B2. Le circonstanciel introduit par *pendant* (*pendant des heures*) est un circonstanciel de durée et implique donc que ct1 = B1 et ct2 = B2; de plus il est non ponctuel, donc ct1 est non confondu avec ct2 (ct1 précède ct2). Un principe général sur les circonstanciels (transcrit sous forme de règles) précise que, pour qu'un procès entre en relation avec un circonstanciel, il est nécessaire que les bornes du procès (intervalle [B1,B2]) soient accessibles à partir de l'intervalle de référence ([I,II]), c'est-à-dire qu'il y ait au moins coïncidence entre les deux intervalles. On déduit de ce principe que I ≤ B1 et que B2 ≥ II. Le résultat de ce principe induit un conflit avec l'aspect inaccompli (B1 < I et II < B2). L'une des résolutions possibles de ce conflit consiste à dupliquer les intervalle de procès et intervalle de

référence pour rendre compte d'une itération. Ceci implique de se donner un nouvel intervalle de référence spécifique ([Is,IIs]) regroupant des itérations de procès aoristiques (compatibles avec les exigences associées au circonstanciel de durée) et sur lequel l'aspect inaccompli peut s'appliquer dans son ensemble (comme en témoigne la possibilité d'énoncer *Il nageait pendant des heures depuis 3 ans*). On peut conclure de cet exemple de résolution de conflit que le verbe *nager* a ici contextuellement pris le sens de *nager régulièrement.*

Le modèle que propose Gosselin pour la temporalité du français contemporain rend bien de compte de phénomènes d'effets de sens dûs aux marqueurs aspectuo-temporels. Il est calculatoire et a donné lieu à une validation par un système expert contenant environ 500 règles (regroupant les règles qui traduisent les propriétés aspectuo-temporelles et les règles de résolution des conflits).

À l'inverse de Gosselin qui étudie uniquement la temporalité, nous ne nous restreindrons pas a priori à une seule dimension du sens dans l'analyse sémantique du co-texte. Comme nous le verrons dans la partie suivante, nous chercherons, par l'étude de la co-présence de lexèmes, à rendre compte de l'adaptation co-textuelle des significations dans les multiples dimensions du sens d'un énoncé, c'est-à-dire aussi bien dans la dimension temporelle que spatiale ou encore catégorielle (dans le sens où la sémantique met à jour des catégories à l'oeuvre dans les processus de co-référenciation).

4. L'effet de la co-présence des lexèmes dans le co-texte

Les effets de sens directement produits par la co-présence de certains lexèmes participent au potentiel de sens d'un énoncé. On fait l'hypothèse que le sens d'une expression n'est prioritairement le fait ni de l'un ni de l'autre des mots qui la composent mais qu'il résulte de leur rapprochement. Ce rapprochement n'est pas lié à une proximité dans la structure syntaxique, ni à une association au niveau de la mémoire sémantique comme par exemple dans les relations métonymiques, mais plutôt à la simple co-présence dans la chaîne syntagmatique, c'est-à-dire à une utilisation conjointe des mots.

Nous commencerons par dégager deux formes d'effets de sens dues à l'utilisation conjointe de deux lexèmes dans un même co-texte : l'influence sémantique et la dépendance sémantique. Ces deux formes sont fonction de la nature du rapprochement des lexèmes. Elles expriment deux niveaux dans les effets du co-texte : l'influence sémantique fait émerger certains effets de sens et la dépendance sémantique pose des contraintes sur des effets de sens possibles.

4.1. L'influence sémantique

L'influence sémantique est le premier niveau de description des effets de sens induits par la co-présence de mots dans un texte ou un dialogue. Elle rend compte de la co-adaptation des significations de certains éléments d'une chaîne syntagmatique.

Considérons, par exemple, l'énoncé suivant :

La vision est fatiguante au pôle, la glace reflète beaucoup de lumière.

Dans cet énoncé apparaissent les mots *pôle* et *glace.* Chacun de ces deux mots est polysémique : le mot *pôle* peut avoir la signification d'un lieu (pôles terrestres ou encore centre de compétence[28]) ou la signification d'une extrémité (pôles plus et moins d'une pile par exemple) et le mot *glace* peut avoir une signification d'eau solide ou de dessert glacé, ou encore de miroir. Dans cet énoncé, les mots bien que polysémiques ne soulèvent pas d'ambiguïtés car la co-présence des mots *pôle* et *glace* conduit l'interprétation à retenir la signification de *pôle* terrestre pour *pôle* et d'eau solide pour *glace.*

Au regard de cet exemple, il apparaît que l'influence sémantique décrit le fait que le co-texte tend à uniformiser des propriétés portées par les mots qui le constituent. C'est-à-dire que la mise en co-présence de mots dans un co-texte engendre deux conséquences :

- d'une part, elle fait ressortir les propriétés que ces mots ont en commun,
- d'autre part, elle tend à effacer, pour chaque mot, les significations qui ne correspondent pas à des propriétés communes ; c'est en cela que les mots précisent effectivement leur signification dans le co-texte qu'ils forment.

On voit d'ores et déjà apparaître en filigrane une notion qui est centrale dans le modèle sémantique computationnel que nous proposons dans cette thèse et qui fera l'objet du chapitre 4, la notion d'isotopie. Ainsi, l'influence sémantique sera considérée comme l'effet des isotopies construites dans le co-texte sur la signification des mots de la chaîne. Comme nous le verrons au chapitre 4, la recherche des isotopies permet de rendre compte des influences sémantiques produites par un énoncé et de les localiser. En ce qui concerne la question de savoir où vont se produire les influences sémantiques, le contenu sémantique des mots hors contexte sera de toute évidence un critère qui va entrer en ligne de compte et son étude sera l'objet du chapitre 3.

[28] C'est le cas dans *pôle modélisation en sciences cognitives de la MRSH.*

4.2. La dépendance sémantique

L'influence sémantique est, comme on vient de le voir, une façon de décrire pour chaque lexème les effets de sens auxquels il participe dans son co-texte. On peut alors considérer que l'influence sémantique fait référence à une contribution du co-texte dans la précision de la signification d'un lexème. La dépendance sémantique exprime une idée de plus par rapport à l'influence. Cette idée est, qu'en plus de l'effet de sens qu'ils engendrent, les lexèmes en cause occupent une place qui entre déjà dans une relation syntagmatique. La dépendance sémantique constitue alors une influence sémantique renforcée par une dépendance syntagmatique. La conséquence consiste à exprimer une contrainte du co-texte pour la signification d'un lexème, et plus uniquement une contribution du co-texte. La dépendance sémantique est donc une relation plus forte que l'influence et, de ce fait, plus facilement prédictible dans une optique calculatoire.

En effet, certaines structures syntagmatiques induisent une influence sémantique prédictible. Elle donnent alors lieu à des phénomènes de dépendance sémantique. C'est notamment le cas des énumérations. Par exemple, l'interprétation de l'énoncé (emprunté à [Bassi Acuña 95, p. 67])

Sur la table, il y a des pommes, des poires et des kiwis.

conduit l'interprétant à donner au référent de *kiwis* les propriétés des fruits comestibles, même si celui-ci les ignorait. Cette dépendance sémantique, liée aux énumérations, est d'ailleurs socialement marquée car la phrase suivante (formant un exemple de *zeugme*[29]), qui montre qu'une énumération n'est pas une simple factorisation, paraît choquante précisément parce que l'énumération ne dégage pas de propriétés communes.

La tour Eiffel est en métal et en hauteur.

En cherchant à rendre compte de mises en relations de signifiés dans la chaîne syntagmatique, Mettinger ([Mettinger 94]) a recensé quelques formes syntaxico-argumentatives, dont précisément les énumérations qui induisent des relations sémantiques fortes sur les lexèmes. Le travail de Mettinger, qui illustre la notion de dépendance sémantique, relève de la sémantique différentielle ([Rastier 87]) où les significations des mots sont exprimées en terme de traits sémantiques qui permettent de décrire des différences entre les mots. Dans ce cadre d'étude, l'hypothèse de Mettinger est que la mise en relation en contexte de significations joue un rôle dans la description et dans la dynamique des traits sémantiques au cours du processus interprétatif, et c'est pour vérifier cette hypothèse qu'il cherche les structures syntaxiques

[29] Le zeugme est une figure de rhétorique correspondant à une coordination grammaticale de deux mots qui possèdent des sèmes opposés. Par exemple /abstrait/ (dans *découragement*) et /concret/ (dans *loques*) dans "On croirait voir deux femelles grises, habillées *de loques et de découragement*." (exemple repris de [Ducrot & al. 72, p. 355]).

particulières qui sous-tendent ces relations. Ainsi, il dégage, à partir d'une étude de corpus portant sur une vingtaine de romans en langue anglaise, une douzaine de formes opposant les significations d'un mot X et d'un mot Y dont par exemple : *X et Y, ni X ni Y, X ou Y, X plutôt que Y, de X à Y, X et en même temps Y*. Nous expliquerons aux chapitres 3 et 4 que nous nous référons également à la sémantique différentielle. Notamment, nous proposerons de rendre compte de la dynamique des traits à l'oeuvre dans la chaîne par la dépendance sémantique en rapport à la cohésion et à la cohérence de l'énoncé[30].

Les phénomènes de dépendance sémantiques ne se limitent pas à des formes linguistiques données de façon exhaustive comme celles que propose Mettinger. Ils recouvrent d'autres relations syntagmatiques comme, par exemple, les relations entre un sujet et son verbe ou encore les relations d'anaphore. Nous allons détailler les formes de dépendance sémantique dans deux types de co-texte : un co-texte intraphrastique[31] correspondant aux co-présences de lexèmes au sein d'un même énoncé et un co-texte extraphrastique correspondant aux co-présences de lexèmes dans des suites d'énoncés.

4.2.1. La dépendance sémantique intraphrastique

Il a été mis en évidence très tôt (dès les travaux de Patañjali [Yagi 84] sur les premières grammaires du Sanscrit au deuxième siècle avant J.C.) que la formation d'une phrase ou d'un énoncé respecte des règles d'organisation. Formuler ces règles pour une langue donnée de façon objective constitue le but d'une étude syntaxique qui vise à rendre compte des contraintes que doit respecter une phrase dans la langue en question.

Différentes approches de la syntaxe ont été formulées. Par exemple, le distributionnalisme ([Harris 54]), développé à partir des travaux de Bloomfield dans les années 20, décrit les régularités d'une langue par les possibilités combinatoires d'enchaînement de ses éléments (ainsi quand deux mots apparaissent dans la même distribution, c'est-à-dire dans la même forme de séquence, alors ils appartiennent à une même classe syntaxique). Ensuite, d'autres approches ont été proposées dans les années 60, avec les théories de Chomsky ([Chomsky 69]) et de Tesnière ([Tesnière 59]).

---

[30] En ce qui concerne les énumérations, nous verons (cf. Chapitre 4) que la dépendance sémantique entre X et Y dans *X et Y*, impose deux effets de sens entre X et Y : une similarité (comme en témoigne l'exemple du zeugme), et une différence (c'est le cas des deux occurrences de *musique* dans *il y a musique et musique*)

[31] On considère ici la phrase comme la chaîne de signes linguistiques formant l'énoncé (i.e. l'énoncé-type chez [Victorri & al. 96]), et non comme une structure propositionnelle ou syntaxique.

La méthode développée à partir des travaux de Chomsky[32] consiste à former, dans une phrase, des groupes syntagmatiques qui englobent des unités linguistiques adjacentes dans la chaîne, puis à les regrouper de nouveau et ce jusqu'à arriver à faire de la phrase un seul et unique groupe. Les règles syntaxiques constituent alors principalement des règles de réécriture et l'ensemble de ces règles forment des grammaires. Le résultat d'une analyse syntaxique, dans l'approche générative, donne lieu à un arbre de constituants dont la racine est un symbole représentant la phrase, où les noeuds représentent les groupes syntagmatiques intermédiaires et où les feuilles sont les mots de la phrase.

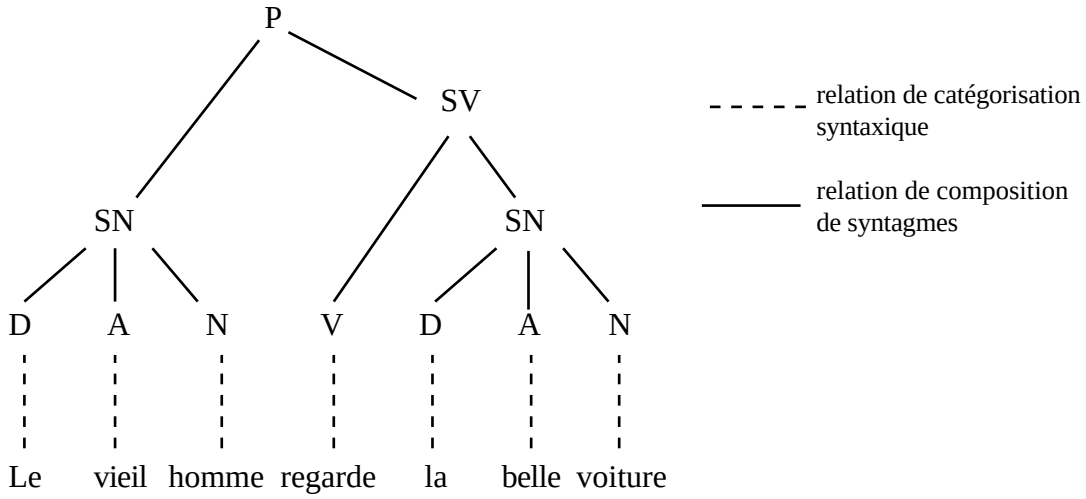


Figure 2.2 : Arbre de constituants de la phrase :

Le vieil homme regarde la belle voiture.

L'approche de Tesnière exprime autrement que par des règles de dérivation les relations entre les mots d'une phrase. La structure syntaxique de la phrase y est mise au jour par des relations de dépendance entre les différents mots de la phrase. Ces dépendances représentent des relations de rection et de subordination entre des mots ou des syntagmes. Ainsi, la règle exprimant le fait qu'un verbe à l'indicatif ne peut se passer de son sujet sera décrite en termes de rection du verbe sur le sujet qui, de fait, lui est subordonné. L'arbre syntaxique est, dans cette perspective, une structure hiérarchique où seuls les mots de la phrase apparaissent et où chaque branche de l'arbre met en relation un terme supérieur régissant et un terme inférieur subordonné. La hiérarchie dans l'arbre détermine l'étiquetage des parties du discours (par exemple un sous-arbre dont la racine est un nom représente un groupe nominal).

[32] L'objectif de Chomsky était de proposer une *grammaire générative*, c'est-à-dire qu'il s'interessait à la génération d'une phrase ou d'un ensemble de phrases à partir d'un axiome et d'un jeu de règles. Son but n'était donc pas, en premier lieu, de proposer un cadre théorique pour l'analyse syntaxique des phrases. Ce sont les linguistes et informaticiens qui ont utilisé les thèories de Chomsky à des fins d'analyse.

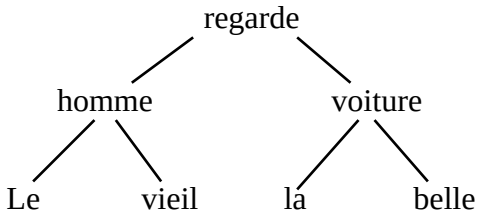


Figure 2.3 : Arbre de dépendance de la phrase :
Le vieil homme regarde la belle voiture.

Il apparaît clairement que dépendance et organisation syntaxique sont deux notions extrêmement liées et Tesnière montre que, dans les phrases, c'est principalement le verbe qui projette ses dépendances. Cependant, ce serait une erreur de penser que la dépendance ne relève que du niveau syntaxique car, dans ce cas, le calcul du sens d'un énoncé à partir des significations de ses constituants ressemblerait à la construction d'un mur à partir de briques. Ce n'est pas le cas, car la signification de chaque mot dans un énoncé est bien souvent contrainte par les autres mots avec lesquels il entretient des relations de subordination ou de rection, ce qui constitue au niveau sémantique une autre forme de dépendance, une dépendance sémantique.

Au lieu de considérer qu'il incombe à la sémantique de venir en aide à la syntaxe quand celle-ci est face à un problème, nous allons utiliser la syntaxe comme un support permettant de prédire les groupes syntagmatiques dans lesquels la sémantique peut mettre au jour des effets de sens co-textuels. Ces effets de sens, du fait de leur justification syntaxique forment des cas de dépendance sémantique. Pour pouvoir rechercher les cas de dépendance sémantique dans les co-textes intraphrastiques, nous allons, dans un premier temps, chercher les relations syntaxiques qu'entretiennent les mots des énoncés. Pour cela, nous faisons l'hypothèse qu'une bonne démarche consiste à exprimer déjà les relations syntaxiques en terme de dépendance. On utilisera donc pour cela un formalisme syntaxique apparenté aux grammaires de dépendance de Tesnière. Plus précisément, nous exploiterons l'analyseur syntaxique de Vergne[33] ([Vergne 94], [Giguet & al. 97a], [Giguet & al. 97b]) qui formule effectivement les relations syntaxiques en terme de dépendance et insiste particulièrement sur la distinction actants - circonstants. Ceci, comme on va le voir dans la suite, est une préoccupation tout à fait en accord avec ce que nous attendons de l'analyse syntaxique.

Reprenons l'exemple d'une énumération:

En faisant mon marché, j'ai vu des poireaux, des concombres et des avocats.

Cette énumération, comme on l'a signalé précédemment, produit une dépendance sémantique issue de la présence des mots *poireaux*, *concombres* et *avocats*. Cette dépendance qui contraint

[33] Des démonstrations de cet analyseur sur quelques selections de corpus sont disponibles sur le WEB à l'adresse *http://www.info.unicaen.fr/~giguet/syntaxique.html*

l'interprétation à rendre récurrentes des propriétés lexicales communes tend à considérer prioritairement que le mot *avocat* aura ici une signification d'aliment plus que d'homme de loi. Dans l'exemple qui suit, le problème de la signification du mot *avocat* est *à priori* également posé :

Jean mange un avocat.

Cet exemple ne met pas en place une énumération, comme dans le cas précédent, et pourtant la précision sur la signification du mot *avocat* est identique et semble apparaître même de façon plus flagrante. Nous expliquons cet effet de sens par une dépendance sémantique entre *mange* et *avocat*. Ces deux mots sont ici en relation de dépendance syntaxique directe étant donné que le groupe nominal *un avocat* est l'objet du verbe transitif *manger*. Du point de vue de ses propriétés sémantiques, le verbe manger indique entre autres que son objet est comestible (ou au moins ingérable) c'est-à-dire qu'il a la propriété d'être un aliment. De plus, on a signalé que parmi les significations du mot *avocat*, il y en existe une en terme d'aliment[34]. L'effet de sens qui consiste à répéter cette propriété d'aliment dans le co-texte précise, de fait, la signification d'*avocat* ainsi que la signification du verbe *manger* qui ne peut être ici considéré dans un sens métaphorique. On a donc, dans le cas présent, une influence sémantique entre des mots co-présents et syntaxiquement dépendants, ce qui renforce d'autant plus l'effet de sens produit et c'est donc bien de dépendance sémantique qu'il s'agit.

Dans certains cas même, la contrainte co-textuelle due à la dépendance sémantique provoque un glissement de la signification d'un mot. Ce glissement est une forme particulière de polysémie car le mot en cause prend alors la signification d'un autre mot (sémantiquement proche), et il ne s'agit plus simplement de retenir dans le co-texte une de ses significations potentielles. Considérons l'exemple suivant extrait de [Gosselin 98] :

Il dormit à 10h40.

La dépendance sémantique met en jeu, dans cet exemple, le verbe *dormir* et le circonstanciel *à 10h40*. *Dormir* est un verbe qui sémantiquement dénote un procès duratif et dans le cas présent ce procès duratif est renforcé par une conjugaison au passé simple qui apporte un aspect aoristique (c'est-à-dire qui réfère à une action qui a un commencement). Or, le circonstanciel *à 10h40*, qui d'après la construction syntaxique est intégré au syntagme verbal, précise un aspect ponctuel. Le syntagme verbal est alors le lieu d'une incompatibilité entre des composants ponctuel et non ponctuel. Gosselin montre dans un tel cas, qu'il appelle conflit, un effet de sens qui consiste à déformer le procès évoqué par l'énoncé. Ce procès se trouve contracté sur sa phase initiale ponctuelle. L'effet de sens, en forçant la répétition de la propriété ponctuelle,

[34] Le trait *aliment* dans la signification comme fruit du mot *avocat* est socialement normé. Pour une société cannibale, on pourrait penser que l'autre signification d'*avocat* porte également le trait *aliment*. Dans l'énoncé *Les cannibales ont mangé l'avocat*, l'ambiguité sur *avocat* ne serait donc pas directement levée, comme c'est le cas dans notre exemple.

déforme la signification du verbe *dormir* qui, du coup, emprunte la signification d'un verbe sémantiquement proche, le verbe *s'endormir*.

Selon Tesnière ([Tesnière 59]), une phrase met en place trois choses : un procès (i.e. *ce qui se passe et que la phrase évoque*), des acteurs (i.e. *les personnes ou objets en cause*) et des circonstances (i.e. *comment cela se passe*). Sur le plan syntaxique, cette distinction implique de rechercher trois catégories dans la chaîne syntagmatique, les *verbes*, les *actants* et les *circonstants*. Les rapports syntagmatiques entre un verbe et des actants sont appelés relations actancielles. Elles peuvent être de deux ordres, en fonction notamment de la nature du verbe : soit elles expriment un rapport entre un sujet et son verbe, on parlera alors de relation actancielle sujet, soit lorsque le verbe est transitif, elles expriment un rapport entre un verbe et son (ses) objet(s), on parlera alors de relation actancielle objet.

Le groupe verbal est le lieu de contraintes syntaxiques fortes que des accords en nombre ou en genre expriment, notamment dans la flexion du verbe :

Accord en nombre : *Il**s** pens**ent** que c'est possible.*
Accord en nombre et en genre **:** ***Elle** est tomb**ée** sur la tête.*

Ces accords sont le plus souvent portés par la relation actancielle sujet (sauf dans les cas où le participe passé s'accorde avec le complément d'objet direct). Il n'en reste pas moins que les relations actancielles objet, mêmes si elles de ne donnent pas toujours lieu à de telles traces, produisent des effets de sens entre le verbe et l'objet comparables aux effets de sens entre le verbe et le sujet, comme le montrent les exemples suivants.

1) Dépendance sémantique dans une relation actancielle sujet :

La France a finalement signé les accords.

Le verbe *signer* évoque une activité et donc contraint son sujet à faire référence à l'acteur (voir les acteurs) de cette activité. La signification de *La France* est ici déformée pour satisfaire cette contrainte (c'est un cas de résolution de conflit comme dans le modèle des relations aspectuo-temporelles de [Gosselin 96a]). Cette déformation guide l'interprétant à retenir dans le contexte une signification du type *le ou les représentant(s) de la France.*

2) Dépendance sémantique dans une relation actancielle objet :

J'ai fini de lire Saussure.

Le verbe *lire* est transitif et contraint son objet à renvoyer à un mot, une phrase, un texte, un livre, une oeuvre ou même métaphoriquement à un signe quel qu'il soit (par exemple *lire*

l'avenir dans les lignes de la main). Le co-texte va ici contraindre l'interprétant à déformer de façon métonymique la signification de *Saussure.* L'interprétant ne va pas retenir la signification comme auteur, mais il va retenir une signification en termes de livre ou d'œuvre pour ainsi permettre une répétition dans le co-texte d'un trait sémantique (retranscrivant un livre ou une oeuvre).

La dépendance sémantique des relations actancielles n'est pas toujours aussi simple car, dans certains cas, elle transmet au sujet un effet de sens dû au verbe et à ses circonstants. C'est le cas de l'exemple qui suit (extrait de [Fauconnier 84]) :

Proust est sur l'étagère à droite.

Proust est à comprendre dans cet énoncé comme un livre en tant qu'objet matériel et non comme son auteur (à l'aspect matériel près, c'est presque le même cas que dans l'exemple précédent où *Saussure* était à comprendre comme un livre ou une œuvre). Cet effet de sens provient du circonstant *sur l'étagère à droite* et est propagé par le verbe *être* qui prend ici un sens locatif du fait de la dépendance sémantique entre le verbe et son circonstant. L'effet de sens est donc plus complexe, il est véhiculé par la relation actancielle mais n'en est pas directement issu. La relation actancielle permet de prolonger jusqu'au sujet de la prédication les isotopies initiées par l'effet de sens de la dépendance sémantique entre le verbe et son circonstant.

Comme on vient de le voir, les relations actancielles, en supportant des effets de sens, produisent des phénomènes de dépendance sémantique. Dans les co-textes intraphrastiques, il en va de même des relations de constitution de groupes nominaux.

Le groupe nominal occupe une place particulière dans les études linguistiques et plus particulièrement en syntaxe. En effet, il indique des traits lexicaux comme le genre, le nombre ou la définitude (définis ou indéfinis). Il porte une marque de cas dans sa relation avec le syntagme verbal (sujet ou objet). Enfin, il contient le substantif, c'est-à-dire l'élément qui a la propriété de désigner. Tous ces aspects en font une structure forte dans laquelle les constituants sont liés, que ce soit au niveau morphologique par des accords en nombre ou en genre ou au niveau sémantique par des effets de sens. Ainsi, les relations de constitution des groupes nominaux ont des influences immédiates sur les significations de leurs constituants (et plus généralement du groupe en entier) et ceci avant même de considérer le groupe nominal en relation avec un groupe verbal. L'organisation des mots au sein d'un groupe nominal exerce une contrainte très forte sur l'interprétation de l'interlocuteur.

Chaque élément d'un groupe nominal peut contraindre un autre élément du groupe, comme le montre la glossématique de Hjelmslev ([Hjelmslev 68]) par trois types de relations : l'*interdépendance*, ou implication réciproque (dans *le garçon* la présence de l'article implique

celle du nom, et réciproquement), la *détermination*, ou implication unilatérale (dans *le grand garçon* la présence de l'adjectif implique celle du nom, mais non l'inverse) et la *constellation*, ou absence d'implication (dans *le grand garçon* l'article n'implique pas la présence de l'adjectif et réciproquement)[35]. Le groupe nominal n'est donc pas exclusivement déterminé par le nom et c'est aussi pour cette raison que notre modèle de la dépendance sémantique recherche des effets de sens dus aux co-présences de mots sans envisager d'importance relative d'un mot sur l'autre au niveau sémantique. À titre d'exemple, nous allons ici donner trois cas de groupes nominaux où la signification de l'un des composants est le fait de la co-présence d'un autre :

1) Quand la signification du nom est contrainte par le déterminant :

1.a :	*un poste*	1bis.a :	*le langage*
1.b :	*une poste*	1bis.b :	*les langages*

Dans 1.a, le genre du déterminant tend à identifier le nom dans une signification qui évoque soit un statut dans une organisation (*un poste d'attaché commercial*), soit un appareil ménager (*poste de radio, poste de télévision, poste téléphonique*). Dans 1.b, le genre du déterminant tend à considérer pour le nom une signification renvoyant à un lieu représentant l'administration postale. Dans les exemples 1bis.a et 1bis.b, le déterminant contraint de la même façon la signification du nom, mais, cette fois, la contrainte n'est plus le fait du genre du déterminant mais le fait de son nombre. Ainsi, dans 1bis.a le nom a la signification d'une faculté générale d'une espèce (par exemple *le langage humain* ou encore *le langage chez les dauphins*) tandis que dans 1bis.b, le nom a une signification de systèmes symboliques artificiels, comme par exemple *les langages de programmation* (on retrouve ce même genre d'effet de sens dans l'opposition des groupes nominaux *la langue* et *les langues*).

2) Quand la signification du nom est contrainte par l'adjectif :

2.a :	*le grand bureau*
2.b :	*le bureau exécutif*

Dans 2.a, l'adjectif conduit l'interprétation à retenir une signification du mot *bureau* soit comme un meuble, soit comme une pièce, alors que dans 2.b, l'influence consiste à comprendre *bureau* dans sa signification de groupes de personnes dirigeantes.

3) Quand la signification d'un adjectif est contrainte par le nom :

3.a :	*un grand garçon*
3.b :	*un grand homme*

[35] Les exemples cités ici sont repris de [Francois 80, p. 71-72].

Dans 3.a, le nom *garçon* donne à l'adjectif *grand* un sens lié à la taille ou à l'âge du garçon, tandis que dans 3.b, le nom *homme* donne à l'adjectif *grand* un sens lié à la célébrité.

La contrainte sémantique d'un composant d'un groupe peut s'exercer sur un autre composant (on vient d'en voir des exemples) mais elle peut aussi directement porter sur la signification du groupe nominal dans son ensemble. C'est le cas des exemple suivants :

un ballon
le ballon

Dans le premier exemple, le déterminant indéfini apporte à la signification du groupe nominal soit un caractère générique, soit un caractère indéterminé, alors que dans le deuxième exemple, le déterminant défini amène un caractère d'unicité à la signification du groupe. Cette influence, due au caractère défini ou indéfini du déterminant, est particulièrement pertinente dans une problématique d'établissement de chaînes de co-références.

Plus la structure du groupe nominal est complexe, plus les effets de sens sont potentiellement nombreux et subtils et donc plus la dépendance sémantique est pertinente pour rendre compte des mises en places des significations par la constitution du groupe. On trouve, par exemple, des constructions complexes dans les structures nominales du type *N1 de N2*. La signification du groupe est complètement dépendante de la nature de *N1* et de *N2* et on ne peut pas en rester à une signification prototypique en terme de possession comme le montre [Bartning 95] dans les exemples suivants :

possession : *la voiture de Jean*
attribution : *la gentillesse de Jean*
localisation : *les bibelots du salon*
origine : *le thé de Chine*
iconique : *la photo de Jean*
… (cette liste n'est pas exhaustive et notamment, l'exemple *l'imbécile de Luc* montre une signification du groupe nominal en terme de qualification)

On ne peut donc pas régler le problème de la signification du syntagme nominal en posant des significations prototypiques attachées à des types de séquences syntaxiquement reconnues. Le problème relève intégralement de la sémantique lexicale.

4.2.2. La dépendance sémantique extraphrastique

Les relations de dépendance sémantique qu'entretiennent des lexèmes n'appartenant pas au même énoncé ne sont pas de nature syntaxique. En effet, les relations syntaxiques sont

propres aux énoncés et aux phrases et il n'y a pas de relations syntaxiques qui relient des lexèmes n'appartenant pas aux mêmes énoncés, même si on parle parfois de grammaires de textes. Toutefois, on peut mettre en évidence des relations de dépendance sémantique quand il y a une influence sémantique entre deux lexèmes en relation de co-référence ou d'anaphore.

Deux entités linguistiques A et B sont en relation de co-référence quand l'objet référent de A est le même que celui de B. Dans certains co-textes intraphrastiques, il y a superposition entre deux effets de sens : les dépendances dues aux relations syntaxiques et les influences entre les entités en relation de co-référence. C'est, par exemple, le cas de *Je* et *me* dans l'énoncé *Je me lave* ou encore c'est le cas de *Le président de la république* et de *l'ancien maire de Paris* dans la phrase *Le président de la république est l'ancien maire de Paris*. Dans des cas comme ceux-ci, on pourra rendre compte des effets de sens des mots les uns sur les autres sans avoir recours à la co-référence car les relations de constitution de groupes nominaux, ainsi que les relations actancielles, suffiront pour les mettre au jour dans le co-texte présent (intraphrastique). Ce n'est pas le cas lorsqu'il y a co-référence entre deux mots ou deux syntagmes n'appartenant pas au même énoncé, c'est-à-dire lorsqu'on veut mettre au jour la dépendance sémantique dans les co-textes extraphrastiques, où les formations de chaînes de co-référence traduisent des reformulations ou des reprises des mêmes objets dans le discours. On touche alors principalement au problème des anaphores.

Considérons la reprise d'un nom par un pronom dans deux énoncés qui se suivent. Le contenu sémantique du pronom est complètement dépendant de la signification du nom qu'il reprend (qu'on appelle son antécédent). C'est même sur ce critère de dépendance qu'est basée la définition du pronom, ce que Milner ([Milner 82]) décrit en une absence de référence virtuelle. Considérons l'exemple :

Jean et Marie sont allés se promener. Il avait oublié son portefeuille.

La relation entre *Jean* et *Il* traduit bien ici une contrainte co-textuelle pour la signification de *Il.* Cependant, l'effet de sens induit par la dépendance sémantique de *Jean* sur *il* est particulier. Il ne s'agit pas, comme on l'a vu jusqu'ici, de mettre en évidence des propriétés communes aux deux mots étant donné que, par définition, le pronom n'en a aucune (au niveau sémantique, car au niveau lexical il porte le genre, le nombre et la personne verbale). L'effet de sens consiste alors à propager au pronom les propriétés de son antécédent pour finalement pouvoir aussi rendre récurrentes les propriétés dans le co-texte. C'est un mécanisme dit d'afférence ([Rastier 87]) que nous détaillerons au chapitre 4. C'est le résultat de cette propagation que Milner évoque lorsqu'il signale que le pronom appelle à une référence actuelle en contexte.

De même, dans les cas de reprise d'un syntagme nominal A par un syntagme nominal B, on peut mettre en évidence une dépendance sémantique entre A et B. Dans l'exemple suivant,

Il a acheté une glace. C'est son dessert préféré.

la relation entre *une glace* et *son dessert* contraint la signification de *glace* qui notamment ne peut plus faire référence à une vitre ou à un miroir. C'est en cela que la dépendance sémantique exprime des contraintes sur les significations des lexèmes (alors que la simple influence sémantique dénote une contribution du co-texte pour la signification d'un lexème).

Ces formes de reprises d'un syntagme par un pronom ou par un autre syntagme constituent des anaphores. La première définition de l'anaphore renvoie à une figure de style de la rhétorique qui décrit une reprise d'un mot, ou d'un groupe de mots, en début de phrases successives. Dans la définition de la notion d'anaphore, aujourd'hui communément admise en linguistique, on retrouve encore cette idée de reprise, mais elle n'est plus littérale, elle concerne la signification et la référence des mots en cause dans la relation d'anaphore. Ainsi, une expression est dite anaphorique lorsque son interprétation est dépendante d'une autre, présente dans son co-texte, ce que Milner exprime par :

> "Il y a relation d'anaphore entre deux unités A et B quand l'interprétation de B dépend crucialement de l'existence de A, au point qu'on peut dire que l'unité B n'est interprétable que dans la mesure où elle reprend, entièrement ou partiellement, A." ([Milner 82, p. 18])

exemples :

(a) ***La veste*** *ne m'a pas coûté cher,* ***elle*** *était en solde.*

(b) ***Le mardi*** *était pluvieux,* ***le lendemain*** *était ensoleillé.*

Dans (a), il y a anaphore entre *la veste* et *elle* et en plus il y a co-référence, la reprise est donc "totale", pour reprendre les mots de Milner. Dans (b), il y a anaphore entre *Le mardi* et *le lendemain* et la reprise est "partielle" étant donné qu'il n'y a pas co-référence.

Le problème central de l'anaphore est que certains mots de la langue ne se suffisent pas pour déterminer leur signification. On les appelle des anaphoriques depuis les travaux de Tesnière ([Tesnière 59]). C'est en quelque sorte un problème dual de la polysémie où un mot polysémique a, à lui seul, plusieurs significations alors qu'un anaphorique, à lui seul, n'arrive même pas à s'en constituer une. Cette non autonomie sémantique est ce qui permet de distinguer les co-références non anaphoriques des anaphores co-référentielles car, dans le cas d'une co-référence, les deux entités sont sémantiquement autonomes. C'est, par exemple, la différence entre *Je me lave*, où l'on a mis au jour une co-référence, et *Il se lave*, où il y a anaphore entre *Il* et *se* étant donné que le pronom réfléchi *se* est sémantiquement dépendant de ce à quoi il se rapporte. De ce point de vue, [Charolles 91] rapporte l'anaphore à un phénomène général de dépendance interprétative et c'est dans cette perspective qu'on cherche à rendre compte des anaphores au moyen de la dépendance sémantique.

Toutes les anaphores ne sont pas co-référentielles. Les anaphores non co-référentielles sont appelées des anaphores associatives. Un exemple classique est :

Nous entrâmes dans ***un village***. ***L'église*** *était située sur une butte.* ([Kleiber 95])

On a ici une anaphore associative entre *un village* et *L'église* car la référence de l'église ne peut être trouvée que dans un rapport à la référence du village étant donné une relation métonymique entre les objets référents. Toutefois, l'anaphore associative n'apporte pas une contrainte uniquement référentielle, mais les significations des pôles de la relation anaphorique sont aussi interdépendantes. Par exemple, dans l'énoncé suivant, on a une anaphore associative (et nominale) entre les syntagmes *l'arbre* ("l'anaphorisé") et *le tronc* ("l'anaphorisant"):

Tu reconnaîtras l'arbre, le tronc est imposant.

Les mots *arbre* et *tronc* sont polysémiques car notamment *arbre* peut avoir une signification de végétal comme une signification de type de structure (comme, par exemple, les arbres généalogiques) et *tronc* peut avoir une signification comme tige rigide ou une signification comme boîte servant à faire l'aumône. Dans la relation anaphorique entre *l'arbre* et *le tronc*, on met en évidence une dépendance sémantique qui permet de réduire les ambiguïtés dûes aux polysémies lexicales des deux mots. Ainsi, dans notre exemple, la dépendance sémantique amène l'interprétant à retenir la signification d'un végétal pour *arbre* et la signification de tige rigide pour *tronc*.

Parmi les anaphores co-référentielles on peut faire une distinction entre des anaphores nominales, quand la reprise anaphorique est un groupe nominal, et des anaphores pronominales[36] quand la reprise anaphorique est un pronom :

Anaphore nominale :	*Ce matin,* ***un chasseur*** *s'est blessé.* ***Le jeune homme*** *a été emmené à l'hôpital.*
Anaphore pronominale :	*Ce matin,* ***un chasseur*** *s'est blessé.* ***Il*** *a été emmené à l'hôpital.*

Dans les deux cas le rapport à l'effet de sens de la dépendance n'est pas le même. L'anaphore nominale tend à uniformiser les traits sémantiques portés par les deux groupes nominaux (comme c'est aussi le cas pour l'anaphore associative) et l'anaphore pronominale, pour réaliser une récurrence (i.e. une isotopie), fait propager au pronom les traits sémantiques du groupe nominal (i.e. l'anaphore pronominale réalise une afférence).

Dans une étude pragmatique de la co-référence (qui sort du cadre de notre thèse), les deux anaphores n'ont pas non plus la même fonction au regard de la prise en compte de l'identité du

[36] Cela ne veut pas pour autant dire que toutes les anaphores pronominales sont co-référentielles, comme le montre l'exemple suivant extrait de [Kleiber 90] : *J'ai acheté* ***une Toyota*** *parce qu'****elles*** *sont robustes.* L'anaphore est ici associative.

référent dans la prolongation de la continuité référentielle. Selon [Kleiber 93], le syntagme nominal qui introduit un référent a deux fonctions : la première est de décrire le référent par un contenu descriptif qui fournit des prédicats sortaux pour permettre de l'identifier, la deuxième est simplement de dénommer le référent. Au regard de la première fonction du syntagme nominal, l'anaphore nominale, qui introduit un nouveau syntagme pour un référent, a pour conséquence d'apporter dans le contexte de nouveaux prédicats sortaux pour ce référent. L'effet pragmatico-référentiel de l'anaphore nominale consiste alors à préciser l'identité catégorielle du référent. Du point de vue de l'interaction, c'est une trace linguistique d'une co-désignation ou d'une co-construction d'une catégorie pour l'objet (qui résulte de la monstration de l'énoncé) que les interlocuteurs évoquent. L'anaphore pronominale n'a pas du tout ce même effet pour la simple raison que le pronom, à l'inverse du syntagme nominal, n'apporte aucun nouveau prédicat sortal pour la construction de l'identité de l'objet. Effectivement, le pronom qui porte des marques de genre et de nombre n'est que très rarement informatif sur son référent, comme dans l'exemple qui suit.

Le premier ministre *nous a rendu visite,* ***elle*** *est charmante.*

L'anaphore pronominale n'apporte donc, en général, pas d'informations nouvelles sur le référent. Son effet sur l'identité n'est plus une construction de cette identité mais son maintien en tant qu'objet de la monstration dans le discours, c'est-à-dire le maintien de son identité individuelle ([Beust & al. 97a]). Dans un bon nombre de cas, l'anaphore pronominale est le moyen de maintenir cette identité individuelle, ce qui fait que le pronom n'est pas une simple reprise de son antécédent. Effectivement, on remarque dans les exemples qui suivent (extraits de [Charolles & al. 93]) que si on remplace le pronom anaphorique par son antécédent, la co-référence n'apparaît plus aussi clairement, voire n'est plus possible du tout et le sens de l'énoncé s'en trouve radicalement changé.

Paul *croit qu'****il*** *a de l'avenir.*
* ***Paul*** *croit que* ***Paul*** *a de l'avenir.*

Tous les concurrents *espèrent qu'****ils*** *vont gagner.*
* ***Tous les concurrents*** *espèrent que* ***tous les concurrents*** *vont gagner.*

Jean a épousé ***une suédoise****. En fait* ***elle*** *est danoise.*
* *Jean a épousé* ***une suédoise****. En fait,* ***une suédoise*** *est danoise.*

Nous avons vu jusqu'ici des cas où la continuité référentielle dans le discours n'est pas problématique dans la mesure où l'objet est stable dans le déroulement temporel de l'interaction. Dans certains cas, les choses sont plus compliquées car l'objet subit des modifications pendant la narration, où la monstration langagière s'instaure sous forme de chaînes de co-référence. Ce sont les cas dits de référents évolutifs dans la littérature linguistique :

> *Prenez **quatre pommes**. Pelez-**les**, coupez-**les** et évidez-**les**. Faites-**les** cuire pendant une 1/2 heure, broyez-**les** jusqu'à ce qu'**elles** soient complètement réduites et, après **les** avoir laissées refroidir, servez-**les** avec des petits gâteaux* (exemple extrait de [Charolles & al. 93])

D'un point de vue lexical, le calcul de l'antécédent des pronoms anaphoriques (*les, elles*) ne pose pas plus de problèmes que dans les cas classiques mais, d'un point de vue pragmatique, la question de la nature de l'objet dans la relation de référence est beaucoup plus délicate. Peut-on encore, par exemple, penser que le référent à la fin de l'énoncé est bien le même que celui du début ? On est bien là face à un problème d'identité. Dans l'exemple précédent, il paraît effectivement impossible dans la fin de l'énoncé d'utiliser le syntagme nominal *quatre pommes* à la place du pronom *les.* Dans ces cas de référents évolutifs, l'apparition d'un nouveau substantif est fréquente et précise une nouvelle identité catégorielle pour le référent. C'est le cas du syntagme *la compote* dans l'exemple de [Charolles & al. 93] cité ci-dessus et légèrement modifié :

> *Prenez **quatre pommes**. Pelez-**les**, coupez-**les** et évidez-**les**. Faites-**les** cuire pendant une 1/2 heure, broyez-**les** jusqu'à ce qu'**elles** soient complètement réduites et, après **les** avoir laissées refroidir, servez **la compote** avec des petits gâteaux.*

En plus des problèmes de continuité référentielle, les référents évolutifs posent des problèmes cognitifs de continuité substantielle. La langue ne peut rendre compte d'une évolution continue de l'état du référent car l'organisation de son système, étant entre autre fondée sur la perception et la catégorisation, procède à une discrétisation des états du monde. La transformation même du référent échappe alors souvent à la verbalisation du phénomène tant que l'objet ne se trouve pas reconnu comme élément d'une nouvelle catégorie. Ceci favorise

- l'emploi d'anaphores pronominales où l'on se raccroche uniquement à l'identité individuelle de l'objet,
- l'usage de déictiques,
- les ellipses complètes de l'objet.

Ceci est particulièrement frappant dans des situations de consignes où ce n'est pas celui qui "dit" qui "fait" et où on assiste souvent à des référenciations directes au processus en train de se faire plutôt qu'aux objets qui sont difficilement appréhendables ([Vivier 92]). C'est ainsi, par exemple, que dans des recettes de cuisine on trouve souvent des séquences comme "laissez cuire à feux doux", "mélanger jusqu'à obtenir une substance homogène" ou encore "salez, poivrez ...".

Dans son étude des référents évolutifs, Tyvaert ([Tyvaert 95]) montre l'importance de la prédication pour la portée dans le discours de la continuité référentielle. C'est, selon lui, une ou plusieurs prédications qui bloquent la possibilité de rementionner la dénomination initiale de l'argument prédiqué, mais c'est également la prédication qui permet d'introduire dans le contexte de nouveaux substantifs pour l'objet en cours de mutation. Ainsi, dans l'exemple suivant, le syntagme *La fonte* est directement interprété comme nouveau substantif pour *le fer* grâce à la prédication précédente *se transforme en fonte*.

> *... Le fer résultant de cette réduction se transforme en fonte sous l'action du milieu fortement carburant. La fonte en fusion, séparée du laitier, est recueillie dans le creuset inférieur de l'appareil. ...* ([Tyvaert 95])

Dans un cas de référence évolutive, le nouveau substantif ne se réduit pas forcément au syntagme mais il inclut, au même titre, la prédication qui autorise l'emploi du syntagme en question. Tyvaert insiste donc sur la prise en compte de l'apport de la prédication dans la construction de la référence, et ce d'autant plus dans les contextes particuliers de référents évolutifs. C'est précisément ce que l'on vise en cherchant simultanément les relations de dépendance sémantique dues aux relations actancielles du verbe (dans les co-textes intraphrastiques) et les relations de dépendance sémantique dues aux anaphores (dans les co-textes extraphrastiques). Dans de tels cas, les dimensions sémantiques de monstration et de prédication se recoupent.

4.3. La dépendance sémantique généralisée

Nous avons détaillé les bases syntaxiques et anaphoriques retenues pour chercher à mettre en évidence les effets de sens produits par la dépendance sémantique dans les co-textes intraphrastiques et extraphrastiques. Si les énoncés isolés que nous avons considérés jusqu'ici sont de bons révélateurs d'un seul type de dépendance sémantique, dans le co-texte, les trois formes de dépendance (qui proviennent de la constitution des groupes nominaux, des relations actancielles et des anaphores) se combinent. En particulier dans les interactions langagières, les énoncés sont le lieu d'une poly-dépendance sémantique, c'est-à-dire une dépendance sémantique généralisée en réponse à la *polysémie contextuelle généralisée* que Gosselin met au jour ([Gosselin 96b]). Considérons, par exemple le petit dialogue fictif suivant.

A : *Hier, Pierre a acheté un livre*
B : *Oui, je sais, il l'a même déjà lu*

Dans cette courte interaction, on peut déjà mettre en évidence dans le co-texte huit relations de dépendances sémantiques :

a) celle due à la constitution du groupe nominal *un livre*
b) celle due à la relation sujet-verbe *Pierre* <-> *a acheté*
c) celle due à la relation verbe-objet *a acheté* <-> *un livre*
d) celle due à la relation sujet-verbe *je* <-> *sais*
e) celle due à la relation sujet-verbe *il* <-> *a lu*
f) celle due à la relation verbe-objet *a lu* <-> *l'*
g) celle due à la co-référence *il* <-> *Pierre*
h) celle due à la co-référence *l'* <-> *un livre*

Ces huit dépendances ne sont pas indépendantes les unes des autres et elles se renseignent mutuellement à travers la constitution d'isotopies qui rendent compte des effets de sens co-textuels. Ainsi, les mises en évidence des relations **g)** et **h)** proviennent directement des résultats des autres dépendances sémantiques. Effectivement, dans ce cas, il n'est pas nécessaire d'avoir recours à un calcul purement linguistique des antécédents anaphoriques des pronoms *il* et *l'* car les effets de sens des autres dépendances éclairent suffisamment la dimension sémantique du calcul des anaphores. Par exemple, la relation **e)** construite sur la relation actancielle sujet entre le pronom *il* et le verbe *lire* crée un effet de sens par la répétition dans le contenu sémantique du pronom, d'un trait que porte le contenu sémantique du verbe pour caractériser son sujet. Ce trait sémantique exprime notamment que le référent du sujet du verbe *lire* appartient à la catégorie des humains. On peut le noter pour l'instant /humain/ (on proposera dans le chapitre suivant une autre représentation des traits sémantiques). Du fait de cette dépendance sémantique, le contenu sémantique du pronom n'est plus vide car il contient le trait /humain/. Ce trait permet ici de résoudre une recherche de co-référence dans le co-texte étant donné que le seul substantif déjà évoqué contenant ce trait est *Pierre*. Cette co-référence permet de mettre au jour la dépendance sémantique **g)** dont l'influence consiste à insister sur les propriétés communes de *il* et de *Pierre*, c'est-à-dire conforter dans le co-texte l'isotopie du trait /humain/ préalablement initiée par la dépendance sémantique **e)**. Le calcul de la dépendance sémantique **h)** s'effectue de la même façon et réalise un prolongement d'une isotopie du trait /objet/.

Un tel exemple nous indique un ordre de calcul pour un processus opératoire de recherche de dépendances sémantiques (nous y reviendrons dans le chapitre n°5) : il convient, en premier lieu, de rechercher les dépendances intraphrastiques et, en second lieu, les dépendances extraphrastiques. Dans la recherche des dépendances intraphrastiques, nous chercherons premièrement à rendre compte des dépendances sémantiques dûes à la constitution des groupes nominaux puis ensuite à rendre compte des dépendances sémantiques dûes aux relations actancielles. On réalise ainsi un ordre dans la recherche des dépendances allant du local vers le

global. On donne alors son importance à la *proximité* entre les constituants de la chaîne dont Lakoff et Johnson soulignent les effets :

> Si la signification d'une forme A exerce une influence sur la signification d'une forme B, alors plus la forme A est *proche* de la forme B, plus *grand* sera l'effet de la signification de A sur la signification de B. ([Lakoff & al. 85, p. 139]).

5. Conclusions

Dans ce chapitre, nous avons voulu montrer certains phénomènes sémantiques spécifiques aux langues naturelles, et plus précisément, des phénomènes propres à la co-détermination des significations des composants d'une chaîne linguistique. C'est parce que la polysémie est omniprésente dans les langues naturelles que le sens d'un énoncé ne peut être calculé de façon purement compositionnelle comme c'est le cas dans la compilation des langages formels. De ce point de vue, nous retenons une approche holiste pour une modélisation computationnelle de la sémantique des langues naturelles.

Dans cette approche de la sémantique, nous nous limiterons à l'analyse de la fonction de monstration des énoncés. Dans le co-texte que forme un énoncé ou une suite d'énoncés, nous cherchons à mettre en évidence les effets de sens dus à la co-présence de lexèmes dans la chaîne et nous nous focalisons plus particulièrement sur les cas où cette co-présence est renforcée par une relation syntagmatique, c'est-à-dire les cas de dépendance sémantique. Il y a deux raisons à cela : premièrement parce que ce sont les cas les plus fréquents et deuxièmement parce que, dans une perspective calculatoire, ils sont prédictibles (notamment par les résultats d'une analyse syntaxique). Les effets de sens que l'on met ainsi en évidence sont considérés comme des contraintes du matériau linguistique qui participent à la mise en forme du potentiel de sens de l'énoncé.

La formulation dans une perspective informatique de la co-présence syntagmatique en termes d'influence et de dépendance sémantiques exige la représentation en machine des contenus sémantiques des lexèmes, c'est-à-dire des significations des mots. Il nous faut donc proposer un modèle sémantique de l'axe paradigmatique. Ce modèle fait l'objet du chapitre suivant. Comme l'a montré la psychologie développementale, l'acquisition des significations est un processus interactif chez les humains car le langage est construit dans et par l'activité dialogique ([Vivier 93]). Dans la construction du modèle sémantique de l'axe paradigmatique, nous placerons l'acquisition des significations dans un processus interactif entre l'humain et la machine, il ne s'agira pas encore d'une interaction dialogique à proprement parler mais d'un processus interactif de co-construction de systèmes de signes.

L'étude de l'influence et de la dépendance sémantiques laisse déjà apparaître une notion centrale de notre modèle, la notion d'isotopie. Nous la détaillerons au chapitre 4 quand nous expliquerons son apport calculatoire au niveau syntagmatique pour rendre compte de la mise en co-texte des représentations paradigmatiques.

CHAPITRE 3 :

UN MODÈLE SÉMANTIQUE DE L'AXE PARADIGMATIQUE

En définissant la dépendance sémantique dans le chapitre précédent, nous avons défendu la position qui consiste à considérer que la variabilité contextuelle des significations est tout à fait première dans la construction du sens. En effet, l'idée de base de la dépendance sémantique (et plus généralement de l'influence sémantique) est qu'on ne peut pas observer la signification d'un mot en tant que telle, mais que cette observation passe obligatoirement par une contextualisation, donc par un effet de sens. C'est-à-dire que les effets de sens sont la seule manifestation des significations[37]. Le calcul du sens par l'étude de la dépendance sémantique s'articule alors suivant deux plans : le premier consiste à représenter les significations en tant qu'elles permettent d'initier ou de renforcer des effets de sens et le second est de mettre en évidence l'émergence d'effets de sens au sein de l'énoncé. Ce deuxième plan est l'objet d'une modélisation de la participation syntagmatique à la construction du sens, c'est-à-dire une sémantique de l'énoncé. Nous la détaillerons au chapitre suivant. Le premier plan, quant à lui, constitue un modèle sémantique de l'axe paradigmatique, c'est-à-dire une sémantique lexicale. Elle constitue l'objet du présent chapitre.

Dans la construction d'une sémantique lexicale, la notion de paradigme (plus précisément de paradigme de significations) est centrale car elle donne un cadre à la représentation des proximités entre valeurs sémantiques. Ainsi, deux expressions sont sémantiquement proches quand on peut, dans un énoncé, remplacer l'une par l'autre en conservant plus ou moins le

[37] Meillet ([Meillet 52]) pousse cette position à l'extrême en posant que les mots n'ont finalement pas de sens mais uniquement des emplois.

même sens. On dit, dans ce cas, que ces deux expressions prennent place au sein du même paradigme. Mais un paradigme est beaucoup plus qu'une simple collection d'éléments. Il comporte une structure interne qui organise ses éléments par des relations particulières et [Ducrot & al. 95, p. 231] lui donnent de ce fait le statut de *catégorie linguistique.*

Les relations qu'entretiennent les éléments d'un paradigme sont, d'un point de vue saussurien, celles qui organisent les significations en un système de valeurs. Les paradigmes ne sont pas indépendants les uns des autres et ils sont notamment reliés par les relations de polysémie car un mot polysémique peut placer ses différentes significations au sein de différents paradigmes. Ainsi, le paradigme peut être considéré comme une zone (locale) dans le système global de la langue. Ce que l'on cherche donc dans un modèle sémantique de l'axe paradigmatique, c'est un modèle le plus global possible de la valeur saussurienne.

La référence à Saussure est, comme on l'a signalé dans les chapitres précédents, un de nos premiers choix théoriques dans la modélisation de la signification des mots. Un second choix est le rapport d'un système de valeurs sémantiques à l'activité de catégorisation. Dans le modèle que nous proposons, la catégorisation est centrale. Ce sont les propriétés des catégories qui permettent de constituer un système où chaque constituant vaut par la place qu'il y occupe, et ce sont les rapports de ces propriétés à un point de vue sur le monde qui permettent d'exploiter les significations en terme d'indices pour la construction des référents du discours.

Dans la première partie de ce chapitre, nous placerons notre approche dans le cadre de la sémantique componentielle ([Pottier 74], [Greimas 66], et plus récemment [Rastier 87]), car les significations seront construites par composition de traits sémantiques. Cependant, les contraintes de rigueur et d'effectivité qui s'exercent sur la modélisation computationnelle font que, dans la plupart des cas, une théorie linguistique descriptive n'est pas directement implémentable en machine. Pour l'interprétation d'énoncés langagiers, ces contraintes nous amènerons à redéfinir la notion de trait sémantique.

Après avoir détaillé nos choix pour la modélisation de la signification des mots, nous proposerons, dans la deuxième partie du chapitre, un modèle de catégorisation fondé sur des relations de différences entre les choses. Ce modèle de catégorisation générique est appelé *Anadia*. Nous expliquerons comment ses principes peuvent être utilisés à des fins de description componentielle des significations des lexèmes.

L'aspect fondamentalement interactionniste de la catégorisation au sein de notre modèle en fait une structure de représentation qui ne peut connaître d'état final car la catégorisation, qui met en place le système de valeurs, est un processus sans fin où les choses ne sont jamais données

de façon exhaustive et où elle peuvent toujours être remises en cause par de nouvelles interactions. Ceci apparaît d'autant mieux quand on examine les processus de catégorisation en rapport à l'activité langagière, car c'est dans et par le dialogue que les représentations des agents évoluent. L'effectivité du modèle sera donc sa capacité à évoluer au fur et à mesure des analyses et il n'y aura pas de phase préparatoire d'apprentissage des données préalable à leur mise en œuvre. L'idée est plutôt de concevoir un système avec une amorce et une boucle d'interaction-apprentissage où l'on commence à exploiter le modèle avec presque aucune connaissance, mais où l'interaction avec l'utilisateur est source d'apprentissage (nous reviendrons plus en détail sur cet aspect d'apprentissage dans les cinquième et sixième chapitres).

1. Modéliser la signification dans une approche componentielle

Dans un bon nombre de théories linguistiques, et plus particulièrement chez les structuralistes, on voit apparaître la notion de trait. Par exemple en syntaxe, ce terme désigne toute propriété syntaxique appartenant à la classe ([Pottier 73, p. 525]), et en phonologie, il désigne un critère de distinction entre deux phonèmes (on l'appelle alors *phème*). Ainsi, le nombre et le genre sont des traits syntaxiques et le voisement et l'aperture sont des traits phonologiques. Cette approche où l'on décrit un "objet" par un ensemble de traits qui le caractérisent est une approche dite **componentielle**. On la retrouve également en sémantique pour caractériser les significations (elle s'oppose par exemple aux sémantiques formelles qui conçoivent la signification de façon atomique ou encore aux modèles connexionnistes où la signification n'est pas explicitement représentée). À l'instar de la terminologie de la phonologie, le trait sémantique est alors appelé *sème* et il se définit comme une unité minimale de signification qui, de même que le phème, n'est pas susceptible de réalisation indépendante.

En sémantique, on peut se placer dans le cadre d'une approche componentielle en considérant la construction de la signification selon deux méthodes qui se distinguent par le choix et la fonction des sèmes.

• La première méthode est référentielle. Elle consiste à définir le sème comme une qualité du référent. Cette méthode, appelée méthode *sémasiologique*, est utilisée dans la lexicographie et part de l'expression pour exprimer le contenu. Selon cette méthode, on décomposerait, par exemple, la signification de l'expression "voiture" en donnant les sèmes /quatre roues/[38], /un moteur/, /un châssis/ ...

• La seconde méthode est différentielle. Elle définit les sèmes comme des critères de différenciation entre les significations d'une même classe sémantique (le taxème). C'est la

[38] Il est d'usage en sémantique componentielle de noter les sèmes entre //.

méthode dite *onomasiologique*. Selon cette méthode, la signification de *voiture* dans la classe des moyens de transport comporterait notamment le sème /sur route/ (par différence avec la signification de *train*), ou encore le sème /privé/ (par opposition avec la signification de *bus*)[39].

Comme nous rejetons une vision référentielle de la sémantique des langues naturelles, nous rejetons par conséquence une description componentielle sémasiologique et nous nous plaçons dans le cadre d'une sémantique componentielle et différentielle. L'approche retenue est en cela profondément différente de celle de Schank, par exemple, qui a développé un système en intelligence artificielle de représentation componentielle (*"dépendances conceptuelles"*, [Schank 72]) fondé sur des relations entre des objets et des actions (de fait, Schank se place dans le cadre d'une approche sémasiologique). À la différence de la méthode sémasiologique qui restreint la sémantique à une ontologie, la méthode onomasiologique permet de fonder la sémantique en tant que système de différences. L'idée de la langue comme un système différentiel des signes est formulée en premier par Saussure ([Saussure 1915]) et elle est reprise ensuite par le courant de la linguistique structurale, tant sur le plan de l'expression (le signifiant) que sur le plan du contenu (le signifié). On doit cette opposition entre le plan du signifiant et le plan du signifié à Hjemslev ([Hjemslev 68]) qui y voit deux dimensions de la conception différentielle car un signifiant n'est signifiant que parce qu'il s'oppose à d'autres signifiants possibles, et il en va de même pour les signifiés.

Au niveau de l'expression, l'approche différentielle a permis de fonder la phonologie en tant qu'étude des sons d'une langue du point de vue de leur fonction dans le système (sons alors appelés *phonèmes*), en opposition à la phonétique qui étudie les sons en eux-mêmes et de façon indépendante de leur fonction. Ainsi, Jakobson dans son étude de la phonologie du français contemporain ([Jakobson 63]) montre que les phonèmes sont le fait d'un système organisé par sept traits de distinction, sept phèmes, opposant les phonèmes entre eux (il s'agit de l'aperture, l'articulation buccale, le voisement, la nasalité, l'antériorité, l'arrondissement et l'ouverture). Les phonèmes occupent dès lors dans la langue des places potentielles prédéfinies par l'organisation des phèmes, ce que montre [Coursil 92] par la *topique des phonèmes*. En fonction de la langue étudiée, toutes les places de ce système ne se réalisent pas forcément en un phonème réel. La question de la réalisation d'un phonème est liée à sa nature signifiante dans la langue en question, que l'on peut mettre au jour par le test de commutation dit *des paires minimales*. Par exemple, en français, le voisement permet d'opposer les deux phonèmes */p/* et

[39] La méthode onomasiologique, en confrontant des significations, induit des représentations en tableaux (par exemple, le tableau qui indique les différences entre les *voitures*, les *trains* et les *bus*) alors que la méthode sémasiologique, en décomposant les significations, induit des représentations en arbres (par exemple l'arbre des *véhicules* comportant les branches *voiture*, *bus* et *train*). De ce point de vue, l'arbre de Porphyre est caractéristique d'une approche sémasiologique.

/b/ qui sont tout deux réalisés car ils permettent de différencier les mots *pain* et *bain.* En revanche, le sons /i:/ (le "i" long) et /i/ (le "i" court) ne forment qu'un seul phonème en français car il ne différencient aucune paire minimale, ce qui n'est notamment pas le cas en anglais, où ils constituent bien deux phonèmes distincts qui permettent, par exemple, de différencier *ship* de *sheep.*

1.1. Principes généraux de la sémantique différentielle

Les théories du signifié en linguistique structurale ont été très fortement inspirées par les résultats systémiques obtenus par la phonologie. Suivant une même approche componentielle et différentielle, l'analyse sémique définit, par analogie avec les phèmes de la phonologie, les sèmes comme des unités minimales de la signification. Selon le même principe de combinaison qu'en phonologie, les sèmes forment des *sémèmes*, équivalents sémantiques des phonèmes. Les sémèmes offrent des représentations componentielles des signifiés (validés par un test de commutation) qui prennent place dans un système que les différences entre sèmes organisent. Un exemple classique d'analyse sémique est celui des sièges de [Pottier 74, § 107, p. 99] où la combinatoire des sèmes /avec dossier/, /pour plusieurs personne/ et /avec bras/ permet de mettre en évidence un système où prennent place les sémèmes 'chaise', 'fauteuil', 'canapé' et 'tabouret'.

	/avec dossier/	/pour plusieurs personne/	/avec bras/
'fauteuil'	+	-	+
'tabouret'	-	-	-
'chaise'	+	-	-
'canapé'	+	+	+

Le sème est le concept de base d'un premier palier de la sémantique, palier qui se limite à la modélisation paradigmatique des signifiés : la microsémantique[40]. La microsémantique a pour limite supérieure la lexie ([Pottier 92]), définie comme un groupement stable de morphèmes constituant une unité fonctionnelle. La lexie est l'unité minimale d'expression (son signifié est la *sémie*). Dans la majeure partie des cas elle correspond au mot mais inclut également les mots composés (ainsi *pomme de terre* est une lexie autant que *poireau*). La lexie élargit donc la notion de mot qui reste très difficile à définir clairement et représente tout signifiant d'un signe linguistique repéré par l'interprète.

[40] Elle se distingue de la mésosémantique qui traite du sens de la phrase et de la macrosémantique qui a rapport au sens des groupes de phrases (textes, récits, ...)(cf. [Rastier & al. 94]).

Dans l'approche différentielle, la microsémantique se donne un double but : une étude componentielle des significations et une étude des systèmes de significations. On retrouve bien là un point de vue structuraliste saussurien : la signification d'un signe linguistique est liée à sa place dans un système. On touche ici à la double fonction du sème : apporter une information sémantique pour un signifié et différencier ce signifié par rapport à d'autres. Plus précisément, [Tanguy 97, p. 42] évoque trois rôles pour un sème :

• La cohérence[41] : les sèmes doivent être la projection sur une unité lexicale de considérations sémantiques globales. Ils devront avoir un rôle organisateur et garantir une uniformité dans la description des contenus sémantiques.

• La relativisation : les sèmes ne doivent pas être utilisés pour capter le sens d'une unité mais pour traduire une interprétation de celle-ci.

• La justification : les sèmes doivent traduire les distinctions et donc aussi les similarités entre les significations des unités lexicales.

La cohérence et la justification sont les raisons qui font que la microsémantique n'envisage pas le signe linguistique de façon isolée mais qu'elle cherche à mettre en évidence un système de signification. La relativisation, quant à elle, permet de préciser une caractéristique importante de la microsémantique : en réponse à la question du nombre des sèmes dans le système de la langue, la relativisation indique que ce nombre ne peut être déterminé de façon préalable. De plus, la question du nombre de sèmes dans les langues amène naturellement la question des universaux : y-a-t-il des sèmes suffisamment généraux pour être des "atomes de signification" dans toutes les langues naturelles ? Ces sèmes, appelés *noèmes*, sont l'objet d'une étude philosophique et sociologique du langage. Cette étude sort du cadre de la linguistique, et plus encore du cadre de la microsémantique. La microsémantique ne peut donc se donner comme but de chercher tous les sèmes possibles qui organisent la langue en système, mais seulement ceux qui résultent de l'interprétation d'un texte ou d'énoncés proférés en contexte[42]. On ne retient donc pas l'approche de Guiraud ([Guiraud 65]) qui cherche à rendre compte d'un système fini de primitives pouvant couvrir toutes les significations. De plus, cet objectif nous semble inadapté à une modélisation de la sémantique d'une langue, car cela revient à vouloir la figer dans un unique plan synchronique défini par une combinatoire en dehors de laquelle elle ne pourrait plus évoluer. Elle en perdrait donc sa caractéristique essentielle de système évolutif. Plus grave encore, considérer que la sémantique des langues est construite sur un ensemble fini de primitives (comme c'est le cas dans [Schank 72]), c'est la concevoir comme une axiomatique, c'est-à-dire comme étant uniquement compositionnelle et isomorphe à un langage

[41] Il s'agit ici d'une cohérence paradigmatique. Nous serons amenés, dans le chapitre 4, à définir une autre notion de cohérence : une cohérence syntagmatique. Notamment, nous l'opposerons à la notion de cohésion.
[42] D'où la nécessité de rapporter la pertinence de représentations componentielles à une analyse de corpus.

formel. La relativisation est donc un argument de plus pour souligner les différences de fond entre les langues naturelles et les langages formels.

Les descriptions componentielles des signifiés des morphèmes, c'est-à-dire les sémèmes, constituent des ensembles de sèmes. Par exemple, on pourrait définir le sémème 'mang-'[43] correspondant au signifié du radical du verbe *manger* par la réunion des sèmes suivants /activité vitale d'absorption de nourriture/, /caractéristique des êtres animés/, /objet comestible/, /source de plaisir/, ... Cependant, pour avoir sa place au sein d'un sémème, un sème doit remplir une des deux conditions suivantes :

• Soit il participe à une différence entre le sémème dans lequel il apparaît et un autre sémème, sémantiquement proche (c'est en cela que l'approche est différentielle). Par exemple, on justifiera le sème /bon/ dans le sémème 'souhait-' (radical de *souhaiter*), formé des deux sèmes /absence/ et /bon/, par le fait que s'il est remplacé par /mauvais/, on décrit alors un autre sémème, 'redout-' (/absence/ + /mauvais/), traduisant le signifié du radical du verbe *redouter* ([Ducrot & al. 95, p. 445]).

• Soit il indexe le sémème dans lequel il apparaît dans une classe sémantique. Par exemple, en reprenant l'analyse des sièges de Pottier, on peut considérer que chacun des sémèmes 'chaise', 'fauteuil', 'canapé' et 'tabouret' contiennent le sème /pour s'asseoir/. Ce sème ne fait pas de différences entre les différents sémèmes (étant donné qu'il est présent dans chacun d'eux) mais il permet de les indexer dans une même classe sémantique, en l'occurrence celle des sièges, où d'autres sèmes permettent de les différencier deux à deux. En sémantique différentielle, on dira alors qu'un tel sème est hérité dans le sémème de son type. Ce type étant lui même un sémème (dans notre exemple, le sémème 'siège'), le sème hérité est finalement aussi justifié par des relations de différences au niveau de la classe du sémème type (par exemple /pour s'asseoir/ permet de différentier 'siège' de 'table').

Selon leur fonction dans leur sémème, on distingue deux sortes de sèmes : des *sèmes spécifiques* et des *sèmes génériques*. Les sèmes spécifiques sont ceux qui permettent de différencier le sémème dans sa classe et les sèmes génériques sont les sèmes non spécifiques d'un sémème, c'est-à-dire que ce sont ceux qui ne le différencient pas directement d'un autre sémème mais qui sont hérités d'autres sémèmes. On définit alors, dans la terminologie de la sémantique différentielle, deux composantes du sémème : le *sémantème* représentant l'ensemble des sèmes spécifiques d'un sémème, et le *classème* représentant l'ensemble des sèmes génériques d'un sémème.

[43] Il est d'usage en sémantique componentielle de noter les sémèmes entre apostrophes.

Les morphèmes en se combinant forment des mots et plus généralement des lexies. Dans la majeure partie des cas, les sémies de ces lexies (i.e. leur signifié) sont construites à l'aide des sémèmes des morphèmes. Cependant, cette construction ne correspond pas nécessairement à une simple composition de sémèmes mais plutôt à des règles cachées selon [Corbin 88], ou à des normes qui ne relèvent pas du système fonctionnel de la langue (on touche ici à la dimension sociale de la langue). Par exemple, on remarque effectivement que *pommade* ne signifie pas "préparation à base de pomme" selon le modèle d'*orangeade* ; de même on peut remarquer que *boulangère* ne signifie pas en général "femme qui fait le pain" mais plutôt "femme du boulanger"[44]. Au niveau d'une description linguistique des signifiés, il est donc pertinent de faire une distinction entre sémème et sémie. Cependant, dans le cadre d'une représentation uniforme des signifiés, nous considérons dans la suite, et par abus de langage, le sémème comme une représentation d'un signifié d'une lexie autant que d'un morphème. De plus, de notre point de vue, le sémème sera la représentation d'une signification plus que d'un signifié. En suivant le modèle du signe saussurien, le signe unit un signifiant à un signifié. Par conséquent, le signe ne peut avoir plusieurs signifiés alors que l'on peut attribuer à un même signifiant plusieurs sémèmes décrivant ses possibles significations lorsqu'il s'agit d'un signe homonyme ou polysémique. Attribuer un sémème à chacune des significations d'un signe constitue une modélisation homonymique de la polysémie[45].

La sémantique différentielle définit plusieurs classes lexicales qui permettent de caractériser la place des sémèmes en fonction de leurs sèmes dans le système de significations qu'ils forment. Ces classes sont les *taxèmes*, les *champs*, les *domaines* et les *dimensions*. Elles permettent d'identifier des normes sociales dans l'usage de la langue et d'affiner la nature des sèmes en fonction de leur rapport à une classe. Ainsi, le taxème est principalement lié aux sèmes spécifiques (c'est la classe sémantique lexicale minimale). Le champ est lié aux sèmes dits micro-génériques (c'est une hiérarchie structurée de taxèmes, par exemple le champ //meubles//[46] contient le taxème //sièges//). Le domaine est lié aux sèmes dits méso-génériques (c'est une classe liée à des connaissances et des pratiques sociales déterminées, où la polysémie et la métaphore sont en général assez peu répandues, c'est par exemple le domaine de l'informatique). Enfin, la dimension est liée aux sèmes dits macro-génériques (les dimensions divisent l'univers sémantique en grandes oppositions, par exemple animé *vs*. inanimé). En fait,

[44] La question de la féminisation des professions est plus que jamais d'actualité avec l'arrivée, de plus en plus fréquente, de femmes au sein des gouvernements (domaine qui était traditionnellement réservé aux hommes, il n'y a pas si longtemps). Ainsi, doit-on dire *Madame "la" ministre* pour faire la différence avec *Madame le ministre* (la femme du ministre) ? Cette question est actuellement débattue à l'académie française (cf. *Le Monde*, numéro du Mercredi 14 Janvier 1998, p. 16). Elle dénote une évolution diachronique actuelle de la langue française.

[45] Dans un tel modèle, la distinction entre l'homonymie et la polysémie depend de l'intersection des différents sémèmes. Quand cette intersection est non vide, il s'agit d'un signe polysémique et, quand elle est vide, il s'agit de deux signes homonymes. Ce que [Martin 83, p. 76] exprime par la phrase : "Contrairement à l'homonymie, la polysémie permet toujours de déceler un sème commun."

[46] Il est d'usage en sémantique componentielle de noter les taxèmes entre doubles //.

les trois dernières classes (dimensions, domaines et champs[47]) ont un but radicalement différent de la première (le taxème). Elles correspondent à des "outils" de description de la langue en synchro-diachronie alors que le taxème occupe une place centrale dans la description et dans la construction des systèmes de significations. Dans sa définition du taxème, la sémantique différentielle se réapproprie et précise celle exprimée par Coseriu comme "structure paradigmatique constituée par des unités lexicales se partageant une zone commune de signification et se trouvant en opposition immédiate les unes des autres" ([Coseriu 76]). Dans le taxème sont regroupés les sémèmes définis différentiellement par leurs sèmes spécifiques et c'est en cela que le taxème regroupe les significations les plus proches. Rastier lui donne une place particulière et signale que le taxème est la seule classe nécessaire car tout sémème comprend au moins un sème générique qui l'indexe dans son taxème de définition ([Rastier & al. 94, p. 61]).

1.2. Un modèle oppositionnel du sème

Notre modèle sémantique de l'axe paradigmatique s'inscrit dans le cadre général de l'analyse sémique telle que la conçoivent [Pottier 74], [Greimas 86], [Rastier 87] ou encore [Martin 83]. En effet, il s'agit bien ici de modéliser les signifiés en tant qu'ils forment un système de valeurs, ce que l'approche componentielle et différentielle permet de réaliser. Cependant, nos contraintes ne sont pas les mêmes que celles des théories purement linguistiques. En effet, une théorie linguistique peut, dans l'absolu, rester descriptive alors qu'une modélisation en intelligence artificielle est vouée à l'échec si elle n'en reste qu'à un niveau de description d'un objet ou d'un phénomène. Il faut donc que les modèles de l'IA soient *projectifs* au sens de [Vernant 98, p. 112], c'est-à-dire qu'ils puissent s'abstraire de l'interprétation du chercheur pour s'exécuter sur une machine. On touche ici à une différence fondamentale entre les deux disciplines que sont la linguistique et l'IA : la place de l'interprétation (humaine) dans leurs modèles. Dans une théorie linguistique, l'interprétation est un préalable à l'analyse et cette analyse ne consiste parfois qu'à simplement décrire cette interprétation. Ce qu'exprime Greimas au niveau de la sémantique par la phrase suivante.

> "La sémantique se reconnaît ainsi ouvertement comme une tentative de description du monde des qualités sensibles." ([Greimas 86, p. 9])

La difficulté consiste alors à essayer de faire la part des choses entre ce qui concerne le fait observé et ce qui concerne l'interprétation même du chercheur. C'est un problème qui touche toute personne qui se donne comme objet d'étude sa propre activité. Au regard de l'activité langagière, ceci provient de plus du fait que la langue est sa propre métalangue et c'est bien ce

[47] Pour de plus amples détails sur les dimensions, les domaines et les champs, on pourra se référer à [Rastier & al. 94, p. 61-63].

que montrent les dictionnaires unilingues qui, en définissant un mot par une suite d'autres mots, constituent des ensembles clos. C'est également l'idée que développe Greimas :

> "La difficulté principale d'une telle description provient, on l'a vu, du caractère privilégié des langues naturelles. Une description de la peinture peut être conçue, de façon très générale, comme la traduction du langage pictural en langue française. Mais la description de la langue française n'est, dans cette même perspective, que la traduction du français en français. L'objet de l'étude se confond ainsi avec les instruments de cette étude : l'accusé est en même temps son juge d'instruction." ([Greimas 86, p. 13])

À l'inverse d'un modèle linguistique inductiviste, un modèle d'IA ne doit pas présupposer une interprétation car simuler partiellement cette interprétation constitue la dimension calculatoire nécessaire du modèle. Le but de la dimension projective du modèle est alors de mettre en place, le mieux possible, un analogue machine de l'interprétation humaine ou au moins un calcul de sens constituant un premier pas vers cette interprétation. Si cette différence du statut de l'interprétation entre les deux disciplines est incontournable, il n'en reste pas moins que la richesse des descriptions linguistiques apporte des informations pertinentes pour la construction d'un système d'IA mais qu'un point de vue calculatoire détaché d'une interprétation humaine amène aussi un autre point de vue sur une théorie linguistique. C'est pour ces raisons que l'interdisciplinarité est fructueuse.

La mise en place de cette dimension projective dans un modèle d'IA pose deux problèmes. Le premier est que les représentations de la machine doivent être signifiantes autant pour l'humain censé interagir avec le système que pour le système lui même. C'est-à-dire, par exemple, que si on décrit un signifié comportant le trait /animal/, significatif pour l'humain, ce trait doit prendre forme sous une représentation qui le rende également significatif (pas forcément de la même façon) pour la machine. Le système de signes de la langue que l'on cherche à mettre en place ne doit donc pas simplement être donné à la machine à titre de description mais il doit être constitutif de l'activité même de la machine. Ceci nous amène au deuxième problème que pose la dimension projective : ne pouvant représenter en machine le système de signes de façon exhaustive, ce système doit être conçu de façon dynamique pour pouvoir assurer sa propre évolution. Cette nécessité de la dynamique du système est d'autant plus cruciale qu'il s'agit de mettre en place une interaction entre l'homme et la machine car, ce qui est le moteur d'une interaction, c'est une co-adaptation mutuelle des systèmes sémiotiques. Si une machine ne peut faire évoluer ses représentations, non seulement tout espoir d'apprentissage est perdu mais le fondement même d'une possible interaction est nié. Ces contraintes propres à notre approche nous ont conduit à reformuler les principaux concepts de la sémantique différentielle et, en premier lieu, celui de sème.

Selon le point de vue classique de l'analyse sémique, le sème est défini comme toute paraphrase métalinguistique, du simple /animé/ au plus compliqué /sert à découper le fromage/, ne contenant rien d'autre que l'interprétation qu'on peut en faire. En somme, les sèmes représentent des marqueurs qui se suffisent à eux-mêmes pour l'interprétation que peut en faire un humain. Cette vision du sème correspond à une approche linguistique des systèmes de signification, c'est-à-dire une approche descriptive du sens qui présuppose une interprétation humaine. Dans le cadre d'une modélisation computationnelle, deux approches peuvent être considérées. La première consiste à limiter le traitement de la machine à une description, c'est-à-dire à lui donner un rôle de suggestion et non de décision, laissant ainsi en fin de compte l'interprétation à la charge de l'utilisateur humain qui voit alors dans les traitements de la machine les raisons de son interprétation. Cette approche, dite anthropocentrée ([Thlivitis 98]), est celle de Tanguy ([Tanguy 97]) dans des domaines d'application plutôt littéraires justement parce que, là plus qu'ailleurs, assister l'interprétation humaine est utile et intéressant. La deuxième approche consiste à tenter de fournir à la machine les premières bases d'une "interprétation" qui soit la sienne et c'est cette approche que nous considérons ici en vue de l'amorce d'une compétence dialogique. Dans cette perspective, on ne peut plus considérer le sème comme une entité sans contenu, une simple chaîne de caractères. Le sème, s'il porte des informations qui orientent l'interprétation humaine, doit de plus structurer les représentations paradigmatiques de la machine. Il forme alors un objet de négociation possible dans l'interaction homme-machine.

Notre première proposition quant à la nature du sème consiste à ne plus l'entrevoir comme une entité positive, dans le sens où elle serait autosuffisante, mais plutôt comme une entité négative au sens de Saussure. Ainsi, ce qui nous paraît pertinent, ce n'est pas le trait sémantique en tant que tel, mais ce à quoi on peut l'opposer. Par exemple, le trait /animal/ en soi n'est pas très informatif mais ce qui est pertinent, c'est de savoir, dans le contexte courant, si par exemple il s'oppose à /végétal/, ou à /humain/, ou à /civilisé/ ... Ainsi, dans l'énoncé suivant, *faune* porte le trait /animal/ et dans ce contexte, ce trait est pertinent pour l'interprétant dans le cadre de l'opposition /animal/ *vs*. /végétal/ :

Il faut préserver la faune et la flore.

Dans d'autres contextes le trait /animal/ se justifie en rapport à d'autres oppositions, par exemple l'opposition /animal/ *vs*. /humain/, comme c'est le cas dans :

L'homme est un loup pour l'homme.

Ici, *loup* porte le trait /animal/ mais, cette fois-ci, le trait prend sens dans l'opposition /animal/ *vs*. /humain/ et non plus /animal/ *vs*. /végétal/. Donner une représentation de la signification du trait /animal/ consiste alors à donner les oppositions grâce auxquelles le trait /animal/ peut s'actualiser en contexte. Par exemple les quatre oppositions suivantes :

/animal/ *vs*. /végétal/
/animal/ *vs*. /humain/
/animal/ *vs*. /objet/
/animal/ *vs*. /divin/
…

Nous proposons donc de rapporter les sèmes à des jeux d'opposition de signes[48](que l'on représente par des oppositions de lexies). Il paraît bien sûr difficile, voire impossible, de donner une liste exhaustive de toutes les oppositions possibles pour un sème donné. Ceci traduit bien le rôle de relativisation du sème et le jeu d'oppositions considéré reste contextuel car lié à une interprétation. La formulation d'un sème consiste donc en une négociation des oppositions et ceci implique que la mise en place de connaissances sémantiques en machine ne puisse se réaliser que dans un processus interactif d'apprentissage. Cette vision oppositionnelle du sème consiste à considérer l'opposition comme élément de base de la signification et ceci coïncide bien avec la dimension défective de la langue décrite par Coursil ([Coursil 92]). L'opposition est ici envisagée comme structurante au niveau même des éléments de base de la signification. C'est, d'une certaine façon, donner à l'opposition une place première, de façon encore plus radicale que dans la sémantique différentielle où les oppositions n'apparaissent que dans les rapports entre les différentes significations.

L'évocation d'une opposition dans un contexte ne se limite pas à la négociation des termes opposés. Pour qu'une opposition soit significative, elle doit être rattachée à un domaine qui représente le cadre de son interprétation. Nous l'appellerons *domaine d'interprétation*[49] . C'est ce qui rend pertinente l'opposition car elle effectue alors une partition significative du domaine d'interprétation. Par exemple, les oppositions animal *vs*. végétal et animal *vs*. humain n'ont de signification que dans le domaine d'interprétation des êtres vivants. Le rapport d'un jeu d'oppositions à un domaine d'interprétation constitue un élément de base de la signification et c'est ainsi que l'on propose de reformuler le sème. Selon ce point de vue, le trait /animal/, tel qu'on l'a détaillé, prend part à la définition des trois sèmes suivants :

sème n° 1 :	domaine d'interprétation :	êtres vivants
	oppositions :	animal *vs*. végétal
		animal *vs*. humain
sème n° 2 :	domaine d'interprétation :	matériel
	opposition :	animal *vs*. objet

[48] On retrouve dans cette position la notion de spécème de Tanguy : "Un spécème est un couple de sémème appartenant à un même taxème, indiquant une opposition sémantique entre ceux-ci" [Tanguy 97, p. 203].
[49] Cette notion de domaine d'interprétation n'entretient *a priori* aucun lien avec la notion de domaine en tant que classe sémantique.

sème n° 3 :	domaine d'interprétation :	imaginaire
	opposition :	animal *vs*. divin

De même, les traits /solide/, /liquide/ et /gazeux/ ne donnent lieu qu'à un seul sème alors que dans le cadre de la sémantique componentielle ce sont trois traits distincts.

sème :	domaine d'interprétation :	état physique
	opposition :	solide *vs*. liquide *vs*. gazeux

Ce modèle du sème (domaine d'interprétation + oppositions) permet partiellement de rendre compte d'un aspect constitutif de la langue : la métaphore ([Lakoff & al. 85]). Par exemple, l'opposition bien *vs*. mal est significative pour le domaine d'interprétation des valeurs qualitatives et pourtant on l'utilise tous les jours de façon métaphorique dans le domaine d'interprétation du bien-être pour exprimer des états de santé. Ces utilisations métaphoriques s'expliquent alors par la projection d'un jeu d'opposition à un autre domaine d'interprétation que son domaine initial. Ceci montre bien que le domaine d'interprétation est tout aussi négociable en contexte que les termes des oppositions. C'est donc le sème tout entier qui est soumis à un consensus lors des processus d'intercompréhension.

La sémantique différentielle définit le sème comme l'extrémité d'une relation fonctionnelle binaire entre significations ([Bachimont 96, p. 133]) alors que, de notre point de vue, le sème est cette relation entre significations. Par exemple, en sémantique différentielle, on différencie les significations de *bistouri* et de *scalpel* en opposant les deux sèmes /pour les vivants/ et /pour les morts/ alors qu'ici nous considérons que *bistouri* et *scalpel* sont différenciés par un seul et unique sème qui résulte de l'opposition vivant *vs*. mort dans le domaine d'interprétation de l'état physiologique. La signification de *bistouri* actualise ce sème dans la valeur vivant et *scalpel* dans la valeur mort. De même, *venimeux* et *vénéneux* sont différenciés par un sème partitionnant le domaine d'interprétation des êtres vivants par l'opposition animal *vs*. végétal et non par deux sèmes (comme dans [Rastier 92]). À la différence de la définition classique du sème (simple paraphrase métalinguistique) qui autorise une description componentielle sémasiologique, la nature oppositionnelle de notre modèle du sème ne nous incite pas à décrire les significations de façon isolée car les sèmes sont justifiés par leur potentiel de différenciation et les sémèmes sont inter-définis de façon relationnelle par leur contenu sémique. Ceci place d'emblée notre modèle dans le cadre d'une méthode de description componentielle onomasiologique.

Dans la suite, nous allons monter comment opérationnaliser notre modèle de description componentielle. Pour ce faire, on utilisera une méthode interactive de catégorisation. Cette méthode est mise en place dans un modèle appelé *Anadia*.

2. Le modèle Anadia

L'objet de cette partie est de présenter une méthode qui rend appréhendable pour une machine la notion de valeur, notion qui permet de fonder un *système de valeurs*. Cette méthode est une méthode de catégorisation différentielle[50], appelée *Anadia*, qui est fondée sur l'interdépendance entre les entités qu'elle permet de représenter. En fonction du but qu'on se donne, on peut, soit l'utiliser comme un modèle de construction d'un système de signes (c'est le cas dans cette thèse), soit l'utiliser comme un modèle de représentation de connaissances. Le point commun entre ces deux types de points de vue est que, quelle que soit la nature des entités considérées, signes ou connaissances, ces entités sont considérées plus pour la place qu'elles occupent dans le système qu'elles forment que pour leurs propriétés intrinsèques. La représentation de connaissances a déjà été maintes fois débattu en IA et nous allons justement montrer en quoi Anadia se distingue des approches connues. Tout d'abord, cette question de la représentation du monde a souvent été abordée du point de vue de l'agent en tant qu'individu ayant la capacité de se décrire mentalement le monde pour ensuite raisonner sur ses représentations qui servent de substitut à la réalité ([Davis & al. 93]). Le point sur lequel Anadia insiste particulièrement, c'est que les représentations du monde ne sont pas le fait d'un agent en tant qu'individu mais qu'elles sont dues aux interactions qu'entretiennent les agents et c'est cette importance de l'interaction qui donne à la connaissance un statut de fait social. De plus, dans les modèles classiques de représentation de connaissances, les choses représentées sont présupposées connues et différenciées. Leur représentation est de fait supposée complète, c'est-à-dire qu'elle est stable et observable, qu'elle ne dépend pas de l'agent qui la construit et encore moins de son activité langagière. C'est précisément cela qu'Anadia remet en cause en mettant au premier plan l'interaction. Nous ne proposerons donc pas de représenter les choses mais plutôt de les présenter, de les montrer à la machine ou *via* la machine, pour que celle-ci soit d'emblée dans une situation d'interaction. Cette importance de la monstration a déjà été évoquée au chapitre 1 comme une base de l'activité de co-référenciation, c'est-à-dire comme une base de l'interaction langagière. D'une façon générale, la démarche que nous proposons avec Anadia consiste à donner à la machine une amorce de l'activité langagière. Les choses en jeu dans l'interaction ne seront alors jamais considérées comme complètement décrites et encore moins de façon exhaustive. Nous ne viserons donc pas, comme c'est souvent le cas, une représentation scientifique ou ontologique du monde mais plutôt un analogue de la mémoire sémantique constituant une compilation des choses mises en commun lors d'une ou plusieurs interactions.

[50] La catégorisation différentielle s'oppose aux méthodes catégorisations par construction de prototypes (par exemple [Dubois 93]).

2.1. Un modèle oppositionnel de la valeur Saussurienne

La méthode de catégorisation sous-jacente à Anadia prend sa source dans la notion de valeur Saussurienne et dans la sémiotique de Hjelmslev, pour qui le premier principe à considérer dans l'analyse d'une structure est que "une totalité ne se compose pas d'objets mais de dépendances" ([Hjelmslev 43, p. 37]). Cette méthode à été formulée plus précisément et de façon opératoire par Coursil pour rendre compte de ce qu'il appelle la dimension défective de la langue ([Coursil 92]), c'est-à-dire le fait que la langue se constitue essentiellement par des *coupures*, des différences. Avant de participer à une première implémentation du modèle ([GIL 95]), il l'a utilisée manuellement à de nombreuses reprises pour l'analyse du discours ([Coursil 93a]). Notamment, il a pu montrer clairement avec Anadia la structure systémique et signifiante de la phonologie du français contemporain telle que l'a étudiée Jakobson, ainsi que la structure des morphèmes dans la constitution syntagmatique des signes.

La méthode mise en place par Anadia ne cherche pas à décrire l'essence des entités en jeu dans une interaction. Il s'agit plutôt de proposer une sorte de grille qui permette de montrer les différences pertinentes entre ces entités, que ce soit pour interpréter ou pour produire des énoncés langagiers. Le point de vue que nous considérons ici n'est pas celui de la production des énoncés mais celui du sujet qui écoute et interprète[51] (que ce soit les paroles des autres ou mêmes les siennes car un sujet qui parle s'entend parler). Il s'agit de donner les bases d'un calcul du sens dans l'activité d'interprétation. La grille offre un certain de nombre de places définies par les enjeux de l'interaction et il est du ressort de l'agent de faire correspondre ces places avec ses propres connaissances. De ce point de vue, la grille modélise un état mémoire de la machine qui structure dans l'interaction langagière un état de ses connaissances mais, en aucun cas, elle ne s'y substitue car elle n'est révélatrice que de l'effectivité de l'interaction où les connaissances de chacun se croisent. Le problème de la représentation des connaissances de chaque agent relève d'une autre approche plus centrée sur les facultés cognitives individuelles. Il s'agit ici de rendre compte, par la grille, d'un système où les places peuvent montrer des objets, sans s'y substituer. Chacune de ces places est, de ce point de vue, révélatrice d'un signe au sens de [Peirce 78, p. 123]. La grille permet donc de représenter un système de signes au centre de l'interaction qui permet de révéler des relations de différences entre chacun de ces signes. Elle met en place un système de valeurs.

On ne peut parler de représentation de systèmes de signes sans évoquer le rapport d'un modèle comme Anadia au carré sémiotique ([Ermine 96]), aussi appelé "carré d'Aristote", massivement utilisé par Greimas dans ses *essais sémiotiques* ([Greimas 83]). La démarche est relativement similaire, c'est une démarche sceptique au sens de Sallantin :

[51] C'est également la position de Jacques Coursil qui la définit en terme de *renversement du sujet* ([Coursil 92]).

> Le carré d'Aristote est précisément une démarche de jugement sceptique. Le sens n'est pas déterminé d'une manière axiomatique. Il est saisi par un réseau de corrélations. Ces corrélations sont des oppositions et des combinaisons. ([Sallantin 97, p. 34])

Le carré sémiotique définit à l'aide de deux relations de différences (les relations abstraites de contrariété et de contradiction) quatre places où l'on peut représenter des systèmes d'opposition de signes. Si l'on considère comme étant contraires *p* et *q*, alors *p* et *q* forment un axe sémantique. Le carré sémiotique représente que l'effet de sens produit sur l'axe sémantique issu de l'opposition des subcontraires *non p* et *non q* est similaire à l'effet de sens produit par l'opposition entre *p* et *q*. Ce qu'on représente par la figure suivante où *p* et *non p*, de même que *q* et *non q* sont contradictoires :

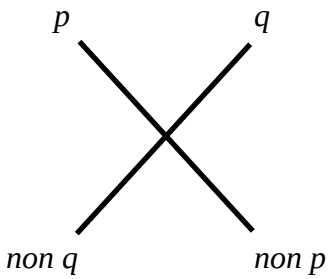


Figure 3.1 : Structure du carré sémiotique

Par exemple, le carré suivant est construit sur l'axe sémantique du *désirable* et du *nuisible* et il concerne la combinatoire des verbes être et vouloir ([Greimas 83, p. 99]).

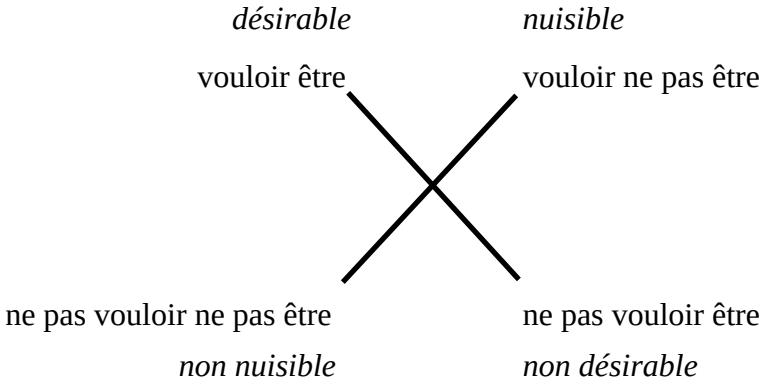


Figure 3.2 : Un exemple de carré sémiotique

Dans le carré sémiotique (comme dans l'hexagone de [Blanché 66] issu des mêmes principes), les relations de différence sont données avant même de considérer les signes qui entretiennent ces relations. Ceci implique effectivement que, n'ayant que deux relations, on ne puisse constituer que des systèmes à quatre places. Les grilles que met en place Anadia ne sont

pas construites sur des relations de différences prédéfinies de façon logique et nommées, comme c'est le cas pour le carré sémiotique, mais sur des relations définies par un certain nombre de traits de distinction entre les places. Le nombre de ces différences n'est pas *a priori* fixé et donc le nombre de places dans la grille non plus. Si l'on ramène les différences que l'on utilise dans Anadia (c'est la notion d'attribut que l'on verra plus loin) à des axes sémantiques, alors les axes sémantiques ne sont plus limités au nombre de deux, et plutôt que de parler d'axes, il conviendrait de parler d'espaces sémantiques étant donné que les différences dans Anadia ne sont pas nécessairement binaires. On peut donc, d'une certaine façon, considérer Anadia comme une généralisation du carré sémiotique car la démarche sceptique est similaire et le but commun est de rendre compte de systèmes d'oppositions de valeurs. Cependant, cette généralisation est algébrique (elle est fondée sur la notion de combinatoire, calculée par produit cartésien) et elle s'abstrait des contraintes de la logique des prédicats.

2.2. La structure d'Anadia

Anadia met en place un processus de catégorisation[52] qui s'appuie sur l'examen des différences entre les catégories et non sur leurs propriétés intrinsèques. Le but n'est pas de mettre au jour des prototypes pour chacune des catégories mais de considérer les attributs discrets qui les définissent différentiellement. Le processus consiste à découper une catégorie en sous-catégories en croisant les valeurs des attributs les plus pertinents pour cette catégorie. Cette pertinence ne dépend pas directement de la catégorie en cause mais dépend d'un projet de catégorisation. Effectivement, on peut toujours sous-catégoriser une catégorie de plusieurs manières en fonction du but qu'on se fixe et ce sont alors les attributs considérés comme pertinents qui changent. Par exemple, si on cherche à sous-catégoriser les êtres vivants pour montrer ce qui caractérise la sous-catégorie des hommes, on ne va pas forcément se donner les mêmes attributs que si l'on cherche à mettre en évidence la sous-catégorie des poissons. On peut penser que, dans le premier cas, l'attribut relevant de la propriété de marcher sur deux pattes sera pertinent alors qu'il ne le sera pas dans le deuxième cas. La modélisation qu'Anadia permet dépend donc essentiellement du projet qu'on se donne et c'est une raison de plus pour ne pas y chercher une description scientifique et ontologique du monde (sauf si cela est explicitement le projet donné).

Le processus de sous-catégorisation est récursif et il est tout à fait possible (c'est même fréquent) de le réappliquer aux sous-catégories qu'il a permis de mettre en évidence. Le résultat obtenu prend alors la forme d'un arbre de grilles de catégorisation où chaque niveau est une sous-catégorisation des places de la grille du niveau supérieur. Chacune de ces grilles dénote

[52] Il ne s'agit pas ici de catégorisation ou de sous-catégorisation au sens de la syntaxe, c'est-à-dire au sens restreint des catégories grammaticales.

une catégorie et montre les sous-catégories obtenues par le processus de croisements des valeurs des attributs localement pertinents. Il n'y a pas de relations entre les sous-catégorisations des places de la même grille, c'est-à-dire entre les grilles "frères" dans l'arbre, car en général les attributs utilisés sont différents et c'est pour cela que la notion de pertinence est locale à une seule sous-catégorisation.

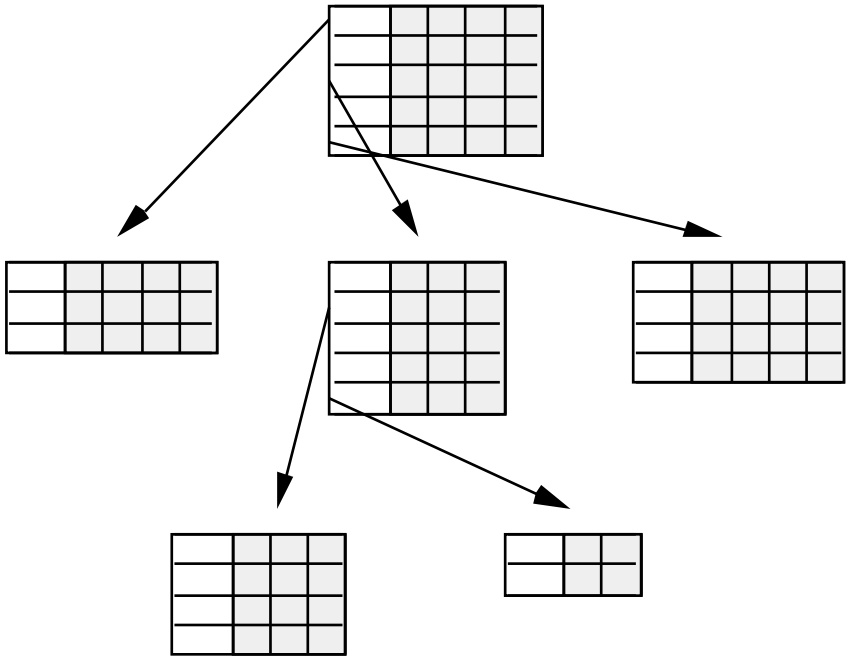


Figure 3.3 : Arbre de grilles de catégorisation

Dans ce schéma, les cases grisées d'une table indiquent les critères de définition des catégories qui figurent dans les cases blanches, que l'on peut sous-catégoriser à nouveau en créant une nouvelle table.

Pour être optimale par rapport au projet, la représentation n'a pas à être complète mais elle doit être suffisante pour différencier toutes les entités qui sont visées comme différentes (et ce avec le moins de charge cognitive possible). Ceci ne peut s'obtenir que dans un échange interactif entre un système qui manipule des concepts et un agent qui les construit et les modifie en accord avec son projet de catégorisation. Nous présenterons, dans un premier temps, les concepts propres à Anadia et, dans un second temps, le processus interactif qu'il met en place.

2.2.1. Les concepts

Le premier concept mis en jeu dans le modèle Anadia est, comme on l'a évoqué, la notion d'*attribut*. Un attribut est une propriété que l'on peut ou non attribuer à une chose. Par exemple, les choses matérielles ont une couleur et les événements n'en ont pas ; de même, les processus

ont une durée mais pas les figures géométriques. Dans cette signification exprimant une qualité que l'on peut attribuer à une chose, on retrouve le sens philosophique du mot catégorie mais, pour nous, une catégorie se limite à un groupe de choses semblables à certains égards.

Un attribut se caractérise par ses valeurs possibles. Dans Anadia, on s'intéresse aux différences entre les sous-catégories que les valeurs d'un attribut permettent de mettre en évidence. Conformément aux remarques d'Aristote sur la catégorisation conceptuelle, les différences à retenir sont des différences de traits et non des différences de grandeur de champs ([Aristote 69, p. 158]). Ce point de vue permet de distinguer deux types d'attributs, les attributs formés de valeurs continues et les attributs formés de valeurs discrètes. En suivant Aristote, seuls les attributs à valeurs discrètes et ayant un nombre fini de valeurs permettent de créer des concepts et ce sont les seuls que nous prendrons en compte dans Anadia.

La notation des attributs comporte un nom et l'ensemble fini et non ordonné des valeurs de l'attribut. On représente par exemple comme le montre la figure suivante, l'attribut noté A, concernant les automobiles et lié au type de leur moteur, à savoir diesel, essence ou électrique.

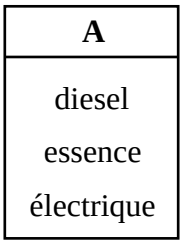


Figure 3.4 : Un exemple d'attribut

La plupart du temps, pour réaliser une sous-catégorisation d'une catégorie donnée, considérer un seul attribut ne suffit pas car, dans ce cas, l'attribut a simplement pour valeurs la liste des sous-catégories et c'est donc qu'on les connaît déjà. L'intérêt de l'interaction qu'Anadia propose est que les sous-catégories cherchées sont à découvrir dans le processus même de sous-catégorisation et que cette découverte est initiée par le croisement de plusieurs attributs. Ceci amène à définir un concept important du modèle Anadia, la notion de *registre*. Former un registre consiste à considérer à un même niveau plusieurs attributs pour que le croisement de leurs valeurs fasse apparaître les sous-catégories cherchées. L'important de cette notion n'est donc pas tant dans sa structure, qui ne représente qu'une union de plusieurs attributs, que dans le choix des attributs qui constituent le registre. Ce choix soulève directement la question de la pertinence d'un attribut à un certain niveau de la sous-catégorisation. Ainsi, les attributs qui ont la même valeur pour tous les éléments d'une catégorie ne sont pas pertinents dans la constitution d'un registre pour cette catégorie car ils ont servi à la faire apparaître (au niveau immédiatement

supérieur dans la catégorisation) et sont maintenant inutiles pour y réaliser des différences. Les attributs qui vont être utiles, et qui vont alors former le registre, sont ceux qui, tout en étant pertinents pour tous les objets tombant sous la catégorie, vont prendre des valeurs différentes d'un objet à un autre. Ces attributs forment le *propre* de la catégorie.

Si par exemple, on s'intéresse à une sous-catégorisation des voitures, on peut choisir de considérer, en même temps que l'attribut A vu précédemment, un attribut B relatif au type de carrosserie et qui comporte les valeurs *berline, break, coupé* et *décapotable*. On forme alors un registre regroupant à un même niveau conceptuel les attributs A et B.

A	**B**
diesel	berline
essence	break
électrique	coupé
	décapotable

Figure 3.5 : Un exemple de registre

Une fois un registre défini, Anadia calcule la grille de catégorisation issue de la combinatoire des valeurs des attributs du registre. Cette combinatoire est calculée par un produit cartésien et cette grille de croisement de valeurs définit la notion de *table* du modèle Anadia. Chaque ligne de cette table représente un jeu de valeurs possibles et offre une place pour une catégorie potentielle. Toutes les places de la table sont une représentation d'un concept, dans le sens d'objet mental résultant d'une opération algébrique (il ne nécessite pas obligatoirement l'existence d'objets). C'est là que se fait la différence entre la notion de concept et celle de catégorie. Pour qu'une catégorie soit attestée, on doit pouvoir donner des exemples d'objets tombant sous son concept. La figure suivante montre la table associée au registre {A, B}. Elle offre 12 places, détermine 12 concepts, 12 catégories potentielles.

	A	**B**
	diesel	berline
	diesel	break
	diesel	coupé
	diesel	décapotable
	essence	berline
	essence	break
	essence	coupé
	essence	décapotable
	électrique	berline

	électrique	break
	électrique	coupé
	électrique	décapotable

Une fois la table calculée, il est du ressort de l'agent menant son projet de catégorisation de faire la distinction entre les places qui restent au niveau de concept et celles qui acquièrent un statut de catégorie effective. Ces catégories constituent ce que l'on appelle la *sélection*, et l'agent va nommer ces places ou en donner des exemples. En rapport à la table précédente, on peut donner beaucoup d'exemples de voitures qui ont un moteur diesel et une carrosserie de berline. C'est donc une catégorie effective et la place qui y correspond fait alors partie de la sélection. En revanche, le jeu de valeurs *électrique* et *break*, ne semble pas actuellement réalisé en tant que catégorie car il apparaît difficile d'en donner des exemples (c'est uniquement un concept).

Lorsqu'une catégorie est effective, l'agent peut la nommer. Il porte alors son nom dans la place correspondante de la table. Nommer une catégorie, c'est beaucoup plus que simplement avoir pu distinguer son concept. Par exemple, les peintres distinguent et nomment fréquemment plusieurs centaines de couleurs alors que les ordinateurs peuvent en distinguer plusieurs millions mais ils ne les nomment pas au même titre que les peintres. Nommer une catégorie, ou même une chose, relève de l'activité humaine qui consiste à se donner une entité qui est un représentant (un "tenant lieu") pour la catégorie ou la chose, c'est-à-dire former un signe. L'intérêt de la démarche que propose Anadia est donc bien d'initier une interaction où l'enjeu est sémiotique : former des signes et les organiser en même temps dans un système. De plus, articuler dans le même modèle la création de signes et la création de catégories permet de reconnaître les catégories comme des objets de référenciation, et donc de co-référenciation possible ([Delépine & al. 95]). On se donne ainsi les moyens de pouvoir initier ou relancer une interaction homme-machine dont les enjeux sont comparables aux interactions en langue naturelle.

La table donne des descriptions en termes de valeurs d'attributs pour chacune des places, donc également pour les catégories de la sélection. De plus, elle structure ces places dans un système où chacune entretient des relations avec les autres. Ces relations entre places sont fondées sur les différences entre leur jeu de valeurs. Ce sont leurs différences qui organisent un système où chacune des places est avant tout "ce que les autres ne sont pas" et c'est cela qui définit la nature propre du signe pour Saussure ([Saussure 1915, p. 162]). Les relations de différence qu'Anadia permet de calculer à partir d'une table sont définies par un nombre de différences de valeurs d'attributs entre les places sélectionnées. On appelle *topique* le graphe de ces relations. Dans la majeure partie des cas, la topique la plus intéressante est la topique des différences "à un trait près". La relation qui la définit est celle qu'entretiennent les catégories dont les jeux de valeurs ne diffèrent que par une seule valeur d'un attribut. Cette relation de

différence est la plus fine qui puisse exister entre deux catégories distinctes. La topique est de ce fait le graphe le plus détaillé qu'on puisse donner pour décrire le système organisé par les relations de différences.

À chaque table correspond une topique à un trait près qui compte autant de sommets que d'éléments dans la sélection. On peut alors considérer que l'arbre de sous-catégorisation révèle également un arbre de topique.

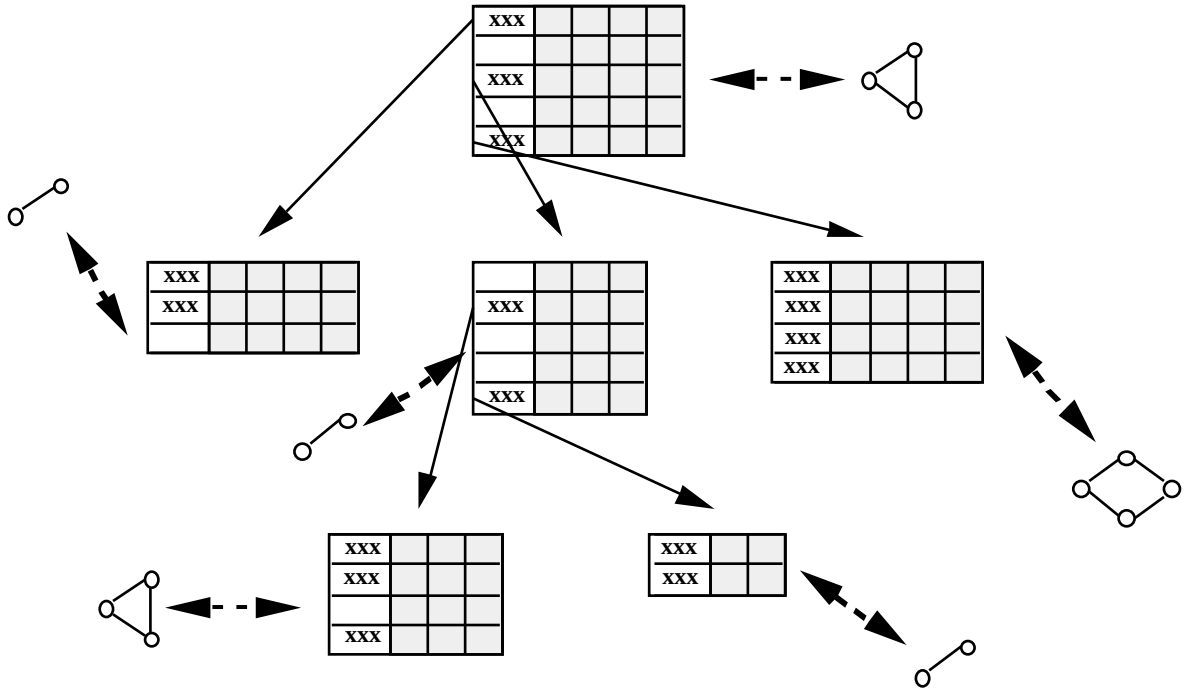


Figure 3.6 : Arbre de topiques

Dans ce schéma, les doubles flèches en pointillés représentent le calcul d'une topique à partir d'une table, les simples flèches marquent le processus de sous-catégorisation et les *xxx* sont des noms de catégories dans les tables.

Les topiques permettent de faire apparaître la structure différentielle engendrée par les attributs considérés et par le choix des catégories. La structure sous-jacente aux attributs considérés est la topique maximale obtenue lorsque toutes les places de la table sont sélectionnées. La topique maximale présente certaines propriétés mathématiques :

• Elle est connexe et tous les sommets ont même degré, c'est-à-dire que tous les sommets ont le même nombre de voisins dans le graphe.

• C'est un produit cartésien de graphes complets ([Andreae & al. 95]). Si on considère une table formée à partir d'un seul attribut à *n* valeurs et que l'on y sélectionne toutes les places, on obtient *n* sous-catégories qui sont toutes en relation de différence à un trait près les unes des autres étant donné qu'il n'y a qu'un attribut. Le graphe de la topique est donc

un graphe complet (tout sommet a pour voisins tous les autres). Lorsqu'on construit une table avec plusieurs attributs, cela revient à construire un graphe qui résulte du produit des graphes complets obtenus par les topiques maximales des attributs considérés un à un.

topique maximale d'un attribut à 3 valeurs

topique maximale d'un attribut à 2 valeurs

topique maximale d'une table construite avec 2 attributs à 2 et 3 valeurs

Figure 3.7 : Un exemple de produit cartésien de graphes

La structure de la topique provient donc d'abord du nombre et de la nature des attributs à l'origine de la table. L'effet de la sélection de l'agent dans la table sur la constitution de la topique est de former un graphe partiel en validant ou non certains sommets ainsi que les arcs qui en sont issus. Aux places de la table correspondent des sommets du graphe maximal. La table donne les valeurs des places en terme d'attributs et la topique montre leur position dans le système. Ce sont deux façons complémentaires de considérer le système de valeurs produit par les attributs.

2.2.2. Le processus interactif de catégorisation

L'intérêt principal de la méthode Anadia ne réside pas tant dans les structures qu'elle met au jour que dans le processus qui aboutit à cette mise au jour. Ce processus consiste en une interaction où se succèdent des opérations algorithmiques du logiciel (calcul de tables par combinatoire, calcul de topiques par examens de différences) et des choix de la part de l'agent qui l'utilise, voire des remises en causes du fait de la présentation des structures produites. C'est une interaction où chacun des deux pôles, le système et l'agent, a ses rôles et capacités propres et où l'enjeu est la co-construction d'un système sémiotique. Elle nous paraît, de ce point de vue, être une première étape intéressante dans l'apprentissage par amorce d'une compétence langagière pour les agents logiciels.

La construction d'une bonne grille de catégorisation pour un domaine est un processus expérimental car, en dehors de certains cas d'école ou de domaines très simples, il convient de choisir plusieurs attributs, d'en examiner la combinatoire et de les remettre en cause jusqu'à ce que le résultat ait de bonnes propriétés algébriques et conceptuelles. L'évaluation des propriétés conceptuelles est directement lié au projet de l'agent, à savoir ce qu'il veut faire d'une

catégorisation de son domaine. L'évaluation des propriétés algébriques est, quant à elle, directement liée aux choix qui ont permis la catégorisation du domaine. Anadia, dans son processus interactif, permet d'évaluer ces propriétés algébriques. Les discussions que la présentation des choses peut susciter, en rendant visibles les catégorisations langagières, peuvent aider l'agent à évaluer les propriétés conceptuelles. Anadia permet donc d'analyser et de structurer un domaine avant de se lancer dans la conception d'un logiciel traitant ce domaine, comme par exemple un système expert. Il s'agit, en fait, d'évaluer la pertinence de représentations du domaine et de les affiner tant que cette pertinence n'est pas optimale ([Beust & al. 96]).

Pour construire une table, la première des choses à faire est de constituer un registre. C'est-à-dire, d'une part, de construire des attributs en fixant leurs valeurs et, d'autre part, de choisir de les considérer ensemble au même niveau de la catégorisation. Le cas où une catégorie est partitionnée par un seul attribut est, comme on l'a vu précédemment, un cas particulier. Si on cherchait à toujours sous-catégoriser chaque catégorie à l'aide d'un unique attribut, l'arbre de sous-catégorisation obtenu ne serait rien de plus qu'un arbre de discrimination comme l'*arbre de Porphyre* par exemple ([Eco 88]). Ceci permet de préciser qu'il y a en fait deux principales différences entre la méthode utilisée dans Anadia et la méthode utilisée pour construire les arbres de discrimination :

1) Pour faire un arbre de discrimination, il faut choisir les attributs discriminants les uns après les autres étant donné qu'un seul est pris en compte à chaque niveau. Ce choix est souvent arbitraire. Si deux attributs apparaissent au même niveau, il faudra en choisir un en premier lieu et redécomposer les catégories obtenues avec le second. On peut toutefois construire un deuxième arbre en considérant en premier le deuxième attribut et le premier en second. Les deux arbres obtenus ne peuvent être départagés (cf. Figure 3.8). À l'inverse des arbres de discrimination, Anadia prend en compte en même temps tous les attributs partitionnant une catégorie et il n'y a donc pas d'alternative. De plus, les partitions de catégories de même niveau utilisent des attributs différents.

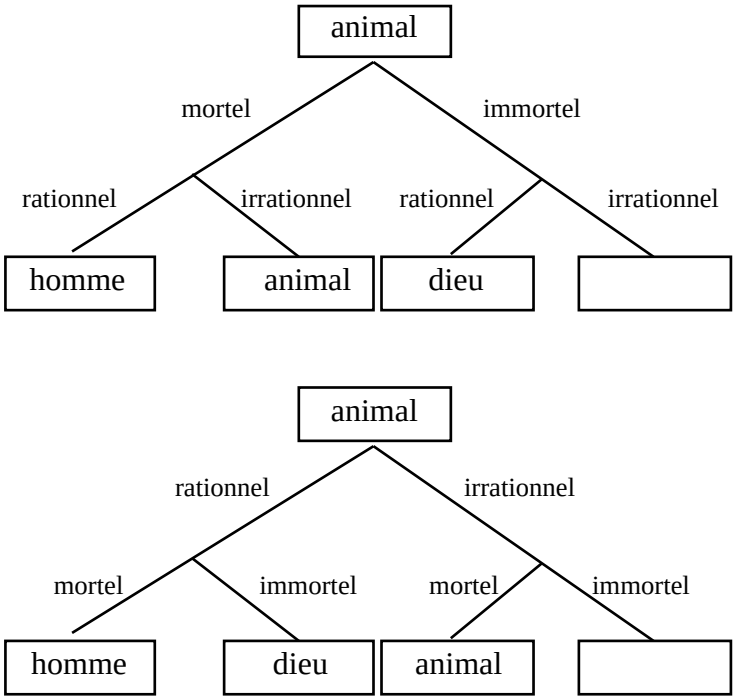


Figure 3.8 : Deux arbres de discrimination alternatifs

2) La deuxième différence avec les arbres de discrimination est une restriction concernant le type des attributs considérés. Comme on l'a déjà signalé, seuls les attributs ayant un nombre fini de valeurs discrètes sont utilisés dans Anadia. Les attributs à valeurs continues, comme l'âge ou la taille par exemple, servent seulement à différencier les objets qui tombent sous une même catégorie mais ces attributs ne peuvent être retenus pour faire des différences conceptuelles entre les catégories. C'est-à-dire qu'ils ne peuvent être pris en compte pour définir une catégorie. Nous faisons ainsi l'hypothèse que seules les différenciations entre les valeurs d'attributs discrets participent à la construction de la mémoire sémantique et que les attributs continus, bien que connus en mémoire sémantique, concernent dans leur valuation la mémoire épisodique ([Nicolle 96]) car ils ne structurent pas les paradigmes (c'est pourquoi le continu est explicitement marqué dans l'axe syntagmatique par des expressions comme "un peu plus de ...", "un peu moins de …"). Toutefois, des attributs continus par nature peuvent être discrétisés par des opérations langagières et peuvent alors servir d'attributs discrets dans un projet de catégorisation. Par exemple, la couleur[53], bien que correspondant à une fréquence (et donc à un domaine continu), est discrétisée par la langue qui permet de nommer différentes couleurs. On peut donc, avec les différentes valeurs de couleurs, constituer un attribut discret pour une catégorisation. D'une manière générale, pour pouvoir construire un

[53] La couleur peut être sous-catégorisée avec Anadia en considérant comme attributs la présence ou non de chacune des trois couleurs primaires. On obtient ainsi les couleurs secondaires comme sous-catégorisation. Cette catégorisation a été présentée dans [Beust & al. 96].

attribut dans Anadia, il faut que les valeurs de cet attribut puissent être énumérables (même si elles ne sont pas, dans l'attribut, énumérées de façon exhaustive). Ceci explique aussi que la mesure n'intervienne jamais dans les catégorisations d'Anadia, seul le jugement compte.

Une fois un registre constitué, la table est construite par une opération algorithmique (produit cartésien) mais les places qui apparaissent sont à remplir et l'étape qui suit dans l'interaction est à la charge de l'agent qui utilise Anadia. Cette étape consiste à reconnaître et à nommer parmi les places de la table les catégories effectives. Cette étape est particulièrement intéressante au regard du choix des attributs considérés et trois questions peuvent être posées.

Question n°1 : Pour chacune des places, est-il possible de trouver des objets qui tombent sous ce concept ?

Si la réponse est oui, alors le concept en question dénote une sous-catégorie effective. Elle peut être nommée, elle apparaîtra dans la topique et pourra être partitionnée dans la suite.

Question n°2 : Y-a-t-il des objets que l'on ne sait pas attribuer à l'une ou l'autre des places ?

Si la réponse est oui, alors les descriptions que réalisent les attributs ne sont pas assez fines et c'est donc qu'il manque au moins une valeur à un attribut qui permette de classer l'objet qui pose problème.

Question n°3 : Toutes les distinctions voulues sont-elles obtenues ?

Si la réponse est non, alors il manque un attribut ou une valeur d'attribut qui permette de faire la différence entre deux catégories qui apparaissent à tort comme confondues.

Lorsque toutes les distinctions voulues ont pu être réalisées et que la sélection a été constituée, c'est-à-dire lorsque les catégories reconnues ont été nommées, on peut demander que soit affichée la topique qui résulte de l'état actuel de la table. En regardant les sélections et les topiques, l'agent peut choisir de supprimer ou de transformer des attributs, ou encore de transformer un attribut pour tenter de refaire différemment la même catégorisation jusqu'à ce qu'il arrive à un résultat satisfaisant.

Les tables et les topiques offrent des indices pour juger de la qualité algébrique des représentations produites ; la connexité de la topique est un de ces indices. En effet, si deux sous-catégories sont voisines dans la topique, c'est qu'il existe un seul attribut qui les différencie par deux de ses valeurs. C'est donc que ces deux valeurs sont bien choisies et que l'attribut a bien sa place à ce niveau de la catégorisation. Plus la topique présente d'arcs, c'est-à-dire plus elle est connexe, meilleur est le registre. Ceci est directement lié à une propriété relative

à la relation entre les catégories et les attributs : *la pertinence*. Un attribut est pertinent pour une catégorie si tous les objets de la catégorie ont la propriété dénotée par l'attribut mais qu'ils n'ont pas tous la même valeur pour cette propriété. Dans ce cas, la connexité de la topique est effectivement accrue. Si tous les objets de la catégorie ont la même valeur pour un attribut, c'est que cette valeur est une caractéristique de la catégorie toute entière et c'est donc qu'elle n'est pas pertinente pour en réaliser une sous-catégorisation, mais qu'elle est pertinente à un niveau supérieur de la décomposition (niveau où l'on fait apparaître la catégorie). Si aucune chose n'a cette propriété, elle n'est pas non plus pertinente pour la catégorie mais peut-être l'est elle pour d'autres catégories du même niveau dans l'arbre de catégorisation. Enfin, si certains objets seulement de la catégorie présentent cette propriété, c'est que l'attribut est pertinent pour une sous-catégorie qu'il convient de déterminer.

Un autre indice utile pour évaluer les représentations concerne l'importance relative dans la table des places vides relativement aux catégories de la sélection. On appelle ici places vides les lignes de la table qui n'ont pu être nommée, c'est-à-dire les lignes qui font référence à un concept qui ne se réalise pas en tant que catégorie. Cet indice met en cause deux types de propriétés :

1) des propriétés de relation entre les attributs : *l'indépendance et la subordination*

2) des propriétés d'un attribut liées à ses valeurs : *la complétude et la minimalité*

1) L'indépendance et la subordination : on sait que deux attributs sont indépendants si toutes les places de la table formée par leur combinatoire sont sélectionnées, c'est-à-dire, si il n'y a pas de places vides. Dans le cas contraire, le doute quant à la dépendance des deux attributs est possible. Dans la plupart des cas, cette dépendance consiste en une subordination des attributs. Un attribut A est subordonné à un attribut B si A est pertinent seulement quand B a certaines valeurs. Par exemple, étant donné que V est subordonné à P, les deux attributs suivants P et V sont dépendants:

P :	propriété :	Pour un bateau, quel est le mode principal de propulsion ?
	valeurs :	moteur, voile
V :	propriété :	Pour un voilier, quel est son type de gréement ?
	valeurs :	sloop, ketch, cotre

Mettre dans un même registre deux attributs subordonnés amène un nombre important de places vides. Par exemple, dans la table formée avec P et V, la moitié des places sont vides du fait de la subordination des attributs.

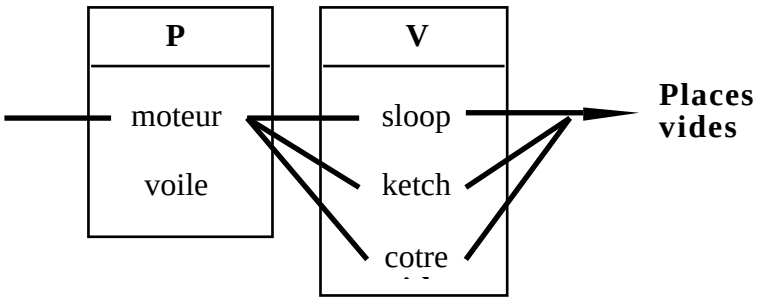


Figure 3.9 : Attributs subordonnés

2) La complétude et la minimalité : l'ensemble des valeurs d'un attribut est complet si tous les objets évoqués peuvent être classés selon les valeurs définies. Cette propriété est triviale si les valeurs de l'attribut sont le couple {oui, non}, mais on risque alors de produire des places vides dans la table car les attributs ne seront plus forcément indépendants. De plus, avec de telles valeurs, il faut veiller à ce que la valeur "non" de l'attribut ne soit pas synonyme de non pertinence de l'attribut dans sa catégorie (de qui introduirait aussi des places vides). L'attribut V, défini précédemment, est un exemple d'attribut incomplet car il manque notamment la valeur goélette, on peut donc vouloir le redéfinir ainsi :

V : propriété : Pour un voilier, quel est sont type de gréement ?
valeurs : sloop, ketch, cotre, goélette

On peut montrer qu'un attribut est incomplet en exhibant des objets qu'il devrait pouvoir décrire (c'est ce qu'on vient de faire avec l'attribut V) ; cependant, on ne peut jamais montrer qu'il est complet car les catégories ne sont pas définies par des propriétés nécessaires mais par des propriétés suffisantes. C'est pourquoi les résultats d'une telle catégorisation ne peuvent être interprétés en termes ensemblistes. L'autre propriété d'un attribut est la minimalité. Un attribut B est minimal si chaque attribut qui lui est subordonné dans l'arbre de catégorisation est pertinent pour une seule valeur de B. Lorsqu'on s'aperçoit qu'un attribut n'est pas minimal, ceci amène une refonte de l'arbre pour le faire remonter au bon niveau.

Le processus que met en place Anadia permet donc de montrer à l'agent des aspects de sa représentation dont il n'avait pas forcément conscience et l'invite (par l'examen des indices précédents) à affiner sa représentation lorsque cela est nécessaire et possible (ce n'est pas toujours le cas, et il peut rester des catégories vides qu'on ne peut pas éliminer). Par un tel processus, l'agent construit et partage avec le système des représentations. Il améliore, par là, sa connaissance du monde et ceci lui évite de considérer dans ses représentations des critères superflus.

2.3. Le lexème dans le modèle Anadia

Le modèle Anadia peut être appliqué à la représentation componentielle et différentielle des signes linguistiques en considérant que la modélisation de la signification suit les mêmes principes que la catégorisation. Seuls les buts des représentations à construire changent mais le processus interactif de mise en place d'un système qui garantisse des propriétés conceptuelles ainsi que des propriétés algébriques reste identique. Qu'il s'agisse de former des catégories ou des classes sémantiques, ce processus est dépendant de l'agent humain et de ses connaissances (relatives au corpus traité). La construction du signifié n'a donc pas pour but de constituer un savoir encyclopédique mais de présenter les raisons de l'interprétation d'énoncés d'un corpus par un agent humain, ce qu'Eco exprime par les mots suivants :

> "Donc, le problème sémiotique de la construction du contenu comme signifié est étroitement solidaire du problème de la perception et de la connaissance comme attribution de sens à l'expérience" ([Eco 88, p. 81])

La tâche qui nous incombe est de représenter le contenu sémantique des mots dans un système afin que leur valeur puisse donner lieu à une description componentielle. Lorsque plusieurs mots sont co-présents dans le même texte ou le même dialogue, ce sont ces descriptions qui vont "s'accorder" dans le co-texte pour mettre en évidence les effets de sens issus de ces co-présences. Dans le chapitre suivant, on décrira ces co-adaptations syntagmatiques des descriptions componentielles à l'aide de la notion d'isotopie, mais il faut au préalable poser un modèle du sème.

2.3.1. Le sème

Le modèle du sème que nous nous sommes donné trouve un mode représentation pour une machine dans le modèle Anadia, présenté précédemment, et plus précisément à travers sa notion d'attribut. Ainsi, le domaine d'interprétation d'un sème est représenté par la propriété d'un attribut et les traits qui constituent les oppositions forment les valeurs de l'attribut. Il faut donc voir dans la liste des valeurs d'un attribut une représentation des oppositions possibles entre ces valeurs. Ainsi, un attribut à trois valeurs *a, b* et *c* représente les oppositions *a vs. b, a vs. c, b vs. c* (on note que l'on ne fait pas différence entre l'opposition *a vs. b* et *b vs. a*) . Par exemple, le sème formé par les oppositions liquide *vs.* solide *vs.* gazeux dans le domaine d'interprétation de l'état physique trouve donc une représentation dans l'attribut dont la propriété est une question sur la nature de l'état physique d'une chose et dont les valeurs sont *liquide, solide* et *gazeux.* De même, le sème dont le domaine d'interprétation dénote les êtres vivants, et contenant les oppositions animal *vs.* végétal et animal *vs.* humain, est représenté par l'attribut

dont la propriété est une question sur l'appartenance à un type d'êtres vivants et dont les valeurs sont *animal, végétal, humain.*

État physique
Solide
Liquide
Gazeux

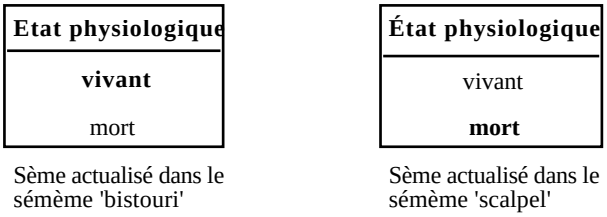
Êtres vivants
Animal
Végétal
Humain

Figure 3.10 : Représentation de sèmes oppositionnels par des attributs Anadia

Les sèmes, considérés comme des attributs, peuvent alors être regroupés en registres. La combinatoire qui en est issue permet de mettre en évidence, par des jeux de valeurs, les significations engendrées par les sèmes retenus ; ce qui conduit à former des sémèmes.

2.3.2. Le sémème

Comme on l'a signalé, la définition la plus immédiate du sémème consiste à le considérer comme un ensemble de sèmes. Pour reprendre un exemple classique de sémantique componentielle, on définira le sémème 'chaise' par l'ensemble de sèmes {/pour une personne/, /sans bras/, /avec dossier/}. Nous ne remettons pas en cause cette définition et, de notre point de vue, le sémème sera également défini comme un ensemble de sèmes, à la différence près que les sèmes dont il s'agit sont tels qu'on les a détaillés dans la partie précédente. Le sémème est composé d'attributs au sens du modèle Anadia qui représentent les sèmes formant le sémème. Ces sèmes sont actualisés au sein du sémème dans le sens où, pour chacun d'entre eux, une valeur de l'opposition de l'attribut est retenue comme représentative de la signification que représente le sémème. Reprenons, par exemple, le cas des significations de *bistouri* et de *scalpel* qui partagent le même sème vivant *vs*. mort résultant de la partition du domaine d'interprétation de l'état physiologique. Le sémème 'bistouri' actualise ce sème dans la valeur vivant tandis que 'scalpel' actualise le même sème dans la valeur mort. C'est donc la différence d'actualisation des sèmes qui différencie les sémèmes construits avec les mêmes sèmes.



Sème actualisé dans le sémème 'bistouri'

Sème actualisé dans le sémème 'scalpel'

Figure 3.11 : Actualisation d'un sème dans un sémème

Constituer un sémème revient donc à considérer un ensemble de sèmes et à les actualiser. Cette double tâche est gérée dans le processus interactif d'Anadia. Le choix des sèmes correspond à la construction d'un registre formé par les attributs qui représentent ces sèmes et leurs actualisations sont réalisées par la reconnaissance du sémème dans la combinatoire du registre. Par exemple, pour constituer le sémème représentant la signification du mot *bus* en tant que transport collectif, on peut construire un registre formé des deux attributs correspondant aux sèmes suivants :

sème n°1 :	domaine d'interprétation :	Portée du transport
	opposition :	intra-urbain *vs*. extra-urbain
sème n°2 :	domaine d'interprétation :	Mode de transport
	opposition :	routier *vs*. ferroviaire

Le développement de la combinatoire du registre fournit une table à quatre places dans laquelle on reconnaît la signification de *bus* dans les valeurs intra-urbain et routier[54]. Ces deux valeurs déterminent l'actualisation des sèmes 1 et 2 dans le sémème 'bus'. De même, on extrait de cette table les sémèmes représentant les significations de *métro, autocar* et *train.*

	Portée du transport	Mode de transport
'bus'	intra-urbain	routier
'métro'	intra-urbain	ferroviaire
'autocar'	extra-urbain	routier
'train'	extra-urbain	ferroviaire

[54] Cet exemple est relativement arbitraire car il existe, de plus, des bus extra-urbains (c'est le cas des *bus verts* dans le Calvados qui relient Caen aux autres villes du département).

Sémème du mot *bus* :

Porté du transport	Mode de transport
intra-urbain	**routier**
extra-urbain	ferroviaire

Figure 3.12 : Un exemple de sémème

Comme le montre cet exemple, le sémème est directement issu du processus interactif d'Anadia qui débute avec le choix des sèmes et la constitution d'un registre. Dans ce processus interactif, le registre est un outil de construction de sémèmes et, dans son rapport à ce registre, le sème acquiert un contenu formel (à la différence de la définition classique du sème) qui correspond au développement de la combinatoire qu'il engendre dans le registre où il apparaît.

Dans cet exemple, nous avons exprimé le sémème que nous recherchions (celui de *bus*) en développant une combinatoire de 2 sèmes. Cet exemple est particulièrement simple mais, dans d'autres cas, on pourrait considérer plus de sèmes (notamment on aurait pu considérer en plus, un sème introduisant l'opposition souterrain *vs*. en surface dans le domaine des transports terrestres, pour faire la différence entre les sémèmes 'train' et 'RER'). La table et la topique comporteraient alors plus de 4 places, formant un système de signification plus complexe. De plus, cet exemple est simple du fait qu'il est isolé de toute autre représentation (ce qui n'est pas le but recherché). Ainsi, si on avait auparavant exprimé un sémème pour *véhicule*, on aurait pu considérer les sèmes retenus comme réalisant une sous-catégorisation des véhicules dans laquelle on aurait exprimé les représentations componentielles de *bus*, *métro*, *autocar* et *train.* Le sémème de *bus* comporterait alors, en plus des sèmes 1 et 2, les sèmes actualisés du sémème 'véhicule'. On rejoint en cela l'approche de Martin ([Martin 83, p. 65]) qui définit le sémème de façon récursive (notamment par ajouts de sèmes).

Les sèmes qui composent un sémème ne sont pas toujours exclusivement ceux qui forment le registre ayant servi à construire le sémème. Comme nous l'avons expliqué en détaillant le modèle Anadia, les tables prennent place dans une hiérarchie de tables formant plusieurs niveaux de catégorisation. Cette hiérarchie permet de calculer une représentation componentielle pour chaque élément de la sélection des tables en considérant les attributs de la table où cet élément apparaît ainsi que les attributs des tables précédentes dans la hiérarchie. Dans le cas général, la table issue du registre s'inscrit dans une hiérarchie ; le sémème est alors formé par la représentation componentielle issue de l'ensemble de la hiérarchie. C'est donc en formant plusieurs registres, en remplissant plusieurs tables et en redéveloppant des lignes des sélections des tables que l'on crée, ou que l'on redétaille, des sémèmes.

Par exemple, considérons un passage du corpus PIC (cf. Chapitre 1) où le rédacteur explique à l'utilisateur les différents types possibles d'aide à l'utilisation du logiciel (ce passage est d'une durée de 5 min. et 15 sec. et se situe dans le deuxième quart d'heure de conversation). Dans ce passage, le rédacteur évoque deux grandes distinctions dans les outils d'aides : ceux dont le support est le papier en opposition à ceux dont le support est logiciel et ceux où la présentation de l'information est globale en opposition à ceux où la présentation de l'information est centrée sur un point précis du logiciel. Ces distinctions permettent d'opposer trois formes d'outils d'aide, verbalisées dans l'interaction entre le rédacteur et l'utilisateur : les *manuels*, les *référentiels* et les *aides en ligne.* En analysant ce passage, on peut se donner deux sèmes qui représentent les distinctions formulées par le rédacteur :

sème n° 1 :	domaine d'interprétation :	Support de l'information
	oppositions :	papier *vs*. logiciel
sème n° 2 :	domaine d'interprétation :	Présentation de l'information
	opposition :	globale *vs*. précise

En formant un registre avec ces deux sèmes, on met en place une combinatoire à quatre places. Dans cette combinatoire, on retrouve les actualisations de sèmes correspondant aux sémèmes décrivant les outils d'aides verbalisés (manuels, référentiels et aides en ligne)[55] :

	Support de l'information	Présentation de l'information
'manuel'	papier	globale
'référentiel'	papier	précise
	logiciel	globale
'aide en ligne'	logiciel	précise

[55] La place vide dans cette table correspond à un jeu d'actualisations de sèmes qui n'est pas verbalisé dans l'interaction (ce pourquoi, on l'a laissée vide). Cependant, on pourrait ramener les valeurs *Support de l'information logiciel* et *Présentation de l'information globale* aux fichiers souvent appelés "lisez moi" qui accompagnent les logiciels. On définirait alors, dans cette place de la table, le sémème de la lexie *lisez moi*.

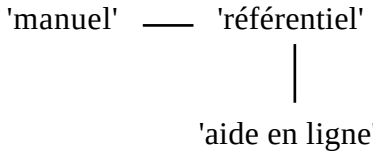


Figure 3.13 : Topiques des outils d'aide

Dans ce même passage du corpus, le rédacteur précise, qu'en fait, l'aide en ligne n'est pas en soi un outil (au même titre que les manuels et les référentiels) mais, en fin de compte, une catégorie d'outils tels que les *assistants*, les *sommaires*, les *index* et les recherches par mots clés (qu'il appelle *recherche full text*). En analysant l'interaction, on peut mettre en évidence deux distinctions entre ces quatre formes d'aide en ligne qui correspondent aux deux questions suivantes :

- Est-ce que le mode d'accès est interactif ou figé ?
- Est-ce que l'information, telle qu'elle est présentée, est hiérarchiquement structurée ou linéaire ?

À ces deux questions, on peut faire correspondre deux sèmes avec lesquels on va pouvoir constituer un nouveau registre, et donc former une nouvelle combinatoire dans laquelle on va apparier les actualisations des deux nouveaux sèmes aux sémèmes 'assistant', 'sommaire', 'index' et 'full text'.

sème n° 3 :	domaine d'interprétation :	Accès à l'information
	opposition :	interactif *vs*. figé
sème n° 4 :	domaine d'interprétation :	Structure de l'information
	opposition :	hiérarchique *vs*. linéaire

	Accès à l'information	Structure de l'information
'assistant'	interactif	hiérarchique
'full text'	interactif	linéaire
'sommaire'	figé	hiérarchique
'index'	figé	linéaire

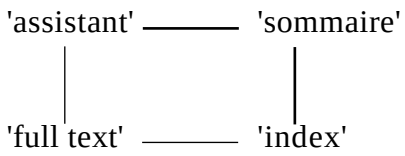


Figure 3.14 : Topiques des aides en ligne

Dans ces deux phases de l'analyse de la séquence du corpus, on a donc mis en forme un système de signification comportant deux niveaux : celui des outils d'aide et celui des aides en ligne.

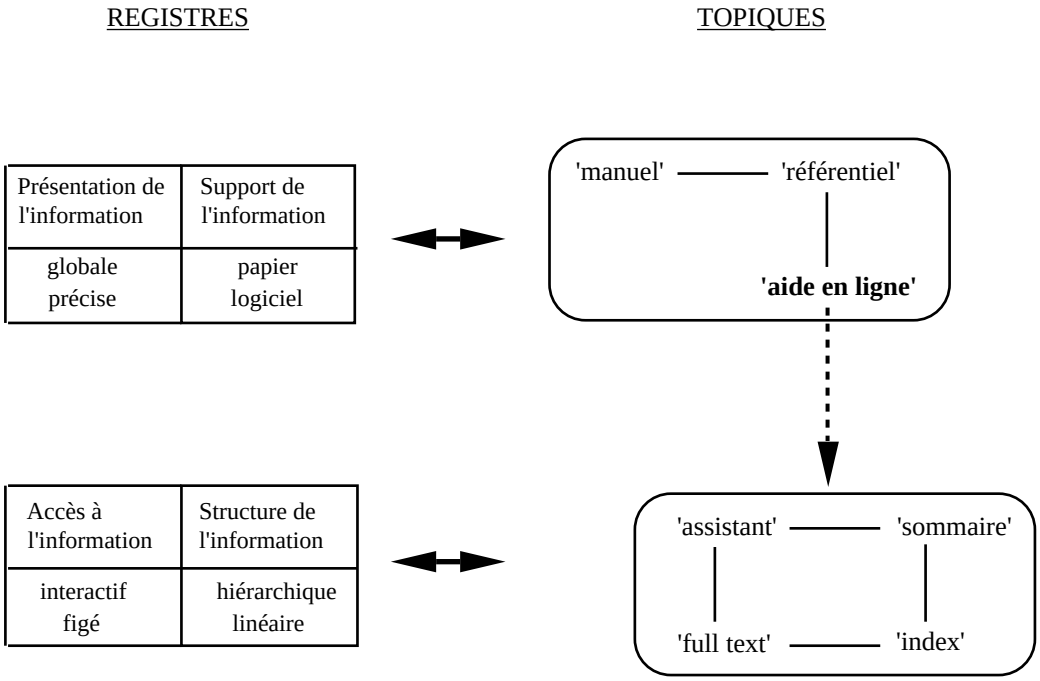


Figure 3.15 : Un système de signification
(les doubles flèches représentent le processus interactif de création d'une topique à partir d'un registre et la flèche en pointillés représentent le processus réappliqué à un élément de la sélection d'une table)

De ce système de signification, on peut extraire la description componentielle des sémèmes indiqués dans les topiques. Par exemple, le sémème du lexème *assistant* comporte quatre sèmes actualisés : ceux de la table des outils d'aide et ceux de la table des aides en ligne :

Présentation de l'information	Support de l'information	Accès à l'information	Structure de l'information
globale **précise**	papier **logiciel**	**interactif** figé	**hiérarchique** linéaire

Figure 3.16 : Un sémème non limité aux sèmes d'un seul registre

Cet exemple n'a aucune valeur dogmatique ou encyclopédique car les sèmes qu'il met en jeu sont choisis et justifiés selon un point de vue relativement arbitraire : ils dépendent de l'interprétation qu'on a de l'interaction et de la situation de l'interaction (en fonction de l'interprétant, ils auraient pu être différents). Cependant, ils peuvent toujours être renégociés par la suite dans le processus interactif d'Anadia. Ils permettent de donner forme au *terrain commun* construit dans les interactions langagières.

Dans cet exemple, on a donné à *aide en ligne* un statut de catégorie, dans le sens où son sémème participe à la construction de plusieurs autres sémèmes, et un statut de lexie en attribuant à la réunion des deux mots un seul sémème. La signification d'une lexie est toujours définie de façon paradigmatique, même quand elle met en jeu plusieurs signes qui ont d'autre part chacun leur signification propre. Dans un tel cas de lexie composée de plusieurs signes, on invoquera directement au niveau syntagmatique la signification de la lexie dans la chaîne où elle apparaît. C'est-à-dire que l'on ne cherchera donc pas de dépendance sémantique d'un signe sur l'autre quand ces signes constituent une lexie. Quand ce n'est pas le cas, les sémèmes de la chaîne seront affectés par les relations de dépendance sémantique de l'énoncé. Les relations entre les dépendances sémantiques et les sémèmes mettent en jeu des opérations interprétatives qui relèvent de l'ordre du syntagmatique et que l'on détaillera dans le chapitre suivant.

Cet exemple montre également que, dans la description componentielle produite par le processus de catégorisation d'Anadia, on donne au sémème une structure. Cette structure fait que le sémème ne se réduit pas à un "paquet de sèmes" (comme dans [Ducrot & al. 95, p. 446]). On retrouve, dans cette structure, la différence introduite en sémantique différentielle entre sèmes génériques et sèmes spécifiques. Ainsi, les sèmes spécifiques d'un sémème sont représentés par les attributs du registre ayant permis de constituer la table dans laquelle le sémème est défini, et les sèmes génériques sont ceux qui proviennent des tables précédentes dans la hiérarchie de catégorisation. Les sèmes génériques du sémème sont donc les attributs dont ses sèmes spécifiques sont des attributs subordonnés[56].

[56] Dans la partie sur le processus de catégorisation mis en place dans Anadia, on a défini la subordination d'attributs en précisant qu'un attribut A est subordonné à un attribut B si A est pertinent seuleument quand B a certaines valeurs.

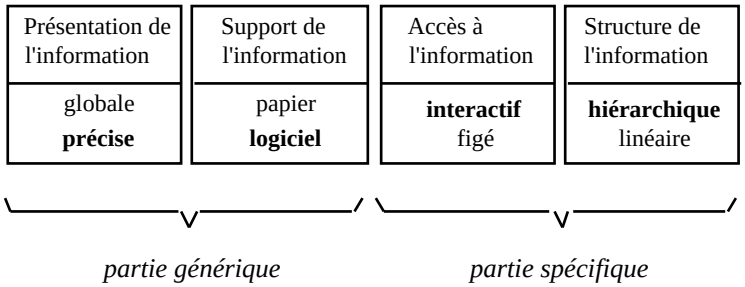


Figure 3.17 : Structure du sémème 'assistant'

Si par la suite on redéveloppait la catégorie pointée par le sémème 'assistant', ce sémème (tel qu'on l'a définit) formerait alors la partie générique des sémèmes que la sous-catégorisation permettrait de mettre en évidence. De même, si on définissait dans une table un sémème pour la lexie *outil d'aide* (ou *documentation*, comme nous le verrons au chapitre 5), alors les sèmes constituant ce sémème viendraient enrichir la partie générique de tous les sémèmes que l'on a défini dans notre exemple ('manuel', 'référentiel', 'aide en ligne', 'assistant', 'sommaire', 'index' et 'full text').

2.3.3. Le taxème

Comme nous venons de le voir avec le sémème, le processus de catégorisation d'Anadia permet de structurer les descriptions componentielles. Il en est de même pour le taxème. Selon la sémantique différentielle, un taxème correspond à une classe de sémèmes indexés par les mêmes sèmes génériques et définis différentiellement les uns et les autres par les mêmes sèmes spécifiques. C'est-à-dire qu'un taxème regroupe des significations sémantiquement très proches. Dans Anadia, nous avons exprimé formellement la proximité des représentations par la relation binaire de *différence à un trait près*. Cette relation nous a permis de constituer les graphes que l'on a appelés topiques. Le taxème, dans le modèle sémantique de l'axe paradigmatique basé sur Anadia, regroupe donc les sémèmes organisés en topique. Le taxème se trouve alors structuré en tant que réseau de différences à un trait près.

Par exemple, la combinatoire des sèmes [Accès à l'information : interactif *vs*. figé] et [Structure de l'information : hiérarchique *vs*. linéaire] nous a permis de mettre en évidence une topique qui montre la structure différentielle du taxème {'assistant', 'sommaire', 'index', 'full text'}. On a ici un cas où la topique constituée par les sélections de la table forme un taxème très fortement structuré car, pour les quatre sémèmes du taxème, on a quatre relations de différences à un trait près entre les sémèmes considérés deux à deux ('assistant' - 'sommaire', 'sommaire' - 'index', 'index' - 'full text', 'full text' - 'assistant'). Ce n'est pas toujours le cas puisque la topique révèle parfois plusieurs sous-classes dans le taxème. En effet, pour qu'une topique ne

représente qu'un seul réseau de différences, il faut qu'elle ne comprenne qu'une seule composante connexe. Lorsqu'une topique est formée de plusieurs composantes connexes, on a autant de "sous-classes taxémiques" que de composantes connexes. Mettre en évidence ces sous-classes permet de préciser un degré de proximité sémantique entre les membres d'un même taxème. Pour un sémème donné (que l'on notera S), on peut alors distinguer trois niveaux de proximité paradigmatique avec les autres sémèmes de son taxème :

• Premier niveau : les sémèmes qui sont des voisins de S dans la topique. Ce premier niveau représente les antonymes de S, différents de S par une seule différence d'actualisation de sème spécifique.

• Deuxième niveau : les sémèmes qui, sans être voisins de S dans la topique, sont des sommets de la même composante connexe que S ; c'est-à-dire, les sémèmes qui se distinguent de S par plusieurs différences, telles que chacune d'entre elles consiste en un rapport différentiel entre deux sémèmes (i.e. il y a un chemin dans la topique entre les deux sémèmes).

• Troisième niveau : les sémèmes des autres composantes connexes de la topique. Ce sont ceux qui, tout en étant définis par les mêmes sèmes spécifiques, sont distincts de S par au moins deux différences d'actualisation de sèmes (spécifiques).

Si, par exemple, on cherche à représenter les significations des mots *bateau, barque, car, avion, planeur* et *vélo*, on peut se donner les sèmes suivants [Milieu : air *vs*. terre *vs*. eau], [Capacité de transport : collectif *vs*. individuel] et [Propulsion : motorisée *vs*. non motorisée]. On forme les sémèmes par les actualisations de sèmes suivantes :

	Milieu	**Capacité de transport**	**Propulsion**
'avion'	air	collectif	motorisée
'planeur'	air	individuel	non motorisée
'car'	terre	collectif	motorisée
'vélo'	terre	individuel	non motorisée
'bateau'	eau	collectif	motorisée
'barque'	eau	individuel	non motorisée

La topique formée comporte deux composantes connexes, l'une regroupant les sémèmes 'bateau', 'car' et 'avion' et l'autre regroupant les sémèmes 'planeur', 'vélo' et 'barque'. On met donc en évidence deux sous-classes du taxème (cf. figure 3.18).

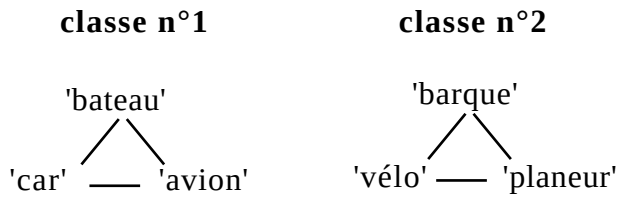


Figure 3.18 : Un taxème divisé en deux classes

Le processus interactif d'Anadia a pour but d'optimiser les propriétés algébriques des représentations ; c'est-à-dire, en ce qui concerne les topiques, chercher le plus possible à augmenter leur connexité. Le but n'est donc pas de subdiviser les taxèmes en plusieurs classes, mais de chercher, au contraire, à les constituer de façon indivisible. Ceci revient à ne pas introduire au sein d'un taxème un troisième niveau de proximité paradigmatique. On rejoint en cela le point de vue de la sémantique différentielle, pour laquelle le taxème est la classe minimale en langue. Lorsqu'un taxème apparaît formé de plusieurs classes, il convient de chercher d'éventuels défauts dans sa formation.

La non-connexité de la topique est liée à l'importance de places vides dans la table. Il y a donc deux raisons possibles pour expliquer un défaut de construction d'un taxème : soit la combinatoire engendre des places vides parce que les sèmes sont sémantiquement dépendants (i.e. certains sont subordonnés à d'autres, ce qui implique que les sèmes choisis ne sont pas tous spécifiques mais que certains sont génériques et doivent être utilisés à un niveau plus haut de hiérarchie de table), soit tous les sémèmes du taxème n'ont pas été décrits dans la table. C'est le cas dans notre exemple des moyens de transport car les sèmes ne sont pas sémantiquement dépendants, et seulement six sémèmes sont pris en compte alors que la combinatoire de la table fournit douze jeux d'actualisation de sèmes. Notamment, parmi les places restée vides, on peut mettre en évidence le sémème 'voiture' (par les valeur d'actualisation de sèmes terre, individuel, motorisée). On réalise ainsi une connexion entre les deux composantes connexes, et le taxème ne représente plus qu'une seule classe.

Figure 3.19 : Un taxème "reconnecté"

Comme on vient de le voir, on conçoit donc le taxème comme un résultat du processus de catégorisation, ce que Rastier souligne également dans le cadre d'une analyse linguistique :

> Par leur structure différentielle, les taxèmes reflètent les conditions perceptibles générales qui font de l'activité linguistique un processus de discrétisation et de catégorisation. ([Rastier & al. 94, p. 62])

Dans le cadre d'une description conponentielle, l'intérêt d'Anadia est d'opérationnaliser dans un modèle informatique une méthode (interactive) de construction de taxèmes.

3 Conclusions

Nous avons montré, dans ce chapitre, qu'il est possible de modéliser la sémantique des langues naturelles avec des principes de catégorisation sans pour autant constituer une sémantique référentielle. Pour faire ce lien entre catégorisation et signification, il faut concevoir la catégorisation comme un processus dirigé par l'examen des différences entre les choses et non comme un classement guidé par les ressemblances (ce qui serait le cas de la classification).

Dans ce but, nous avons proposé le modèle de catégorisation Anadia et nous avons montré les liens avec les principes de base de la sémantique componentielle. Cependant ces principes, tels qu'ils apparaissent dans les travaux des linguistes, ne constituent qu'un outillage descriptif de la langue. Nous avons, de ce fait, dû les réexprimer pour les rendre opérationnels au regard d'un système de représentation pour une machine. Ceci nous permet de renforcer le statut de la différence. Pour la linguistique, le sème est "atomique" et doit son existence au sein d'un sémème à sa pertinence, c'est-à-dire à sa place comme extrémité d'une relation de différence entre deux significations. De notre point de vue, le sème est constitué par une relation de différence (non nécessairement binaire), il n'est donc plus "atomique" et constitue une entité négative. La question de la pertinence du sème est toujours en rapport avec des différences de signification et elle est, comme on l'a expliqué, directement analysable dans le processus interactif de construction d'un système de signification mis en place dans Anadia (notamment par un examen de la connexité des topiques produites).

Enfin, notre démarche de conception par amorce se justifie par le fait qu'un modèle de l'axe paradigmatique ne peut être conçu de façon autonome. Il est nécessairement construit dans un rapport à un modèle syntagmatique du sens des énoncés car il apparaît, avec le concept de taxème, que l'apprentissage d'une langue est indissociable de sa pratique. Le taxème, en matérialisant des significations formellement très proches, indique que ces significations sont semblables à un certain niveau. C'est-à-dire que si dans une chaîne, deux éléments d'un même taxème apparaissent, ils vont donner lieu à de fortes récurrences de sèmes (i.e. des isotopies). On peut alors considérer le taxème sous deux points de vue : soit comme une "cristallisation" paradigmatique d'effets de sens fortement marqués au niveau syntagmatique, soit comme un moteur de l'activité d'interprétation d'une chaîne linguistique où il permet les co-adaptations co-textuelles. Ces deux points de vue reviennent à considérer la langue dans ses deux dimensions inséparables que sont l'acquisition et l'usage d'un système sémiotique. Notre modèle de l'axe

paradigmatique va permettre de mettre en évidence des effets de sens du niveau syntagmatique (notamment les cas de dépendance sémantique que nous avons décrit au chapitre 2). En retour, ces effets de sens vont permettre la construction d'un système de valeurs paradigmatique.

La modélisation de l'axe syntagmatique fait l'objet du chapitre suivant. La notion centrale sur laquelle est fondée cette modélisation est la notion d'isotopie. Alors que le paradigmatique est un ordre déterminé par la différence, nous expliquerons que le syntagmatique est un ordre déterminé par la récurrence. Il restera ensuite à montrer comment les deux axes déterminent conjointement la production du sens (cf. chapitre 5).

CHAPITRE 4 :

UN MODÈLE SÉMANTIQUE DE L'AXE SYNTAGMATIQUE

Le chapitre précédent concluait sur la nécessité de mettre en place un modèle sémantique de la chaîne pour fonder la modélisation différentielle des signifiés. Auparavant, au cours du chapitre 2, nous avons montré que, pour la mise en place du potentiel de sens de l'énoncé, l'analyse de la chaîne ne se réduit pas à sa syntaxe mais que la co-présence de ses constituants est un facteur essentiel. Nous allons ici faire le lien entre ces deux chapitres en proposant un modèle opératoire de la co-adaptation syntagmatique des significations. C'est donc dans ce chapitre que sera mise en oeuvre l'approche interprétative annoncée dans le premier chapitre.

Le modèle de l'axe syntagmatique proposé ici est fortement influencé par la sémantique interprétative ([Rastier 87]). Ceci explique que nous emprunterons à Rastier un bon nombre d'exemples pour montrer en quoi, dans une perspective informatique, nous nous inspirons et nous nous démarquons de ses travaux. La première partie décrit comment la sémantique interprétative rend compte des effets de sens syntagmatiques. Les effets de l'ensemble de la chaîne sur les significations qu'elle met en jeu y sont considérés comme le fait *d'opérations interprétatives.* Ces opérations interprétatives induisent des récurrences de sèmes le long de la chaîne. Dans une deuxième partie, cette notion de récurrence sémique sera précisée, et l'on rappellera la définition du concept clé du modèle de l'interprétation : l'**isotopie**. Dans une troisième partie, l'isotopie sera rapportée à la récurrence de sèmes oppositionnels (cf. chapitre précédent), ce qui permet de rendre compte des effets de sens syntagmatiques. L'isotopie, loin d'être limitée à un simple résultat des redondances de la chaîne, sera vue comme le "moteur" de son interprétation par une machine. Il en découle deux propriétés d'un énoncé : sa cohésion et sa

cohérence. La cohésion sera liée à l'intelligibilité de l'énoncé (sa mise en évidence est le but de l'analyse interprétative de la machine) et la cohérence permettra d'envisager les extensions du modèle pour le traitement des schémas argumentatifs du discours et de l'enchaînement conversationnel.

1. Les opérations interprétatives

C'est parce que les langues naturelles ne fonctionnent pas sur un mode strictement compositionnel comme les langages formels, mais qu'elles procèdent par répétitions et différenciations, que l'analyse d'une chaîne linguistique ne se réduit pas à la mise au jour de sa structure syntaxique "décorée" par la sémantique. Comme nous l'avons montré au chapitre 2, les significations qui composent la chaîne sont l'objet d'effets de sens directement dûs à l'enchaînement syntagmatique. Pour décrire ces effets de sens, la sémantique interprétative définit des modes d'adaptation des significations dans leur co-texte. Ces modes d'adaptation sont les résultats d'une interprétation autant qu'ils la constituent et sont le fait du sujet interprétant. Ils constituent des opérations interprétatives.

L'étude des opérations interprétatives concerne la dynamique des sèmes qui apparaissent dans les significations des constituants de la chaîne. Ce n'est donc pas directement de la constitution du signifié qu'il s'agit mais de sa mise en co-texte. C'est-à-dire que l'objet de l'analyse n'est plus le même. On ne se limite plus exclusivement à la sémie (signifié de la lexie), comme c'est le cas en microsémantique, mais on considère le palier d'analyse qui s'étend de la formation du syntagme aux suites d'énoncés en relation de dépendance sémantique. On se focalise donc sur une zone de localité sémantique définissable par l'étendue dans le discours des relations de prédication et d'anaphore. Cette zone, appelée *Période* ([Rastier & al. 94, p. 116]), définit le champ d'étude de la mésosémantique.

Pour poser la question de la dynamique des sèmes dans la période, il convient de préciser quels sont les sèmes en cause. Ce problème de la sélection des sèmes dans la chaîne syntagmatique constitue une première opération interprétative qui conduit à retenir les sémèmes les plus pertinents pour les composants de la chaîne. Ce premier pas de l'interprétation amène à réduire, dans le co-texte, la polysémie lexicale des éléments de la chaîne. Comme l'ont montré de nombreux exemples du chapitre 2, cette réduction de la polysémie lexicale est le fait des dépendances sémantiques du co-texte. Ainsi, dans *Je mange un avocat*, c'est parce qu'il y a une dépendance sémantique entre *mange* et *avocat* que la polysémie de *avocat* ne pose pas de problème et que l'on ne retient pas *a priori* le sémème décrivant la signification d'homme de loi.

Comme on l'a expliqué au chapitre 2, les relations de dépendances sémantiques tendent à renforcer les propriétés communes des composants de la chaîne et à effacer celles qui ne donnent pas lieu à une répétition dans le co-texte. Ce sont ces renforcements et ces effacements que l'on cherche à mettre au jour en étudiant la dynamique des sèmes dans le co-texte. Ils correspondent à deux formes d'opérations interprétatives : *l'actualisation* et la *virtualisation.* L'actualisation consiste à identifier un sème dans un contexte. Par exemple, dans l'énoncé

Le facteur m'a donné une lettre.

Le sème /courrier/ est actualisé dans le sémème 'lettre' parce qu'il se répète dans le sémème 'facteur'. Cette actualisation permet de retenir la signification pertinente de *lettre* dans l'énoncé (on ne retient donc pas, par exemple, la signification de *lettre* en tant que caractère de l'alphabet) et précise une sélection du co-texte sur une partie du signifié de *lettre.* Ainsi, dans cet exemple, le sème /courrier/ est renforcé par le co-texte alors que ce n'est notamment pas le cas du sème /en papier/ appartenant également au sémème 'lettre' ; à l'inverse, ce sème serait probablement actualisé dans *Il a chiffonné sa lettre* et pas /courrier/. La virtualisation est l'opération interprétative duale de l'actualisation. Elle décrit une neutralisation d'un sème en contexte. Par exemple, dans le syntagme *Le chat immortel,* on dira que le sème /mortel/ appartenant au sémème 'chat' est virtualisé car non seulement il n'est pas répété dans l'énoncé mais, de plus, il est invalidé par le contenu sémique de 'immortel'. La virtualisation ne doit pas être considérée comme un simple retrait d'un sème dans un sémème. L'idée d'une neutralisation temporaire est plus juste car, si dans la suite du discours le sème virtualisé venait à réapparaître dans d'autres sémèmes, il serait alors actualisé. Dans notre exemple, ce serait alors le contenu sémique de 'immortel' qui serait virtualisé.

Comme on vient de le voir, l'actualisation et la virtualisation jouent un rôle important sur la mise en co-texte de contenus sémiques définis dans l'ordre paradigmatique, c'est-à-dire sur les sèmes inhérents des lexèmes de la chaîne. Cette notion de sème inhérent est à opposer à celle de sème afférent. Cette distinction repose sur le fait que, dans des contextes particuliers, on peut actualiser dans des sémèmes des sèmes qui ne font pas partie de leur définition paradigmatique. Ces sèmes sont dits afférents et l'opération interprétative qui consiste à les actualiser porte le nom d'afférence. L'afférence peut prendre deux formes décrivant chacune une actualisation particulière : l'assimilation et l'afférence dite socialement normée.

L'assimilation est l'opération interprétative qui décrit l'effet de sens que l'on a constaté au chapitre n°2 dans la dépendance sémantique liée aux énumérations. Ainsi dans l'énoncé

Voici des choux, des concombres et des scoubidous.

où le sémème 'scoubidous' est à préciser, l'assimilation consiste en une afférence d'un sème à 'scoubidou' où le but est de renforcer une répétition déjà initiée dans le co-texte. En

l'occurrence, il s'agit d'enrichir le contenu sémique de 'scoubidou' avec le sème afférent /légume/ pour rendre le thème de l'énoncé uniforme[57]. L'affférence due à l'assimilation consiste à propager à un sémème un ou plusieurs sèmes inhérents des autres sémèmes de l'énoncé. Ces sèmes sont des sèmes génériques dans les sémèmes où ils sont inhérents et c'est cette récurrence de généricité qui rend le thème uniforme (dans notre exemple, c'est le sème /légume/ qui est afférent et pas le sème /vert/ inhérent à 'choux' et 'concombre' mais spécifique, ce qui a pour conséquence qu'il n'y ait pas de défaut d'assimilation dans *Voici des choux, des concombres et des carottes*). L'assimilation est donc une afférence co-textuelle (i.e. le sème afférent est inhérent dans d'autres sémèmes du co-texte) d'un sème générique par présomption d'une récurrence sémique (ce que nous appellerons bientôt isotopie générique).

L'assimilation est aussi l'afférence qui décrit le rattachement des anaphoriques à leurs antécédents. Par exemple, le sème /objet/ n'est pas défini dans le contenu sémique du pronom 'il' mais il y est afférent dans l'énoncé :

J'ai retrouvé mon stylo, il était caché.

Le cas des pronoms anaphoriques est, il faut l'avouer, un peu particulier étant donné que, par définition, leur contenu sémique est vide et qu'ils demandent à être sémantiquement saturés en contexte. Ainsi, les anaphoriques n'ont pas de sèmes inhérents et n'ont (en contexte) que des sèmes afférents. Le cas de mots inconnus est sensiblement similaire car leurs contenus sémiques sont également vides (étant donnés qu'ils sont inconnus). Ainsi, la seule façon d'y actualiser des sèmes est de réaliser des afférences[58]. Cependant, à la différence des anaphoriques, l'afférence a ici une incidence paradigmatique car les sèmes afférents sont "cristallisés" dans le contenu sémique du mot et y deviennent inhérents. De ce point de vue, c'est en contextualisant des mots que l'on fixe leurs contenus. L'afférence, en décrivant un effet du syntagmatique sur le paradigmatique, permet l'acquisition des langues par leur pratique.

L'opération interprétative inverse de l'assimilation est la dissimilation. Alors que l'assimilation diminue, par afférence, les contrastes forts, la dissimilation, quant à elle, augmente les contrastes faibles. La dynamique des sèmes en cause dans la dissimilation n'est plus une afférence de sème générique, comme c'est le cas pour l'assimilation, mais une afférence de sèmes spécifiques pour différencier, dans un co-texte, les sémèmes appartenant au même taxème. C'est le cas lorsqu'un mot générique est utilisé pour exprimer une idée

[57] Ce type d'assimilation permet une forme d'humour lorsque le sème afférent est inattendu dans le contenu sémique du lexème en cause. Le titre du livre de G. Lakoff *Women, Fire and Dangerous Things : What Categories Reveal about the Mind* (1987) en témoigne en provocant l'afférence du sème /dangereux/ au lexème *Women*.

[58] C'est également de cette façon que la langue permet des jeux d'interprétation comme ceux à l'œuvre dans le langage des *stroumphs*, ou encore dans "Le grand combat" de H. Michaux dont est extraite la phrase *Il l'emparouille et l'endosque contre terre* ([Michaux 66, p. 14]).

spécifique. Par exemple, dans *routes et autoroutes,* le sémème 'route' doit décrire une signification spécifique qui exclut la signification de *autoroute* et non une signification générique qui inclut cette signification. La dissimilation est encore plus flagrante dans l'exemple suivant où il y a répétition de la même lexie dans une relation de dépendance sémantique de coordination :

Il y a musique et musique.

Ici, l'effet de sens que produit la dissimilation consiste, dans une approche descriptive, à distinguer par des sèmes spécifiques les deux significations de *musique*. Ainsi, on peut afférer à la première le sème spécifique /agréable/ et à la seconde /désagréable/. Dans le cadre de notre modèle où le sème résulte d'une opposition (cf. chapitre précédent), on considère qu'une telle dissimilation produit une afférence d'un même sème (portant l'opposition agréable *vs*. désagréable) dans les deux sémèmes 'musique' mais que, pour chacun d'eux, l'afférence actualise le sème de façon différente. Le résultat d'une telle afférence a pour effet paradigmatique de renforcer l'aspect polysémique du mot *musique* en formant deux sémèmes pour une seule lexie. Ces deux sémèmes s'opposent par une unique différence d'actualisation, ce qui a pour conséquence de constituer un taxème (le taxème //musique//) regroupant ces deux sémèmes. Ce genre de dissimilation n'est pas obligatoirement portée par une coordination comme le montre la dissimilation dans l'énoncé *Une femme est une femme*. Elle tient à la répétition d'un même mot dans la chaîne.

Une telle dissimilation due à la répétition d'une même lexie est une opération interprétative qui conduit à ne pas interpréter l'énoncé comme une tautologie. C'est parce que le contenu sémique de la lexie est dédoublé dans l'interprétation de l'énoncé que la loi d'informativité qu'exprime la maxime de quantité de Grice n'est pas mise en défaut. De plus, ce genre de dissimilation semble traduire une propriété générale des signes qui indique que, quand un signe se succède à lui même, alors ses deux occurrences diffèrent. Cette propriété a notamment été mise en évidence par Coursil ([Coursil 92], [Coursil 98]) en ce qui concerne les fonctions portées par les phonèmes dans les syllabes. Par exemple, dans la chaîne de phonèmes *appa,* les deux occurrences du phonème *p* diffèrent par leur fonction dans la constitution des syllabes, le premier est explosé tandis que le second est implosé[59].

L'assimilation et la dissimilation sont des formes d'afférences co-textuelles mais la sémantique interprétative définit également, comme on l'a signalé, une autre forme d'afférence : l'afférence socialement normée. Dans de tels cas d'afférence, il y a bien enrichissement d'un sémème dans un énoncé par un (ou plusieurs) sème(s) (c'est pour cela qu'il s'agit toujours d'une afférence) mais cette fois, le ou les sèmes afférents ne sont pas inhérents dans d'autres sémèmes de l'énoncé. L'afférence est alors le fait d'une norme sociale partagée au

[59] Il en résulte qu'entre les deux *p*, on a une frontière syllabique.

sein d'une communauté linguistique (un *topos* selon [Raccah 97]). C'est, par exemple, le cas du sème /tristesse/ afférent au sémème 'noir' dans *il broie du noir* ou encore le cas du sème /bonheur/ dans 'rose' dans *la vie en rose.*

Ce type d'afférence, décrite au niveau syntagmatique, met en évidence la cristallisation paradigmatique des pratiques sociales. Dans notre approche oppositionnelle du sème, faire une telle afférence (par exemple afférer au sémème 'rouge' le sème résultant de l'opposition violence *vs*. douceur, actualisé dans la valeur violence) consiste à rajouter une différence entre le sémème en cause (en l'occurrence 'rouge') et les autres sémèmes de son taxème. C'est-à-dire que les sémèmes avec lesquels il était en relation de différence par une seule actualisation de sèmes (on les appelle ses antonymes) ne le sont plus forcément étant donné que, du fait du sème afférent, on rajoute entre eux une possible différence. On a donc ici restructuré un taxème (dans notre exemple, si 'rouge' et 'vert' étaient voisins dans la topique et que 'vert' actualise la valeur douceur du sème afférent, les deux sémèmes présentent deux différences d'actualisation et ne sont donc plus des antonymes au sein de leur taxème).

Avec les opérations interprétatives, et plus précisément avec le concept d'afférence, un outil théorique très complet est posé pour la description de la dynamique des sèmes dans la chaîne syntagmatique. Ce n'est donc pas étonnant de voir la sémantique interprétative fréquemment exploitée à des fins de traitement automatique (cette thèse en témoigne, tout comme [Tanguy 97] ou encore [Bassi Acuña 95]). Quand il s'agit d'opérationnaliser les opérations interprétatives (pour ne plus s'en tenir à des descriptions, mais pour établir de manière projective les récurrences sémiques), on trouve l'idée de considérer un sémème type au niveau paradigmatique et un sémème occurrence au niveau syntagmatique, qui hérite du type, et qui s'en distingue par les afférences réalisées en contexte (c'est notamment le cas dans [Bassi Acuña 95]). Ce point de vue sur la contextualisation de la signification nous semble être une "remontée" au niveau conceptuel de considérations liée à l'implémentation. Au niveau conceptuel, il n'y a pas, à notre avis, à considérer de sémèmes types et de sémèmes occurrences parce qu'il n'y a pas de significations occurrences (comme il n'y a pas non plus de significations types). Ce que montrent les opérations interprétatives, ce sont des enrichissements sémiques de sémèmes dûs à l'ensemble du co-texte, et ces enrichissements jouent directement sur la définition même des significations au niveau paradigmatique. C'est ce que l'on a montré quand une afférence amène à restructurer un taxème. De plus, si l'on fait la distinction conceptuelle entre un sémème type et un sémème occurrence, cela revient à dire que c'est le sémème type qui est premier, étant donné que l'occurrence hérite du type et que l'occurrence n'a pas d'effet retour sur le type. C'est-à-dire que le sens en langue naturelle est compositionnel et que les opérations interprétatives sont les "outils" qui réalisent la composition du sens. Si nous avons choisi un point de vue holiste dans cette thèse, c'est parce que ni les significations, ni le

sens ne sont premiers conceptuellement ; ils se construisent les uns par les autres. C'est bien ce que montrent les opérations interprétatives dans leur double effectivité de contextualisation de la signification et de construction de la signification dans l'enchaînement syntagmatique.

De plus, cette distinction entre sémème type et sémème occurrence, conduit à abuser des opérations interprétatives, et notamment de l'afférence dans l'étape d'instanciation du sémème. Ainsi, selon [Bassi Acuña 95, p. 66], dans l'énoncé *La voiture rouge de mon voisin*, l'instanciation du sémème 'voiture' transite par l'afférence des sèmes /rouge/ et /propriété de mon voisin/ au sémème type 'voiture'. Cette condensation sémantique conduit à détruire la structure syntagmatique et à résumer le discours. Tous les sèmes se trouvent alors être récurrents, ce qui ne fait plus de la récurrence sémique une forme remarquable des chaînes linguistiques. Plus gênant encore par rapport à nos hypothèses, un tel point de vue sur le sémème offre une vision référentielle de la signification au niveau syntagmatique et, dans ce cas, on voit difficilement l'intérêt de conserver un appareillage théorique différentiel au niveau paradigmatique. De notre point de vue, le sémème n'entretient aucune relation avec la référence, ni au niveau paradigmatique, ni au niveau syntagmatique.

Donner aux opérations interprétatives une effectivité référentielle pose le problème de la limite entre la sémantique et la pragmatique. Nous ne rediscuterons pas des problèmes théoriques que cela pose, mais nous remarquerons que ceci a pour incidence de retirer sa légitimité à la fonction de monstration des énoncés langagiers. C'est pourtant bien la monstration qui permet de passer d'une sémantique syntagmatique non référentielle aux actualisations pragmatiques de l'énoncé en offrant la possibilité d'exploiter les traits sémantiques comme des critères définitoires d'un possible prototype. Ainsi, de notre point de vue, dans l'énoncé *La voiture rouge de mon voisin*, les interactions sémiques éventuelles précisent (si besoin est) les significations de la chaîne (par exemple on ne retient pas ici pour 'rouge' le sémème comportant le sème afférent actualisant la valeur violence) et la monstration, en mettant en commun les critères définitoires de 'voiture', de 'rouge' et de 'voisin', permet la construction pragmatique d'un objet prototype représentant une voiture dont le champ couleur porte la valeur rouge et le champ propriétaire une référence au voisin du locuteur.

Les opérations interprétatives permettent donc de rendre compte des effets de sens de la chaîne syntagmatique mais, afin d'en conserver la pertinence, il ne faut pas les généraliser abusivement. C'est en conservant leur caractère local que l'on donne toute sa place à la dynamique sémique de la chaîne et que l'on insiste sur l'importance de la récurrence sémique. Nous allons maintenant approfondir cette notion de récurrence.

2. La récurrence dans la chaîne

Quand on compare un énoncé en langue naturelle à une formule logique ou à un énoncé d'un langage formel (par exemple, un langage de programmation), une première différence qui apparaît immédiatement est, qu'à l'inverse des langages formels, la langue naturelle est redondante. Par exemple, la marque du nombre (pluriel ou singulier) est l'objet d'une très forte redondance en langue car elle est marquée dans tous les composants du groupe nominal et dans la désinence du verbe. Ce n'est pas le cas dans les langages formels où le nombre fait éventuellement l'objet d'une unique prédication ou d'une quantification préalable à la prédication.

Langue naturelle :	Les oiseaux chantent.	**3 marques du pluriel**
Logique des prédicats :	$\forall x, Oiseau(x) \rightarrow Chante(x)$	**1 marque de "pluriel"**

Figure 4.1 : Un exemple de la redondance en langue

Plutôt que de considérer la redondance comme une particularité, *a priori* inutile, des langues naturelles, nous allons en faire une base du modèle opératoire que nous proposons. Pour cela, nous utiliserons principalement le concept d'isotopie. L'isotopie fait son apparition dans les travaux de Greimas ([Greimas 66]) pour rendre compte de l'homogénéité du discours par la récurrence de certains traits sémantiques. Dans la terminologie de Greimas, ces traits sont des *classèmes* et ils correspondent à la notion de sème générique. Par la suite, la notion d'isotopie est reprise par Rastier qui l'étend à toutes les sortes de traits sémantiques (i.e. sèmes génériques et sèmes spécifiques) et la définit comme *l'effet de la récurrence syntagmatique d'un même sème* ([Rastier 87, p. 274]). Il fait la différence entre deux formes d'isotopies : celles qui ont une fonction syntaxique et qui sont "obligatoires" (dans la terminologie de Rastier) car prescrites par le système fonctionnel de la langue (par exemples, les accords en nombre ou en genre) et que l'on appellera *isosèmies* ; et celles qui ont une fonction sémantique (qui par rapport aux isosémies sont "facultatives" selon la même terminologie), pour lesquelles on gardera le terme général d'*isotopies.*

2.1. L'isosémie

L'isosémie constitue un premier niveau de description des phénomènes de récurrence dans une chaîne linguistique. Elle décrit les récurrences qui sont imposées aux lexèmes dans leur enchaînement syntagmatique. Par exemple, dans la chaîne *la grande montagne*[60], le genre

[60] Exemple extrait de [Rastier & al. 94, p. 120].

féminin est récurrent et cette récurrence forme une isosémie car le genre est marqué dans les morphèmes "-a" (du lexème *la*) et "-e" (du lexème *grande*) et même, selon [Martinet 60, p. 101], dans le lexème *montagne*. Cette récurrence est obligatoire en tant qu'elle est imposée par un accord prescrit par le système fonctionnel de l'enchaînement syntagmatique (ce qui explique l'impossibilité de *le grande montagne*[61]).

L'isosémie peut prendre deux formes en fonction de la contrainte de l'enchaînement syntagmatique qui en est la cause. Il s'agit, soit d'une isosémie d'accord quand la contrainte syntagmatique est morphologique, soit d'une isosémie de rection quand la contrainte syntagmatique est syntaxique.

• Les isosémies d'accord sont marquées au niveau de l'expression des morphèmes (les redondances que l'on a mis en évidence dans la comparaison entre les énoncés au pluriel dans les langues naturelles et les expressions des langages formels sont des isosémies d'accord). Elles sont liées à la concordance du genre (ex : *la grande montagne, le grand bureau*), du nombre (ex : *les oiseaux*) ou de la personne verbale (*vous chantez*). Le trait récurrent dans une isosémie d'accord est inhérent (i.e. non propagé) aux lexèmes impliqués dans l'isosémie.

• Les isosémies de rection sont produites par des dépendances syntaxiques. C'est le cas des récurrences issues de la projection des rôles thématiques d'un verbe sur ses actants et circonstants. Par exemple, un verbe transitif régit syntaxiquement son objet et lui projette le trait /accusatif/ qui de fait devient récurrent (c'est le cas de *il mange une pomme*, où *pomme* prend le trait /accusatif/). Le trait récurrent dans une isosémie de rection est donc inhérent dans le lexème régissant et afférent dans le(s) lexème(s) régit(s). De ce point de vue, l'isosémie de rection est une prescription syntagmatique moins forte que l'isosémie d'accord.

Les isosémies sont obligatoires dans le sens où elle mettent en évidence des traces des contraintes de formation des chaînes linguistiques. Au regard de l'opposition forme *vs*. substance de Hjelmslev ([Hjelmslev 43])[62], ces contraintes s'exercent sur la forme (et non sur la substance) du contenu des constituants de la chaîne. La forme du contenu correspond à la valeur au sens de Saussure. On retrouve la notion d'isosémie dans la grammaire des valeurs pures de [Coursil 92] (qui retravaille les principes saussuriens dans le cadre d'une modélisation en

[61] L'exemple de *la grand mère* ne constitue pas un contre-exemple au critère obligatoire de la récurrence dans l'isosémie car *grand mère* constitue une lexie. Il n'y a donc que deux lexèmes qui s'accordent dans le syntagme *la grand mère* (et pas trois comme dans *la grande montagne*)

[62] La forme d'un contenu est définie par le réseau relationnel où les contenus se définissent mutuellement. C'est la définition de la place du contenu dans son système. La substance est définie par les caractéristiques intrinsèques du contenu qui lui viennent de sa forme. La substance est la manifestation de la forme.

informatique) et, plus précisément, à travers la notion "d'anaphore des catégories sous le signe". Selon Coursil, la forme du contenu est analysable au moyen d'un jeu de valeurs différentielles (*valeurs pures*) qui partitionnent des catégories (par exemple le nombre, le genre ou la personne verbale) qui déterminent le signe. Ces valeurs perdurent le long de la chaîne syntagmatique en s'instanciant dans les catégories portées par les morphèmes. L'instanciation d'une même valeur au sein de plusieurs morphèmes crée une suite de relation d'identité de catégories dans la chaîne syntagmatique, ce que Coursil appelle *anaphore*. Les anaphores du nombre, de genre ou encore de la personne verbale que montre Coursil sont des exemples d'isosémies (isosémies d'accord).

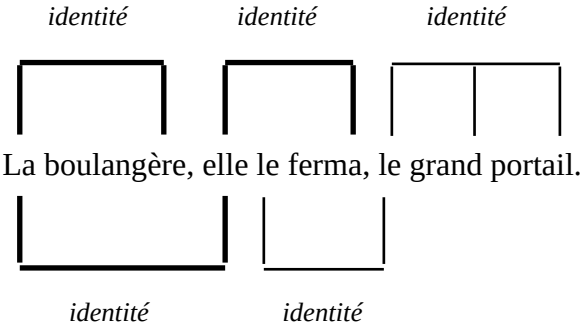


Figure 4.2 : Un exemple de l'anaphore du nombre de Coursil ([Coursil 92, p. 233])

Exemple de la phrase *La boulangère, elle le ferma, le grand portail* où deux "fils d'identité d'instances de catégories" montrent deux anaphores du nombre, c'est-à-dire deux isosémies.

L'isosémie permet de rendre compte de dépendances syntagmatiques. Nous avons vu au chapitre 2 que les dépendances syntagmatiques font la différence entre les dépendances sémantiques et les influences sémantiques. Dans le cas d'une simple influence sémantique, il y aura donc un effet de sens sans isosémie (c'est le cas des lexèmes *pôle* et *glace* dans l'énoncé *La vision est fatiguante au pôle, la glace reflète beaucoup de lumière*). C'est-à-dire que l'influence sémantique révèle une co-adaptation de la substance des contenus des lexèmes mais sans co-adaptation de leurs formes. À l'inverse, dans le cas d'une dépendance sémantique, l'effet de sens constitue une co-adaptation des substances et l'isosémie traduit, en plus, une co-adaptation des formes.

On peut donc considérer l'isosémie comme une trace de phénomènes de dépendance sémantique. D'une certaine façon, l'isosémie rend la dépendance sémantique repérable d'un point de vue opératoire car, là où il y a isosémie, il convient de chercher un effet de sens relevant d'une dépendance sémantique. Dans la suite, nous allons décrire les effets de sens par la notion d'isotopie. Les isosémies sont donc génératrices d'isotopies.

2.2. L'isotopie

L'isosémie est une récurrence de trait qui a rapport à la forme[63] ; l'isotopie (l'isotopie sémantique par différence avec l'isosémie) est, quant à elle, une récurrence de trait qui a rapport à la substance du contenu (c'est en cela qu'elle est sémantique). À l'inverse de l'isosémie, elle n'est pas prescrite par des règles d'enchaînement syntagmatique mais elle résulte entièrement de l'interprétation d'une chaîne linguistique (elle ne lui préexiste pas), et c'est en ce sens qu'elle est facultative par rapport à l'isosémie. Les traits récurrents dans une isotopie sont des sèmes (traits sémantiques) et, à l'inverse des traits morphologiques, ils ne peuvent être marqués dans l'expression des morphèmes (par exemple, il n'y a pas de marque de /animé/ en français).

La sémantique interprétative décrit plusieurs types d'isotopies en fonction de la nature du sème qui est récurrent dans les sémèmes de la chaîne : les isotopies spécifiques, génériques, ou mixtes[64] :

• Isotopie spécifique : comme son nom l'indique, une isotopie est spécifique quand le sème récurrent est un sème spécifique (i.e. un sème qui différentie plusieurs sémèmes au sein de leur taxème). C'est le cas de l'isotopie du trait /inchoatif/ dans le vers d'Éluard *L'aube allume la source.* Le sème /inchoatif/ est inhérent au sémantème (car il est spécifique) des sémèmes 'aube', 'allume' et 'source'.

• Isotopie générique : une isotopie est générique quand le sème récurrent est un sème générique (i.e. un sème qui indexe un sémème dans une classe lexicale). En fonction du degré de généricité du sème, une isotopie générique peut être microgénérique, mésogénérique ou macrogénérique. Elle est microgénérique si le sème en jeu dans l'isotopie est microgénérique, c'est-à-dire qu'il indexe des sémèmes appartenant au même taxème (c'est le cas de l'isotopie du sème /degré de cuisson/ dans les sémèmes 'bleue', 'saignante', 'à point' et 'bien cuite' de l'énoncé *Et l'entrecôte, bleue, saignante, à point ou bien cuite ?*). L'isotopie est mésogénérique quand le sème est mésogénérique, c'est-à-dire qu'il indexe des sémèmes appartenant au même domaine (c'est le cas du sème /navigation/ dans les sémèmes 'amiral', 'carguer', et 'voiles' de l'énoncé *L'amiral Nelson ordonna de carguer les voiles*). Enfin, l'isotopie est macrogénérique quand le sème récurrent est macrogénérique, c'est-à-dire qu'il indexe des sémèmes appartenant à la même dimension (c'est le cas de /animé/ dans les sémèmes 'hérisson' et 'porc-épic' de *Le hérisson insectivore n'est pas de la même famille que le porc-épic*). Les isotopies génériques sont

[63] C'est notamment pour cette raison qu'un défaut d'isosémie ne provoque pas obligatoirement un non sens car le sens est en rapport à la substance du contenu plus qu'à sa forme. Par exemple, sans qu'il y ait isosémie d'accord (de genre) dans l'anaphore suivante, elle reste tout à fait interprétable : Le premier ministre nous a rendu visite, elle est charmante.

[64] Les exemples que nous allons donner pour éclairer les isotopies génériques et spécifiques sont extraits de *Sémantique Interprétative* ([Rastier 87, p. 111-113]).

liées à des paradigmes codifiés en langue. Elle induisent les *impressions référentielles.* Elles sont donc particulièrement déterminante pour la monstration langagière.

• Isotopie mixte : une isotopie est mixte quand elle n'est ni purement générique, ni purement spécifique tout le long de la chaîne, c'est-à-dire quand le sème récurrent est générique dans certains sémèmes de la chaîne et spécifique dans d'autres. C'est le cas de l'isotopie du sème /qui vole/ dans l'énoncé *Le Boeing 747 est un bon avion*, qui est générique dans le sémème 'Boeing 747' et spécifique dans le sémème 'avion'. Dans un tel énoncé exprimant l'appartenance d'un élément à sa catégorie, l'isotopie est forcément mixte car le sémème où le sème récurrent est spécifique ('avion') correspond également au taxème dans lequel est indexé le sémème où le sème récurrent est générique ('Boeing 747').

Parmi ces trois types d'isotopie, on peut encore faire une distinction qui consiste à savoir si l'isotopie est inhérente ou afférente. Dans les exemples que l'on a cités jusqu'ici, les isotopies sont inhérentes car les sèmes récurrents sont inhérents aux sémèmes où ils apparaissent. À l'inverse, l'isotopie est afférente lorsque le sème récurrent est afférent aux sémèmes de la chaîne. C'est notamment le cas lorsqu'il y a une afférence socialement normée dans plusieurs sémèmes de la chaîne, comme l'afférence du sème microgénérique /carrière/ dans les sémèmes 'rouge' et 'noir' du titre du roman de Stendhal *Le rouge et le noir* ([Rastier 87, p. 113]). Lorsque l'afférence est co-textuelle (i.e. n'est pas socialement normée), l'isotopie ne sera ni purement inhérente, ni purement afférente, dans la mesure où le sème récurrent est inhérent dans un sémème et afférent dans un autre, car co-textuellement propagé (c'est le cas des isotopies produites par les récurrences de sèmes lors du rattachement des anaphoriques à leurs antécédents car, par définition de l'anaphore, les sèmes afférents dans le sémème de l'anaphorique sont inhérents dans le sémème de l'antécédent).

L'isotopie (quelque soit son type) permet d'insister sur l'importance d'un sème (celui qui est récurrent) en ce qui concerne la thématique du discours. C'est une trace essentielle du processus d'interprétation qui forme une contrainte forte de l'énoncé sur ses actualisations pragmatiques possibles. De ce fait, l'isotopie montre des aspects du potentiel de sens d'un énoncé et, plus précisément, ses contraintes de monstration (comme en témoigne l'isotopie génériques du sème /navigation à la voile/ dans l'énoncé *Le skipper et son trimaran restaient encalaminés dans le pot-au-noir*[65]).

[65] Dans cet exemple (extrait de [Rastier 91, p. 220]), on a souligné les lexèmes dont les sémèmes portent le sème générique /navigation à la voile/.

Plus qu'un constat *a posteriori*, l'isotopie forme la base du processus d'interprétation. C'est principalement ce qui lui confère son intérêt dans le cadre d'une modélisation d'une compétence interprétative en informatique. En effet, interpréter un énoncé consiste à en extraire les contraintes sémantiques, c'est-à-dire à mettre en évidence une ou plusieurs isotopies. Il y a donc, dans le processus d'interprétation d'un énoncé, un préalable qui consiste en une *présomption d'isotopie* :

> En général, on considère l'isotopie comme une forme remarquable de combinatoire sémique, un effet de la combinaison des sèmes. Ici au contraire, où l'on procède paradoxalement à partir du texte pour aller vers ses éléments, l'isotopie apparaît comme un principe régulateur fondamental. Ce n'est pas la récurrence de sèmes déjà donnés qui constitue l'isotopie, mais à l'inverse la présomption d'isotopie qui permet d'actualiser des sèmes, voire *les* sèmes. ([Rastier 87, p. 11])

L'afférence (qu'elle soit co-textuelle ou socialement normée) doit certainement beaucoup à cette fameuse présomption d'isotopie car, d'une certaine façon, son but est d'installer ou de prolonger une isotopie dans la chaîne. Dès lors, l'apprentissage de mots nouveaux s'explique par cette présomption. De même, l'interprétation de messages énigmatiques fait appel à l'isotopie car l'enjeu dans la résolution de l'énigme est de reconstruire une isotopie qui n'apparaît pas lors d'une première interprétation de la chaîne ([Mauger 98]). La présomption d'isotopie est ce grâce à quoi l'interprétant entre dans le jeu de l'énigme, en cherchant une isotopie cachée. De même, on peut considérer que la paraphrase consiste à formuler par une autre chaîne syntagmatique la ou les mêmes isotopies qu'une première chaîne.

L'isotopie, vue comme la base de l'interprétation, permet de décrire une articulation entre les ordres paradigmatique et syntagmatique. Dans cette articulation, elle est la clé de voûte d'une sémantique unifiée car elle relie les différents paliers d'analyse de la sémantique : la microsémantique, la mésosémantique et la macrosémantique. Au niveau microsémantique, les isotopies s'installent dans les contenus par afférence et les modifient dans leur définition paradigmatique. Au niveau mésosémantique, elle rend compte des effets de sens co-textuels et insiste sur la monstration de l'énoncé (par les influences et les dépendances sémantiques). Enfin, au niveau macrosémantique, elle touche la rhétorique dans l'étude des figures de style qui proviennent de corrélations d'isotopies le long d'un texte (deux isotopies peuvent être parallèles quand elles touchent les mêmes sémèmes, relayées quand dans certains sémèmes l'une prend fin quand l'autre commence, successives quand l'une commence une fois l'autre terminée, ou encore alternées quand l'une et l'autre se prolongent dans le texte sans jamais affecter les mêmes sémèmes [Tanguy 97, p. 80]).

3. Un modèle calculatoire des effets de sens guidé par l'isotopie

Dans les parties précédentes, en exposant rapidement les principes de base de la sémantique interprétative de Rastier, nous en avons montré l'intérêt. Nous allons maintenant expliquer comment exploiter ces principes linguistiques en relation avec le modèle sémantique de l'axe paradigmatique (cf. chapitre n°3). Plus précisément, nous allons donner les bases d'un modèle informatique de l'interprétation des effets de sens guidé par la notion d'isotopie.

Comme on l'a signalé, l'isotopie est définie par la récurrence d'un même sème dans les lexèmes de la chaîne. Mais les sèmes que l'isotopie rend récurrents sont vus ici comme des sèmes oppositionnels (dans le sens où ils résultent d'un jeu d'oppositions qui partitionnent un domaine d'interprétation). De même qu'en sémantique différentielle, les sèmes oppositionnels structurent le sémème en termes de sémantème (sèmes spécifiques du sémème) et de classème (sèmes génériques du sémème) ; on retrouve ici les trois types d'isotopies : générique, spécifique et mixte.

En analysant la première demi-heure du corpus PIC, où il est question des formes d'aide à l'utilisation d'un logiciel, nous avons exhibé les sémèmes 'manuel', 'référentiel', 'aide en ligne', 'assistant', 'sommaire', 'full text' et 'index' par les sèmes [Présentation de l'information : globale *vs*. précise], [Support de l'information : papier *vs*. logiciel], [Accès à l'information : interactif *vs*. figé] et [Structure de l'information : hiérarchique *vs*. linéaire] (cf. Chapitre 3, § 2.3.2.). Ces sèmes sont récurrents dans de nombreux énoncés de cette demi-heure de conversation et forment des isotopies, tantôt spécifiques, tantôt génériques, ou tantôt mixtes. Par exemple, dans l'énoncé :

D :
U :
R : vous pouvez avoir comme vous dites des informations très précises du genre un **référentiel**, où plutôt un **manuel** utilisateur

Figure 4.3 : Un extrait du corpus (21ème minute)

apparaissent les lexèmes *référentiel* et *manuel* qui contiennent dans leur sémantème les deux sèmes S1 = [Présentation de l'information : globale *vs*. précise] et S2 = [Support de l'information : papier *vs*. logiciel]. On a donc ici deux isotopies spécifiques.

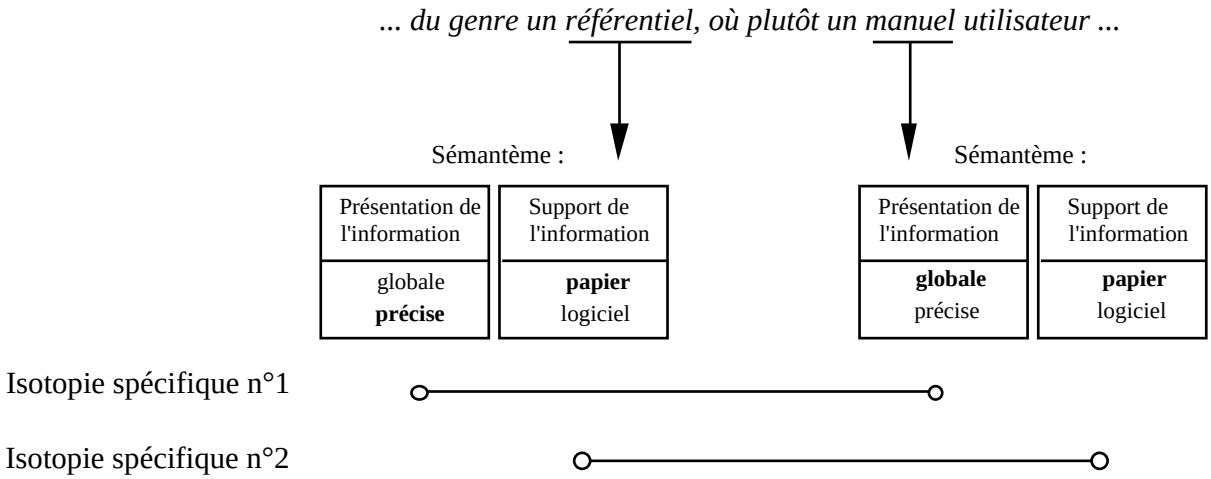


Figure 4.4 : Deux isotopies spécifiques (parallèles)

De même, dans l'énoncé

> *... vous connaissez les outils dans les **aides en ligne**, les **assistants** …* (énoncé du rédacteur à l'intention de l'utilisateur, à la 20ème minute)

on trouve les lexèmes *aides en ligne* et *assistants*. Le sémème 'aide en ligne' contient les sèmes S1 et S2 dans sa partie spécifique et le sémème 'assistant' contient S1 et S2 dans sa partie générique, ce qui donne lieu à deux isotopies mixtes.

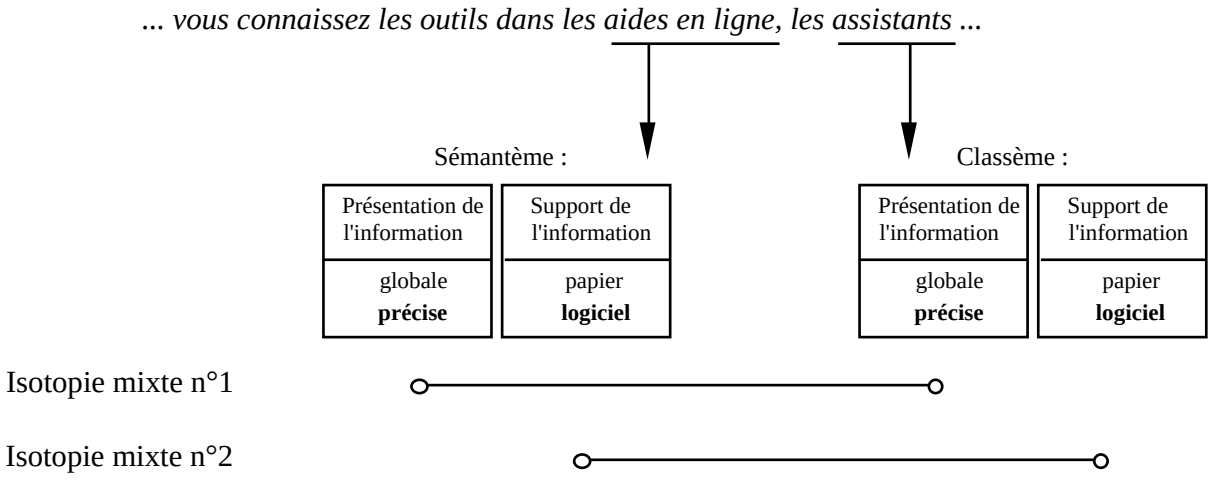


Figure 4.5 : Deux isotopies mixtes (parallèles)

De même, si l'on considère que les sémèmes 'manuel', 'référentiel', 'aide en ligne', 'assistant', 'sommaire', 'full text' et 'index' contiennent à un niveau supérieur de généricité le sème [Produit informatique : documentation *vs*. application] actualisé dans la valeur documentation, alors ce sème est récurrent dans toute cette partie du corpus (il l'est même tout le long des deux

heures de dialogue étant donné que la tâche des trois participants est de réfléchir à la conception de la documentation technique d'un logiciel). Ceci forme une isotopie générique.

Ces isotopies marquent ce sur quoi les conversants insistent particulièrement dans les énoncés qu'ils produisent. Elles rendent compte d'effets de sens qui constituent de fortes contraintes de monstration des énoncés que nous rapportons à leur potentiel de sens. Par exemple, pour l'énoncé précédent (... *vous connaissez les outils dans les aides en ligne, les assistants* ...), les deux isotopies nous permettent de formuler les deux contraintes suivantes :

• Il est question de la présentation de l'information et celle-ci est précise (en opposition à une présentation globale)

• Il est question du support de l'information et celui-ci est logiciel (en opposition à un support papier)

Plus une isotopie est dense (i.e. plus le sème récurrent est présent dans les différents sémèmes de la chaîne) et plus elle s'étend dans la suite des énoncés, plus la contrainte qu'elle produit est forte (et plus cette contrainte rend compte de l'effectivité de l'interaction). De ce point de vue, les isotopies génériques permettent de mettre en évidence des contraintes plus fortes que les isotopies spécifiques car les différents sémèmes qui partagent les mêmes sèmes génériques les rendent plus souvent récurrents, alors que des sèmes spécifiques sont moins partagés.

Comme nous l'avons déjà signalé précédemment, l'isotopie n'est pas seulement un constat *a posteriori* de l'interprétation, elle en est un principe régulateur. C'est-à-dire, qu'en plus de montrer des effets de sens, elle les met en place. C'est grâce à l'isotopie que l'on va pouvoir réduire la polysémie lexicale des constituants de la chaîne pour retenir les sémèmes contextuellement pertinents. Pour cela, on formule la règle calculatoire suivante :

Règle de réduction de la polysémie lexicale : Pour un morphème ou un lexème d'une chaîne, son sémème en contexte est celui qui renforce le plus d'isotopies.

Par exemple, dans le dialogue suivant entre le rédacteur et l'utilisatrice, il apparaît une polysémie du mot *support.*

D :
U : *bah* non dans la
R : est-ce que vous appelleriez quand même *heu* le support parce que c'est une méthode que vous aimez bien

D :
U : mesure où je me débrouille toute seule avec mon support papier
R :

Figure 4.6 : Un extrait du corpus (12ème minute)

Pour le rédacteur, le support est un technicien (appartenant à la société qui commercialise le logiciel) que l'on peut contacter par téléphone lorsque l'on a des problèmes d'utilisation d'un logiciel (il l'explique à l'utilisatrice dans la minute précédente du dialogue). Cette prise de position du rédacteur est marquée dans le co-texte par la dépendance sémantique due à la relation actancielle objet entre le verbe *appeler* et le syntagme *le support*. À l'inverse, dans la réponse de l'utilisatrice, support à un sens d'objet matériel ayant une signification métonymique de documentation. On peut donc considérer que le mot *support* est polysémique (au moins dans ce contexte) et l'on va définir son contenu sémique par deux sémèmes différents représentant chacun un des sens possibles du mot[66]. Ainsi, dans la description componentielle du premier sens du mot *support* en tant que technicien, on pourra avoir le sème spécifique suivant [Agent commercial : technicien de maintenance *vs*. vendeur] actualisé dans la valeur technicien de maintenance. Dans son deuxième sens, en terme de documentation, on aura le sème spécifique [Support de l'information : papier *vs*. logiciel] actualisé dans la valeur papier. Dans la situation de notre corpus, on peut également affecter au sémème du lexème *papier* le même sème [Support de l'information : papier *vs*. logiciel] (évidemment actualisé dans la valeur papier). Dans la réponse de l'utilisatrice, ce sème est alors récurrent dans les descriptions componentielles des lexèmes *support* et *papier* du syntagme *mon support papier*. Par application de la règle de réduction de la polysémie lexicale, l'isotopie formée par cette récurrence, permet de désambiguïser le mot *support* en considérant uniquement, dans cet énoncé, le sémème qui crée cette isotopie. Cette levée de l'ambiguïté du mot *support* par l'isotopie permet de rendre compte d'une dépendance sémantique due à la constitution du groupe nominal *mon support papier*.

[66] Ceci ne revient pas à adopter une approche sémasiologique, à laquelle nous préférons une approche onomasiologique, car les deux sémèmes ne sont pas définis l'un par rapport à l'autre par le fait qu'ils aient une même expression mais différentiellement au sein de taxèmes distincts.

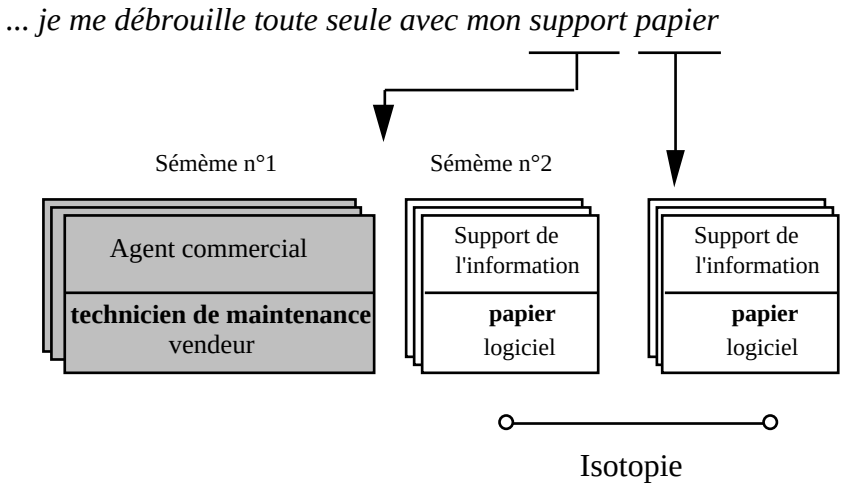


Figure 4.7 : Réduction en contexte de la polysémie lexicale par l'isotopie

Le sémème qui est grisé (sémème n°1) n'est pas retenu du fait de l'application de la règle car, contrairement au sémème n°2, il n'entre ici en compte dans aucune isotopie.

Comme l'ont montré [Victorri & al. 96], la polysémie est très fréquente (cf. Chapitre 2), et on la retrouve massivement dans notre corpus de dialogue. Ceci explique que l'on puisse souvent appliquer la règle que l'on vient de formuler car elle permet de rendre compte de la monstration des énoncés en contexte. C'est-à-dire que son application donne des indices sur les négociations et les incompréhensions entre les différents partenaires dans l'établissement du terrain commun de l'interaction.

Examinons, par exemple, une séquence du corpus où apparaît une incompréhension entre les interlocuteurs, qui est directement due à la polysémie d'un mot. Cette séquence est située à 1 heure et 9 minutes du début de l'interaction, lorsque les partenaires essayent le logiciel en manipulant son interface. Dans cette séquence, les trois conversants se posent le problème de remplir un champ de l'interface graphique du logiciel et ce champ est présenté par l'identifiant *poste à l'IUT*.

D :	non poste à l'IUT, c'est le poste téléphonique
U :	poste à l'IUT, *ben* on va mettre vacataire donc *ah* le poste téléphonique
R :	téléphonique

Figure 4.8 : Un extrait du corpus (1h. 09 min.)

Dans la premier énoncé de U (l'utilisatrice) *"poste à l'IUT, ben on va mettre vacataire"*, le mot *poste* est à interpréter comme un statut professionnel. On peut considérer que son signifié contient un sémème ayant comme sème spécifique le sème [Activité professionnelle : statut *vs*. fonction] actualisé dans la valeur statut (en opposition, le sémème 'profil de poste' actualiserait la valeur fonction). De même, dans cet énoncé, *vacataire* est à interpréter comme un poste

possible (en opposition à *titulaire*). C'est-à-dire que 'vacataire' est un sémème extrait de la sous-catégorisation de 'poste' (de même pour 'titulaire'). En d'autres termes, les sèmes de 'poste' forment le classème de 'vacataire' qui contient dans son sémantème le sème spécifique [Durée du contrat : indéterminée *vs*. déterminée] opposant 'vacataire' (actualisant la valeur déterminée) à 'titulaire' (actualisant la valeur indéterminée). On a donc, dans cet énoncé, une isotopie mixte du sème [Activité professionnelle : statut *vs*. fonction] ; il est générique dans 'vacataire' et spécifique dans 'poste'. Cette isotopie marque une influence sémantique entre les mots *poste* et *vacataire* étant donné que leur co-présence (sans dépendance syntagmatique), dans le même énoncé, permet de préciser leurs contenus sémiques.

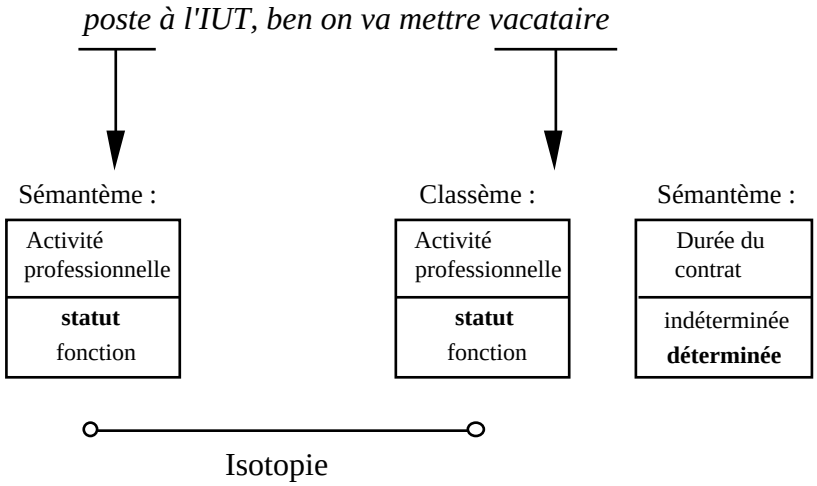


Figure 4.9 : Une influence sémantique

Il n'y a pas jusqu'ici de polysémie flagrante. Cependant, l'énoncé qui suit dans le dialogue, *non poste à l'IUT, c'est le poste téléphonique*, marque une prise de position de la part de D (la développeuse) qui signale à ses deux partenaires une incompréhension sur le sens de *poste à l'IUT*. Dans cette prise de position, D va actualiser un nouveau sens du mot *poste* en tant qu'appareil servant à téléphoner. Ce nouveau sens fait de *poste*, à ce moment de l'interaction, un mot polysémique ayant produit une incompréhension. Pour représenter la nouvelle signification de *poste*, on va former un nouveau sémème. Ce sémème est issu d'une sous-catégorisation de 'téléphone' et contient dans sa partie générique le sème [Communication à distance : téléphone *vs*. courrier] actualisé dans la valeur téléphone. Dans sa partie spécifique, on peut actualiser la valeur appareil du sème [Matériel téléphonique : appareil *vs*. ligne] ; ce sème permettant également de définir le sémème 'réseau' (en actualisant la valeur ligne). L'adjectif *téléphonique*, quant à lui, reprend dans sa substance le sémème 'téléphone' (dans sa forme, il précise qu'il est qualificatif ce qui peut donner lieu à des isosémies mais pas à des isotopies). C'est-à-dire qu'il contient le sème spécifique [Communication à distance : téléphone *vs*. courrier] actualisé dans la valeur téléphone. Il apparaît donc une isotopie (mixte) dans le groupe nominal *le poste téléphonique* traduisant une dépendance sémantique (i.e. une influence sémantique en plus d'une dépendance syntagmatique) entre les mots *poste* et *téléphonique*.

L'application de la règle de réduction de la polysémie lexicale permet de résoudre cette dépendance sémantique en ne sélectionnant pas ici le sémème de *poste* en terme de statut professionnel, car il ne valide pas d'isotopie dans le groupe nominal (cf. Figure 4.10). Dans un second temps, ceci implique de considérer également pour la première occurrence du mot *poste* dans l'énoncé, le sémème décrivant la signification en terme d'appareil téléphonique. En effet, les deux occurrences de *poste* produisent des isotopies par répétition dans l'énoncé de leur contenu sémique. Cependant, les sèmes du sémème en termes de statut professionnel ne sont pas récurrents étant donné que, dans la seconde occurrence de *poste*, ils n'ont pas été retenus par la dépendance sémantique. Ceci traduit une influence sémantique entre les deux occurrences du mot *poste* qui renforce l'isotopie du sème [Communication à distance : téléphone *vs*. courrier] et crée une isotopie du sème [Matériel téléphonique : appareil *vs*. ligne] (cf. Figure 4.11). Ces deux isotopies entrent en contradiction avec celle initiée par l'énoncé de U (i.e. l'isotopie du sème [Activité professionnelle : statut *vs*. fonction]). Ceci montre la prise de position de D par rapport à U et rend compte d'un malentendu évité dans l'interaction (c'est le signe d'une interaction réussie). La suite du dialogue marque bien un consensus final entre les inter-actants sur le sens de *poste à l'IUT* car les deux énoncés suivants dans la séquence renforcent les isotopies initiées par l'énoncé de D : *téléphonique* (énoncé par le rédacteur) et *ah le poste téléphonique* (énoncé par l'utilisatrice).

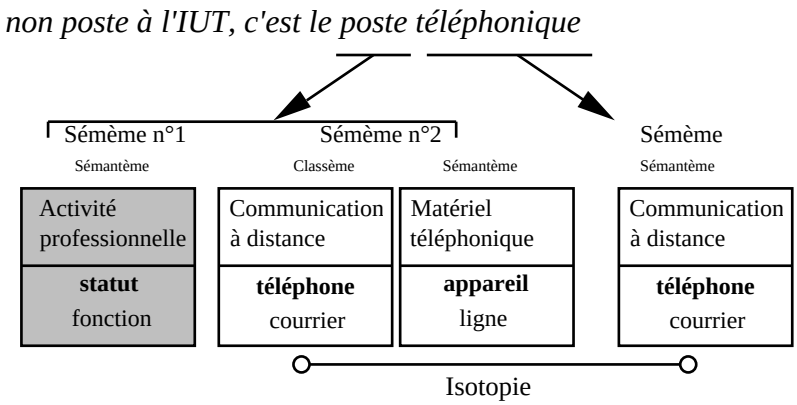


Figure 4.10 : Premier temps : la dépendance sémantique *poste - téléphonique*

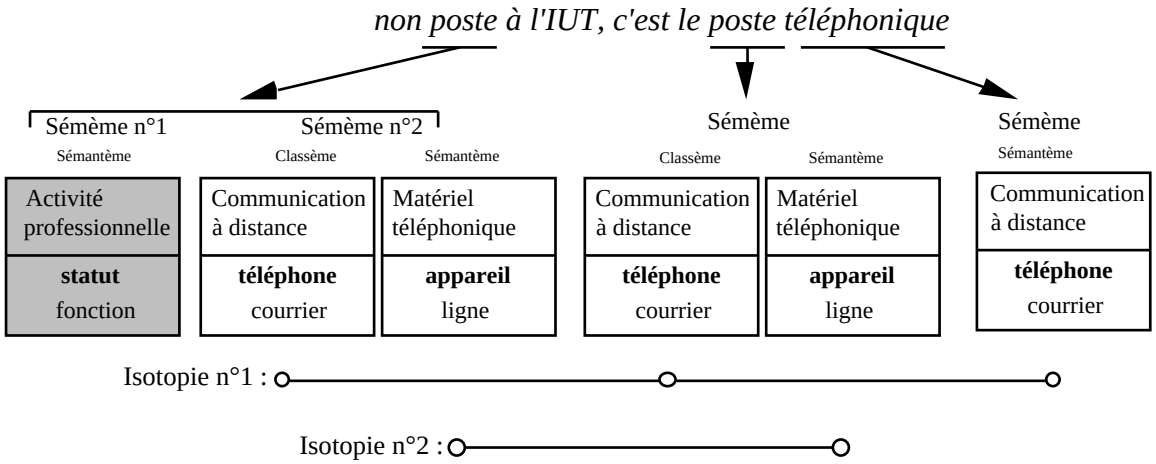


Figure 4.11 :Second temps : l'influence sémantique *poste - poste.*

L'effet de sens est propagé de la seconde à la première occurrence de *poste*, c'est-à-dire qu'il est propagé en sens inverse de la chaîne.

Comme on vient de le voir sur des exemples de notre corpus, on modélise, grâce à l'isotopie, un calcul des effets de sens co-textuels dus aux co-adaptations des lexèmes. Ces effets de sens consistent à renforcer des propriétés communes et permettent de desambiguïser les mots polysémiques dans les énoncés où ils apparaissent, mais il ne s'agit pas à proprement parler de modéliser l'afférence, dont on a signalé l'importance dans la première partie de ce chapitre. Par rapport à un problème de sélection de contenus sémiques en contexte, le problème de l'afférence se pose différemment car il n'est pas, en général, essentiellement syntagmatique (sauf dans des cas d'afférence co-textuelle due au rattachement d'un pronom anaphorique). La question de l'afférence articule les ordres syntagmatiques et paradigmatiques. En quelque sorte, nous rendons compte ici de la contextualisation des contenus alors que l'afférence relève de la construction des contenus par contextualisation. Dans la perspective d'un modèle interactif où la machine tient le rôle d'un apprenant, l'afférence sera à la charge du partenaire humain à travers la description des contenus sémiques en fonction de la situation.

Nous allons maintenant expliquer ce qu'un modèle oppositionnel du sème apporte en plus d'un modèle classique (en tant que simple chaîne de caractères) dans la recherche syntagmatique d'effets de sens. Comme c'est le cas dans la sémantique interprétative, l'isotopie sera la trace de la cohésion textuelle d'un énoncé (ou d'un discours), mais nous allons voir également qu'un modèle non atomiste du sème permet d'interpréter l'isotopie, non plus simplement en terme de cohésion, mais aussi en terme de cohérence interne de l'énoncé. Ce premier niveau de la cohérence de l'énoncé est le fait des actualisations relatives des sèmes le long d'une isotopie.

3.1. La cohésion d'un énoncé

Ce n'est pas parce qu'un énoncé est syntaxiquement correct qu'il est sémantiquement correct. Cette différence d'acceptabilité au niveau des plans syntaxiques et sémantiques permet de mettre au jour la cohésion d'un énoncé. La cohésion est ce qui fait qu'un énoncé bien formé syntaxiquement est en plus, intelligible (ou encore non absurde). C'est la différence entre les énoncés (1) et (2) suivants, où (1) a une cohésion alors que (2) n'en a pas :

(1) *Ma voiture est garée sur le parking.*

(2) *Le silence vertébral indispose le voile licite* (Tesnière)

L'énoncé (2), que l'on doit à Tesnière, est produit par un procédé de codage assez simple qui consiste, à partir d'un énoncé "normal", à remplacer chaque lexème par un de ceux qui le suivent dans le dictionnaire en respectant les accords et les rections. Ainsi, (2) est formé à partir de l'énoncé *Le signal vert indique la voie libre*. Dans ce procédé, en respectant les accords et les rections, on conserve les isosémies (i.e. la forme de l'énoncé). Cependant, on supprime complètement les isotopies (i.e. la substance de l'énoncé) qui étaient instituées dans l'énoncé d'origine. Par là, on montre que la cohésion d'un énoncé provient des isotopies (comme en témoigne l'isotopie générique de /automobile/ dans (1)). Lorsque l'énoncé ne présente pas d'isotopies, il est dénué de cohésion. Plus précisément, c'est le défaut d'isotopies génériques qui produit le manque de cohésion car l'énoncé ne porte pas d'impressions référentielles.

> L'absurdité d'un énoncé syntaxiquement bien formé (recevable pour ce qui concerne la forme du contenu) est un effet de l'absence d'isotopie générique : l'énoncé est alors irrecevable en ce qui concerne la substance du contenu. ([Rastier 87, p. 156])

En fait, cela revient au même de dire que, quand un énoncé ne porte pas d'isotopies génériques, il ne porte pas d'isotopies. En effet, s'il porte une isotopie spécifique, c'est qu'un sème spécifique est récurrent dans au moins deux de ses lexèmes. C'est-à-dire que ces lexèmes font partie du même taxème. Les sèmes génériques qui définissent le taxème sont donc également récurrents. Quand il y a isotopie spécifique, il y a donc isotopie générique (de même et pour des raisons identiques quand il y a une isotopie mixte). Le défaut de cohésion lié à une absence d'isotopie générique s'explique donc par une absence d'isotopie (qu'elle soit générique, spécifique ou mixte).

Le manque de cohésion d'un énoncé n'est pas synonyme d'une interprétation impossible, bien que l'interprétabilité soit (comme on l'a vu précédemment) liée à la mise en forme d'isotopies dans l'énoncé. En effet, s'il n'y a pas d'isotopie produite par les contenus inhérents des lexèmes de la chaîne, il peut y avoir des isotopies produites par des contenus afférents que l'on peut attribuer aux lexèmes. Attribuer ces contenus consiste précisément en une

interprétation d'un énoncé qui apparaissait initialement obscur. Ainsi, on trouve dans [Bassi Acuña 95, p. 122] une interprétation de l'énoncé sans cohésion apparente *D'incolores idées vertes dorment furieusement*[67] qui conduit à former la paraphrase *Un vague espoir s'agite dans le subconscient* (incolore est ici interprété comme mal défini ou vague, vert comme espoir suivant la maxime selon laquelle le vert est la couleur de l'espérance, dormir comme faisant référence au subconscient, et furieusement comme ayant rapport à une agitation). Une telle interprétation d'un énoncé obscur met en jeu des parcours interprétatifs complexes qui nécessitent plusieurs afférences dans les contenus des lexèmes. Ces nombreuses afférences traduisent des constructions tropiques[68] de sémèmes en contexte. Elles sont fréquentes dans les interprétations de textes poétiques.

Interpréter consiste à relever les isotopies d'un énoncé, c'est donc mettre en évidence la cohésion de l'énoncé, même si celle-ci est volontairement dissimulée (l'interprétation consiste alors à la dévoiler). Il ne faut alors pas considérer la cohésion comme un préalable à une possible interprétation, mais comme son résultat autant que son moyen. Ainsi, rendre compte d'un effet de sens dans un énoncé, c'est prouver sa cohésion textuelle car c'est mettre en évidence une ou plusieurs isotopies. Il y a donc cohésion, quand il y a possibilité de former une contrainte sémantique de monstration précisant une opposition dans un domaine, contrainte issue de l'isotopie d'un sème oppositionnel. C'était le cas des exemples donnés dans la partie précédente.

3.2. La cohérence interne

Si la sémantique interprétative définit la cohésion d'un énoncé par ses relations sémantiques internes, elle définit également une autre forme d'unité d'une séquence linguistique : sa cohérence. Dans le cadre de la sémantique interprétative, la cohérence d'un énoncé est définie par ses relations avec son *entour* ([Rastier 87, p. 273]). En fonction de ce que l'on considère comme *entour*, on décrit la cohérence à différents plans d'analyse. Ainsi, à un niveau d'analyse linguistique textuelle, la cohérence d'un énoncé est déterminée par les relations qu'entretiennent cet énoncé et son co-texte (par exemple en terme de compatibilité thématique). À un niveau d'analyse pragmatique, la cohérence est déterminée par les relations entre l'énoncé et son contexte non linguistique, analysables en terme de compatibilités entre des connaissances linguistiques convoquées par l'énoncé (plus précisément par son interprétation) et des connaissances sur le monde et la situation d'énonciation ([Bassi Acuña 95, p. 79]). On peut donc dégager deux formes de cohérence, sans que l'une exclue l'autre : une cohérence

[67] Cet énoncé est une traduction française de l'énoncé *Colourless green ideas sleep furiously* construit par Chomsky comme un exemple d'énoncé logiquement indéterminable.
[68] Un trope est une figure de rhétorique où il apparaît immédiatement que le sens littéral n'est pas le sens intentionné.

pragmatique (qui concerne l'actualisation dans un contexte du sens d'un énoncé) et une cohérence sémantique (qui concerne le sens "potentiel" d'une suite d'énoncés).

Plutôt que de considérer exclusivement la cohérence sémantique comme un rapport entre plusieurs énoncés, nous allons définir un premier niveau de cohérence dans les relations sémantiques internes de l'énoncé et nous en expliquerons la différence avec la cohésion. Nous appelons ce premier niveau de la cohérence sémantique, la cohérence interne.

La cohérence interne est ce grâce à quoi le potentiel de sens d'un énoncé est formé de contraintes compatibles, dans le sens où elle n'entrent pas en contradiction les unes envers les autres. C'est ce qui différencie les énoncés "normaux" des énoncés qui, bien qu'induisant une ou plusieurs impressions référentielles, demeurent étranges à l'interprétation. Considérons par exemple, l'énoncé suivant :

Le train disparu, la gare part en riant à la recherche du voyageur.

[Rastier & al. 94, p. 127] (dont cet exemple est extrait) analysent cet énoncé comme étant porteur d'une isotopie renvoyant au domaine des transports et présentant une rupture d'isosémie dans la relation sujet-verbe *la gare part en riant* car le sujet est non animé alors que le procès renvoie à une activité d'un sujet animé (et même humain du fait de *en riant*). De notre point de vue, il n'y a pas ici de rupture d'isosémie car l'énoncé est syntaxiquement correct et les sèmes /animé/ et /non animé/ ne sont pas liés à la forme mais à la substance du contenu (à l'inverse du trait /accusatif/ par exemple). Ceci conduit à proposer une autre analyse de l'énoncé, en termes de défaut de cohérence interne. Il n'y a pas, dans cet énoncé, de défaut de cohésion textuelle car on peut mettre en évidence au moins deux isotopies :

- celle qui renvoie au domaine des transports, que l'on montre par la récurrence du sème [Transports : collectifs *vs*. individuels] dans les sémèmes 'train', 'gare' et 'voyageur' et

- celle due à la récurrence du sème [Actants : animé *vs*. non animé] dans les sémèmes 'gare', 'part' et 'riant' (cf. Figure 4.12).

La première isotopie permet de formuler la contrainte exprimant qu'il est question de transports collectifs (en opposition à des transports individuels). La seconde isotopie exprime deux contraintes incompatibles : il est question d'actants animés (en opposition à non animé) et il est question d'actants non animés (en opposition à animé). Ceci manifeste un défaut de la cohérence interne de l'énoncé.

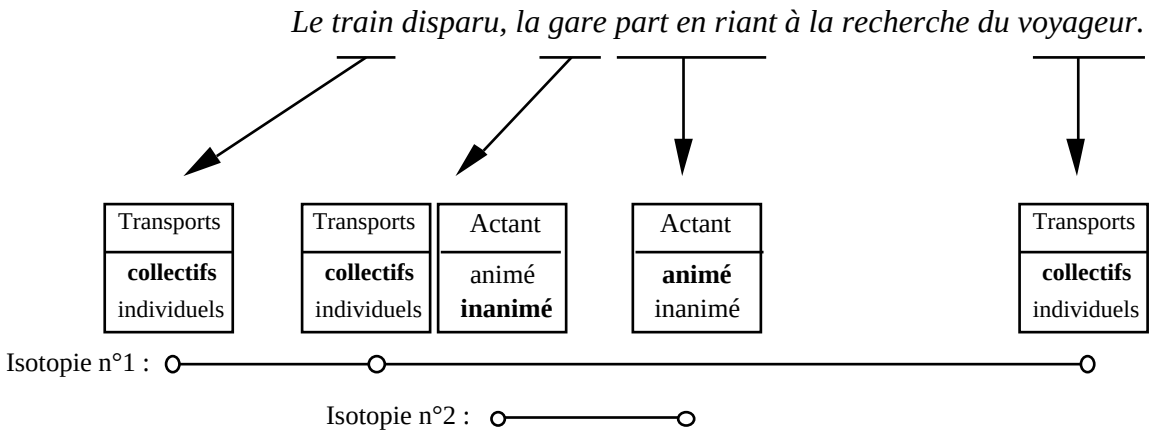


Figure 4.12 : Un énoncé sans cohérence interne

La distinction entre la cohésion textuelle et la cohérence interne apparaît avec le modèle non atomiste du sème que nous avons proposé. À la différence de la cohésion qui est uniquement induite par les isotopies de l'énoncé, la cohérence interne provient de l'actualisation du sème récurrent dans les différents lexèmes que "touche" l'isotopie. Cela implique que l'on ne puisse décider de la cohérence (ou de l'incohérence) d'un énoncé que si celui-ci est porteur d'isotopies. En d'autres termes, il n'y a cohérence interne que si, au préalable, il y a cohésion. La cohésion est induite par la récurrence de la partie "constante" d'un sème, c'est-à-dire son domaine d'interprétation (vu comme un descripteur de taxème), et la cohérence par la compatibilité des valeurs actualisées dans les sèmes récurrents le long de l'isotopie.

Cette incohérence, due à l'incompatibilité des valeurs actualisées, n'est pas issue de la construction même de l'isotopie. En effet, on a montré précédemment (cf. Figure 4.4, isotopie spécifique n°1) une isotopie où le sème récurrent est actualisé de façon différente dans deux sémèmes reliés par l'isotopie sans que cela soit la trace d'une incohérence. Cette compatibilité est prescrite par les relations syntagmatiques qu'entretiennent les lexèmes reliés par l'isotopie. C'est en cela qu'une dépendance syntagmatique est génératrice d'une dépendance sémantique. Dans l'exemple précédent, le défaut de cohérence provient de la différence d'actualisation du sème dans l'isotopie alors que la relation actancielle sujet qui porte cette isotopie "demande" à ce que les actualisations des sèmes soient compatibles. L'isotopie qui ne respecte pas les prescriptions de la dépendance syntagmatique qui la porte réalise alors une disjonction exclusive entre deux sémèmes. Ce genre d'isotopie est appelée *allotopie*[69].

[69] Au regard de la rhétorique, l'allotopie forme une figure de style particulière (un trope) où, justement, le but est de montrer une incompatibilité de sémèmes pour orienter l'interprétation de façon métaphorique (c'est l'exemple de *Un cyclone de tendresse* où l'allotopie entre les sémèmes 'cyclone' et 'tendresse' conduit à un usage métaphorique de *cyclone*)

Il convient donc de faire un inventaire des prescriptions syntagmatiques qui contraignent la formation des isotopies dans les dépendances sémantiques. Pour cela, nous proposons deux familles de dépendances sémantiques :

1) Les dépendances sémantiques construites sur des relations syntagmatiques qui prescrivent une compatibilité des actualisations de sèmes dans l'isotopie. Cette famille regroupe :

- les relations actancielles sujet (comme on l'a vu dans l'exemple précédent),
- les relations actancielles objet (ex : dans *Je lis Saussure*, l'interprétation métonymique de l'objet consiste précisément à mettre en forme une isotopie dont les actualisations sont compatibles),
- les relations de constitution des groupes nominaux (ex : l'interprétation du syntagme *Le chat immortel* conduit à considérer un monde contre-factuel (ou des référents artéfactuels) où, justement, les contenus sémiques du nom et de l'adjectif ne soient plus contradictoires),
- les relations de détermination (ex : l'énoncé *La girafe s'est échappée de la réserve où elle vivait en liberté*[70] montre un défaut de cohérence interne car la proposition relative demande une compatibilité entre le lexème qu'elle détermine et les lexèmes qui la composent, alors que dans le cas présent on a l'isotopie du sème [Liberté : enfermé *vs*. libre] dans les sémèmes 'réserve' et 'liberté', mais 'réserve' l'actualise dans la valeur enfermé et 'liberté' dans la valeur libre, d'où l'allotopie[71]),
- les anaphores (car l'anaphorique, sémantiquement non autonome, reprend par afférence les sèmes de son antécédent ; les actualisations de sèmes sont donc les mêmes).

2) Les dépendances sémantiques construites sur des relations syntagmatiques qui prescrivent une différence d'actualisation de sèmes dans les isotopies spécifiques[72]. Cette famille regroupe :

- les relations de coordination (on retrouve ici le phénomène de dissimilation car la coordination prescrit une isotopie spécifique avec une différence d'actualisation, ce qui explique que l'énoncé *Il y a musique et musique* ne soit cohérent que si l'on considère deux significations de *musique* différentiées spécifiquement l'une de l'autre),

[70] Cet exemple est extrait de [Coursil 97] où il est décrit comme un récit logiquement incohérent.
[71] On ne retrouve effectivement pas d'incohérences dans les variantes suivantes : *La girafe sort du jardin où elle était en liberté* (ici le sème est actualisé dans la valeur libre tout le long de l'isotopie), *La girafe sort de la réserve où elle était détenue* (ici le sème est actualisé dans la valeur enfermé tout le long de l'isotopie) et *La girafe sort de la réserve où elle n'était pas en liberté* (ici la détermination de la proposition relative inclue une prédication qui rend compatible l'actualisation de la valeur enfermé dans 'réserve' avec la négation de l'actualisation de la valeur libre dans 'liberté').
[72] En plus bien sûr, d'une isotopie générique d'où provient la cohésion.

- les relations de disjonction (il ne peut effectivement pas y avoir disjonction entre contenus sémiques n'ayant rien en commun, comme il ne peut y avoir de disjonctions entre contenus sémiques identiques),
- les relations d'énumération (pour les mêmes raisons liées aux dissimilations que la coordination et la disjonction),

Comme on l'a précédemment signalé pour le manque de cohésion, un défaut de cohérence interne n'est pas cause de non interprétabilité d'un énoncé. Au contraire, cela peut être un indice pour mettre en place une stratégie particulière d'interprétation. C'est, par exemple, le cas des énoncés métaphoriques (de même pour les énoncés métonymiques), où une interprétation littérale conduit à un défaut de cohérence interne, signe pour l'interprétant de la nécessité d'une interprétation ayant recours à un trope.

4. Conclusions

Nous avons proposé, dans ce chapitre, de modéliser la recherche des effets de sens par une analyse de la récurrence des sèmes dans le co-texte. Cependant, à la différence de la sémantique interprétative, les sèmes ne sont pas ici considérés comme atomiques. Cela permet de mettre en évidence une cohérence (ou une incohérence) interne de l'énoncé. La notion de cohérence interne permet d'exploiter conjointement l'apport des dépendances syntagmatiques et l'apport des isotopies dans l'analyse de la variabilité co-textuelle des significations. Ceci montre qu'il n'y a pas, dans l'enchaînement syntagmatique, de distinction franche entre le niveau syntaxique et le niveau sémantique. Dans une perspective informatique, la prise en compte de la cohérence interne des énoncés se situe donc à l'interface entre un module d'analyse syntaxique et un module de recherche de contraintes sémantiques (cela montre que les deux modules en cause ne sont pas indépendants).

Étudier la notion de cohérence interne conduit à étendre le champ de l'analyse sémantique. Si nous nous sommes limités aux effets de sens de monstration, nous sommes conscients de n'aborder qu'un aspect de la sémantique. Il conviendrait également de faire une étude des connecteurs argumentatifs cas par cas (dits opérateurs chez [Ducrot & al. 95, p. 449]) car ils sont porteurs de prescriptions pour l'actualisation des sèmes dans les isotopies. Ils conditionnent donc directement la cohérence interne. De plus, quand ils prescrivent une différence d'actualisation, ils permettent de retenir l'actualisation principale dans l'isotopie. C'est-à-dire que l'argumentation éclaire aussi la monstration de l'énoncé. Considérons, par exemple, le connecteur argumentatif *mais* dans les deux énoncés suivants dont le contenu prédicatif est identique :

(a) *Jean est intelligent mais brouillon.*

(b) *Jean est brouillon mais intelligent.*

Dans le cadre d'une sémantique de l'argumentation ([Raccah 97]), les contenus de *intelligent* et *brouillon* sont représentés par des *topoï,* c'est-à-dire des champs graduels de la forme //antécédent, conséquent//. Ainsi, *intelligent* porte le *topos* //plus il est intelligent, mieux c'est// et *brouillon* le *topos* //plus il est brouillon, moins bien c'est//. La sémantique du connecteur *mais* est expliquée de la façon suivante :

Dans "P mais Q",
(1) le conséquent de P est opposé au conséquent de Q, et
(2) le topos retenu pour l'énoncé est celui de Q.

La proposition (1) consiste à imposer que les conséquents de P et de Q doivent être opposables, c'est-à-dire qu'ils n'ont pas la même valeur au regard d'un domaine d'interprétation où l'on puisse les comparer. C'est donc signaler qu'il y a une isotopie entre P et Q mais que, dans cette isotopie, P et Q actualisent différemment le sème récurrent. La contrainte sémantique de *mais* est en cela similaire aux contraintes induites par les coordinations, les disjonctions, les énumérations et les anaphores nominales. De plus, (2) indique l'actualisation qui fait l'objet de la monstration de l'énoncé ; ce qui explique que l'énoncé (a) insiste sur les défauts de *Jean* alors que l'énoncé (b) insiste au contraire sur ses qualités.

De même, cette notion de cohérence permet de mettre en évidence des phénomènes propres aux dialogues. Dans le déroulement normal d'un dialogue, où il est question d'un terrain commun, l'interprétant attribue aux énoncés de son interlocuteur une cohésion et une cohérence par défaut. L'absence de l'une ou de l'autre va affecter le déroulement de l'interaction. Si l'absence de cohésion et/ou de cohérence est évidente et qu'elle fait consensus, il n'est pas pertinent d'y revenir (c'est notamment le cas des lapsus). C'est le cas dans l'extrait du corpus (et plus précisément dans l'énoncé de l'utilisatrice) suivant :

D :
U : **j'appelle le support et si il est pas**
R : vous appelez le support, le support étant le technicien qui vous répond

D : *rires*
U : là, je le recherche dans la documentation *rires*
R : d'accord très bien

Figure 4.13 : Un extrait du corpus (12ème minute)

Dans cette séquence, il apparaît une incohérence interne dans l'énoncé de U. Elle est due à une incompatibilité de l'actualisation des contenus sémiques dans une dépendance sémantique

anaphorique et une dépendance sémantique actancielle. La relation actancielle objet du verbe *rechercher* demande une actualisation compatible dans l'isotopie entre son objet (*le*) et son circonstant (*dans la documentation*). La valeur ici actualisée est matériel en opposition à humain. Or, l'objet (*le*) est anaphorique et reprend la substance du contenu de son antécédent (*le support*). Cet antécédent est également l'objet du verbe *appeler* et cette relation actancielle sélectionne le contenu sémique de *support* (par application de la règle de réduction de la polysémie lexicale) qui actualise la valeur humain (en opposition à matériel). Dès lors, l'anaphorique est le lieu d'une incompatibilité humain - matériel, d'où l'incohérence. Cette incohérence, due à une polysémie dans la situation du mot *support* non levée par les dépendances sémantiques, produit une incohérence interne qui fait immédiatement l'objet d'un consensus (d'où les rires qui suivent dans l'interaction). Elle ne remet pas en cause le terrain commun et ne donne donc pas lieu à un dialogue incident.

Le modèle proposé de l'axe syntagmatique éclaire la dimension sémantique des dialogues. Inversement, l'étude des dialogues permet d'expérimenter le modèle pour en tracer les limites. Dans cet objectif, nous allons décrire dans le chapitre suivant le fonctionnement d'un outil logiciel permettant cette expérimentation en articulant dans une interaction homme-machine des stratégies syntagmatiques et des représentations paradigmatiques.

CHAPITRE 5 :

RÉALISATION LOGICIELLE

L'objet de ce cinquième chapitre est double : il s'agit, d'une part, d'expliquer comment les deux modèles de l'axe paradigmatique et de l'axe syntagmatique se co-déterminent dans un modèle opératoire d'analyse des dialogues et, d'autre part, de proposer une implémentation informatique. Le but de cette implémentation n'est pas de mettre au point un analyseur sémantique robuste, comme le sont notamment les applications informatiques d'analyse syntaxique. Il s'agit, pour l'instant, de proposer un outil d'expérimentation logicielle pour tester sur les corpus un premier niveau d'une compétence interprétative artificielle. Ce premier niveau est conçu comme une amorce pour l'apprentissage ([Nicolle 96]) et la compétence interprétative évolue qualitativement au fur et à mesure des analyses menées, car celles-ci engagent une interaction homme-machine dont le but est de faire acquérir au système des signes linguistiques pertinents dans la situation étudiée.

Le traitement interprétatif des énoncés par une articulation entre des stratégies syntagmatiques et des connaissances paradigmatiques est proposée dans la première partie de ce chapitre. Dans la seconde partie, les principes de la réalisation informatique sont donnés et expliqués sur une analyse d'une séquence de corpus. La troisième partie du chapitre présente une expérimentation de l'outil logiciel en détaillant un système hiérarchique différentiel construit dans l'interaction homme - machine qui rend compte de la situation étudiée dans le corpus PIC. Pour conclure nous ferons un premier bilan du modèle proposé et de sa réalisation informatique.

1. Articuler le syntagmatique et le paradigmatique

Le lien entre les modèles paradigmatique (cf. chapitre 3) et syntagmatique (cf. chapitre 4) permet une analyse des énoncés en convoquant des connaissances paradigmatiques dans des stratégies interprétatives syntagmatiques. Ce lien est double car les stratégies interprétatives visant à rendre compte de la contextualisation des contenus sémiques des lexèmes (i.e. les sémèmes qui lui sont rattachés) présupposent une modélisation paradigmatique des signifiés de ces lexèmes, mais cette modélisation paradigmatique est également fondée sur l'analyse syntagmatique. Il ne s'agit pas de mettre en place au niveau paradigmatique une connaissance encyclopédique de la sémantique lexicale avant d'aborder le niveau syntagmatique. Au contraire, les significations qui retiennent notre attention sont celles qui font l'objet de la situation de dialogue étudiée, et c'est par l'analyse d'énoncés que la question de la pertinence de leur représentation est posée.

> Rien ne peut être représenté en langue qui n'ait auparavant été décrit en contexte. ([Rastier 87, p. 62])

L'analyse des énoncés impose deux relations entre le paradigmatique et le syntagmatique. La première, allant du paradigmatique au syntagmatique, vise un premier niveau d'une compétence interprétative confrontée à la situation, et la seconde, allant du syntagmatique au paradigmatique, vise l'apprentissage des systèmes différentiels de significations évoqués dans une situation.

Comme nous l'avons montré dans le troisième chapitre, le modèle Anadia est basé sur un processus interactif de catégorisation qui ne connaît pas de fin car les connaissances données à la machine via ce processus n'ont jamais valeur de dogme et peuvent toujours être raffinées au cours d'autres interactions. Le modèle de l'axe paradigmatique étant fondé sur Anadia, l'acquisition de nouvelles connaissances sémantiques fait l'objet de la même observation. C'est-à-dire que le système informatique proposé aura toujours un rôle d'apprenant autant que d'expert. Il n'y aura pas de distinction entre une première phase d'apprentissage de concepts et une seconde phase d'analyses d'énoncés utilisant ces concepts, mais les analyses seront d'autant plus pertinentes qu'au cours des interactions avec l'humain de nouveaux concepts seront acquis.

Grâce à la structure différentielle du modèle paradigmatique, en soulevant dans l'analyse d'un énoncé un manque de connaissances sémantiques, on est amené, bien sûr, à combler ce manque mais, en plus, à représenter également des connaissances évoquées par celles-ci. Les interactions entre le système et son utilisateur permettent de faire progresser le système beaucoup plus vite qu'une simple description des situations rencontrées. Par exemple, si dans l'analyse d'un énoncé le besoin de décrire la signification du mot *bus* se pose pour la première fois, on va utiliser des sèmes qui, en décrivant un taxème, permettrons de former les sémèmes

proches de 'bus' ('métro', 'train', 'autocar', et pourquoi pas aussi 'RER' si la combinatoire des sèmes le permet). À chaque interaction, on décrit donc finalement beaucoup plus de connaissances que la signification des mots rencontrés. D'analyses en analyses, la machine acquiert et structure (par leurs différences) de nouvelles significations. Les systèmes de significations sont co-construits dans et par l'interaction homme-machine. La sémantique est envisagée comme une connaissance partagée et non comme un savoir individuel issu de processus cognitifs. Ce qui est visé en définitive, c'est la construction d'un système de valeurs partagé où les connaissances soient toujours contextuellement renégociables (comme l'est le terrain commun[73] dans un dialogue). C'est en cela que ce modèle est interactionniste.

Comme l'interaction entre humains est centrale dans l'acquisition de la langue, l'interaction homme-machine est centrale dans l'acquisition de la langue par la machine. Cette interaction est motivée par un enjeu syntagmatique. Elle est le moyen d'une co-construction de systèmes de signification (i.e. de taxèmes) par une adaptation réciproque. La finalité de cette interaction est d'abord paradigmatique mais, ensuite, les sèmes, sémèmes et taxèmes acquis permettent de mettre en évidence de plus en plus d'isotopies et donc d'affiner, en retour, l'analyse syntagmatique.

Pour monter l'articulation des deux niveaux, considérons par exemple l'énoncé suivant, alors qu'aucune connaissance sémantique n'a encore été donnée à l'agent.

> *... vous faites une recherche par sommaire, par index, ou une recherche full text ...*
> (énoncé du corpus PIC du rédacteur à l'intention de l'utilisatrice à la 17ème minute de dialogue)

Étant donné l'absence de connaissances, l'analyse interprétative (i.e. la recherche des isotopies) menée ne conduit à aucun résultat. Cela amène le partenaire humain à se lancer dans une interaction avec le système pour décrire des contenus sémantiques afin d'initier une analyse plus fructueuse. Il peut par exemple fournir une description du sémème de la lexie *sommaire*, comme celle proposée au chapitre 3. Les distinctions verbalisées dans le corpus permettent de construire le sémème de *sommaire* dans la combinatoire des sèmes [Présentation de l'information : globale *vs*. précise], [Support de l'information : papier *vs*. logiciel], [Accès à l'information : interactif *vs*. figé] et [Structure de l'information : hiérarchique *vs*. linéaire] :

[73] Cf. [Clark & al. 86].

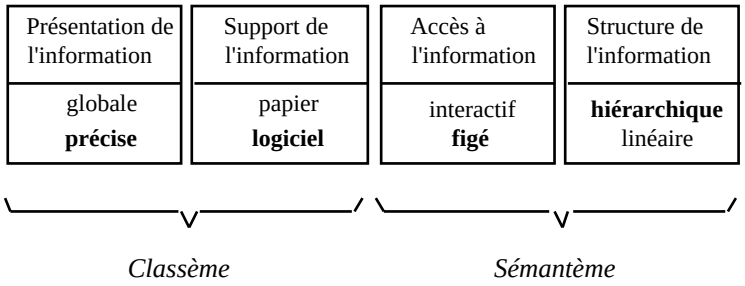


Figure 5.1 : Sémème de *sommaire*

Pour mettre en forme ce sémème, on a construit le taxème représentant la catégorie des aides en ligne (cf. Chapitre n°3) dans laquelle on a exprimé les sémèmes 'sommaire', 'assistant', 'index' et 'full text'. À l'issue de la construction du sémème de *sommaire*, la relance de l'analyse syntagmatique de l'énoncé apporte des résultats bien plus satisfaisants car, en construisant le taxème des aides en ligne, on a défini, en plus du sémème 'sommaire', les sémèmes des lexies *index* et *full text.* Il s'en suit que l'analyse met au jour 4 isotopies (deux génériques : celles des sèmes [Présentation de l'information : globale *vs*. précise] et [Support de l'information : papier *vs*. logiciel], et deux spécifiques[74] : celles des sèmes [Accès à l'information : interactif *vs*. figé] et [Structure de l'information : hiérarchique *vs*. linéaire]) donnant lieu aux 4 contraintes du potentiel de sens suivantes.

- Il est question de la valeur *précise* (actualisée dans 'sommaire', 'index' et 'full text')
  dans le domaine d'interprétation *Présentation de l'information*
  et dans l'opposition *précise vs. globale*.

- Il est question de la valeur *logiciel* (actualisée dans 'sommaire', 'index' et 'full text')
  dans le domaine d'interprétation *Support de l'information*
  et dans l'opposition *logiciel vs. papier*.

- Il est question de la valeur *figé* (actualisée dans 'sommaire' et 'index')
  dans le domaine d'interprétation *Accès à l'information*
  et dans l'opposition *figé vs. interactif.*

- Il est question de la valeur *linéaire* (actualisée dans 'full text' et 'index')
  dans le domaine d'interprétation *Structure de l'information*
  et dans l'opposition *linéaire vs. hiérarchique.*

[74] Cet exemple montre la cohérence interne de l'énoncé (cf. chapitre 4) car les sèmes récurrents dans les isotopies spécifiques ne sont pas actualisés de la même façon dans les sémèmes 'sommaire' (figé & hiérarchique), 'index' (figé & linéaire) et 'full text' (interactif & linéaire), ce qui est en accord avec les prescriptions syntagmatiques de la disjonction (*recherche par sommaire, par index, **ou** une recherche full text*).

Ceci montre le retour au niveau de l'analyse syntagmatique de la description paradigmatique initiée par la mise en forme d'un sémème pour la lexie *sommaire*. La mise en forme d'un contenu sémantique pour *sommaire* est un choix du partenaire humain dont la nécessité lui est apparue suite à une première analyse syntagmatique infructueuse. En exhibant, en interaction avec la machine, les raisons de son interprétation de l'énoncé, il a permis une seconde analyse interprétative plus satisfaisante. La première analyse était infructueuse parce que la machine n'avait aucune connaissance paradigmatique. Si elle en avait eu, mais que ces connaissances aient été insuffisantes, le résultat n'aurait pas non plus était satisfaisant du point de vue du partenaire. Par exemple, si l'on avait défini les sémèmes 'sommaire', 'index' et 'full text' sans les faire apparaître dans une table issue de la sous-catégorisation de 'aide en ligne' (i.e. si on les avait défini dans une table indépendante de la table des outils d'aide), les classèmes de ces sémèmes auraient été vides. Ceci aurait eu pour conséquence de ne trouver dans l'interprétation de l'énoncé que des isotopies spécifiques. C'est-à-dire que les contraintes liées aux domaines d'interprétation *Présentation de l'information* et *Support de l'information* qui proviennent des isotopies génériques n'auraient pas été mises en forme, d'où l'insuffisance de l'analyse. Cet exemple montre l'articulation des deux niveaux d'analyse (paradigmatique et syntagmatique) dans un processus interactif où le système informatique mène un calcul interprétatif soumis à la validation du partenaire humain.

Dans cette articulation, ce sont moins les compétences analytiques de la machine qui sont visées que la mise au jour de ses défauts. En effet, ces défauts sont révélateurs de contenus sémantiques insuffisants ou inexistants et ils donnent lieu à des interactions avec l'humain afin d'expliciter les contenus sémantiques en cause, pour ainsi proposer une analyse plus satisfaisante. Les incomplétudes des analyses sont alors le moteur de la constitution de l'amorce interactive de l'interprétation. Le statut même de l'analyse, et donc du calcul qui est réalisé, sont changés par rapport aux approches qui envisagent l'analyse comme révélant un monde préalablement donné. Le résultat a une valeur intersubjective et l'intérêt du processus est dans l'apprentissage de la compétence linguistique. La connaissance sémantique ne peut être donnée à une machine "d'un seul coup" car personne ne la possède mais l'apprentissage d'une compétence linguistique peut venir de la pratique de cette même compétence (même à un état embryonnaire).

2. Réalisation de l'outil logiciel

Un outil logiciel, qui effectue l'articulation entre les modèles de l'axe paradigmatique et de l'axe syntagmatique, opérationnalise les principes de base retenus dans cette thèse qui, de fait, peuvent être testés. Il met ainsi en place les premières bases d'un analogue machine de la

compétence interprétative humaine qui peut être expérimentée sur le corpus ([Beust 98], [Beust & al. 99]). L'outil logiciel proposé a pour but de mettre en forme, à partir d'un énoncé, un ensemble de contraintes linguistiques sur les sens possibles en contexte de cet énoncé. Plus précisément, ces contraintes sont des contraintes de monstration (cf. Chapitre 1) et c'est donc une partie du potentiel de sens que recherche actuellement la machine, étant donné qu'elle ne produit pas les contraintes de prédication, d'argumentation et de modalité.

Le traitement interprétatif est effectué sur des séquences d'énoncés de dialogue dans une situation interactive de collaboration avec un partenaire humain. L'objectif de cette interaction est d'examiner les énoncés un à un afin d'arriver à un consensus entre l'interprétation de l'humain et les contraintes sémantiques proposées par l'outil logiciel (cf. Figure 5.2). À l'issue du calcul interprétatif de chacun des énoncés, un jeu de contraintes est soumis à la validation du partenaire humain. Si les contraintes ne sont pas jugées satisfaisantes au regard du sens de l'énoncé perçu par l'humain, c'est qu'il manque au système des connaissances paradigmatiques, c'est-à-dire des descriptions de contenus sémiques. Il convient alors de renseigner la machine sur des significations en jeu dans l'énoncé et de relancer le processus interprétatif en reformant les nouvelles contraintes sémantiques qui en découlent. Dans cette situation interactive, l'interprétation n'est pas considérée comme un processus individuel mais comme une activité conjointe d'analyse d'actes de langage. Son but est l'acquisition de représentations sémantiques et l'amélioration de la compétence interprétative artéfactuelle.

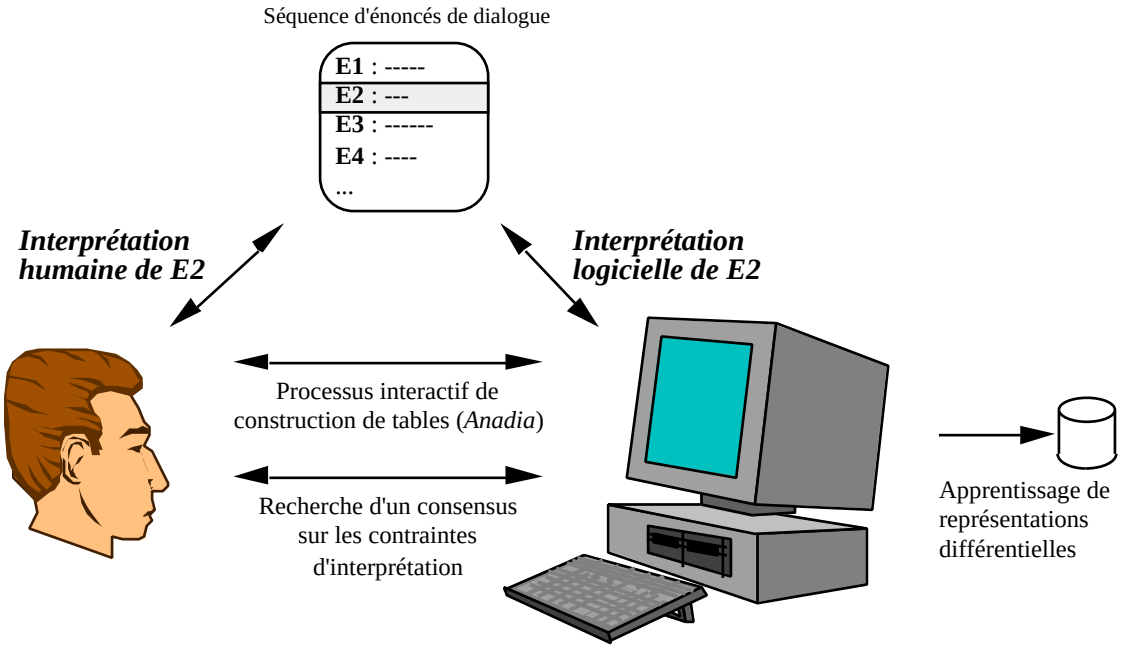


Figure 5.2 : La situation interactive d'utilisation de l'outil logiciel

Comme on l'a montré au cours du chapitre n°4, le calcul des contraintes du potentiel de sens est réalisé par une analyse des isotopies de la chaîne. Le calcul des isotopies convoque deux types de données :

- Les relations syntagmatiques donnant lieu à des phénomènes de dépendance sémantique. Elle proviennent d'une l'analyse syntaxique de l'énoncé.

- Les contenus sémiques des lexies composant la chaîne. Ils proviennent des connaissances sémantiques paradigmatiques du système.

Le processus opératoire d'interprétation est réalisé en trois phases : la première est un étiquetage morpho-syntaxique de l'énoncé, la deuxième est une recherche dans la base de connaissances des contenus sémiques des lexies de la chaîne et la troisième est l'extraction d'isotopies dans les dépendances sémantiques (cf. Figure 5.3). La mise en forme des contraintes, à partir des isotopies, forme le résultat du traitement interprétatif d'un énoncé.

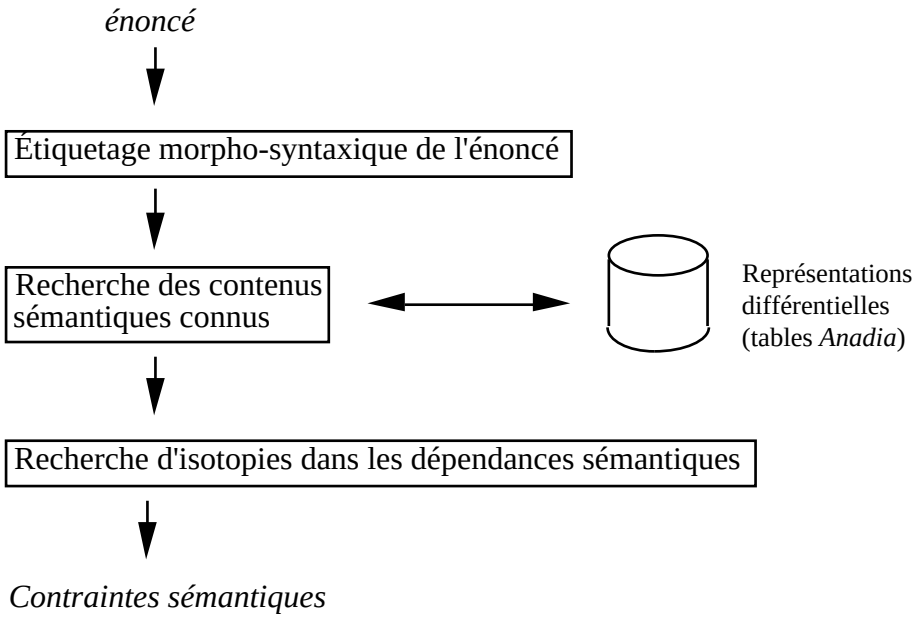


Figure 5.3 : Phases du traitement interprétatif de l'outil logiciel

L'implémentation présentée ici est réalisée en CommonLisp et elle vient se greffer à l'environnement de programmation de MCL™ version 3.0[75]. Le menu correspondant offre plusieurs possibilités : commencer une analyse, afficher la liste des lexies dont un contenu sémique a été précisé (option *Dictionnaire* du menu, cf. Figure 5.4) et en inspecter les descriptions, ou formuler de nouveaux contenus (options *Nouveau sème* et *Nouveau sémème* du menu).

[75] *Macintosh Common Lisp* version 3.0, Copyrigth © 1995 Digitool, Inc.

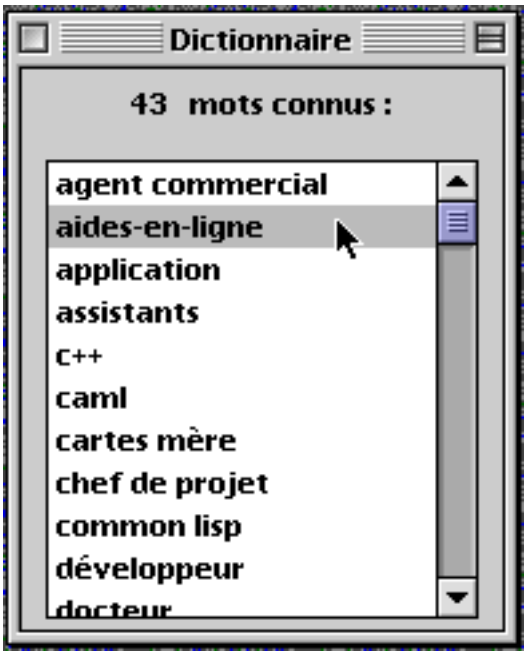


Figure 5.4 : Fenêtre de consultation des lexies connues du système

En cliquant sur un lexie de la liste, on a accès à ses propriétés lexicales (cf. figure 5.7) et à son contenu sémique (cf. figure 5.8).

Pour détailler le déroulement du processus d'analyse syntaxico-sémantique, reprenons un exemple abordé dans le chapitre n°4. Il s'agit de l'énoncé suivant issu du corpus PIC. Considérons qu'alors aucun contenu sémantique n'a encore été précisé.

... Non, poste à l'IUT, c'est le poste téléphonique ...
(énoncé du développeur à la 69ème minute de dialogue)

Pour commencer l'analyse de cet énoncé, l'usager est invité à sélectionner un fichier texte contenant la séquence du corpus où l'énoncé figure. Une fois chargé ce fichier, les énoncés apparaissent dans une fenêtre où il est possible de les sélectionner un à un en cliquant dessus dans une liste (cf. Figure 5.5) pour déclencher une analyse sémantique de l'énoncé sélectionné.

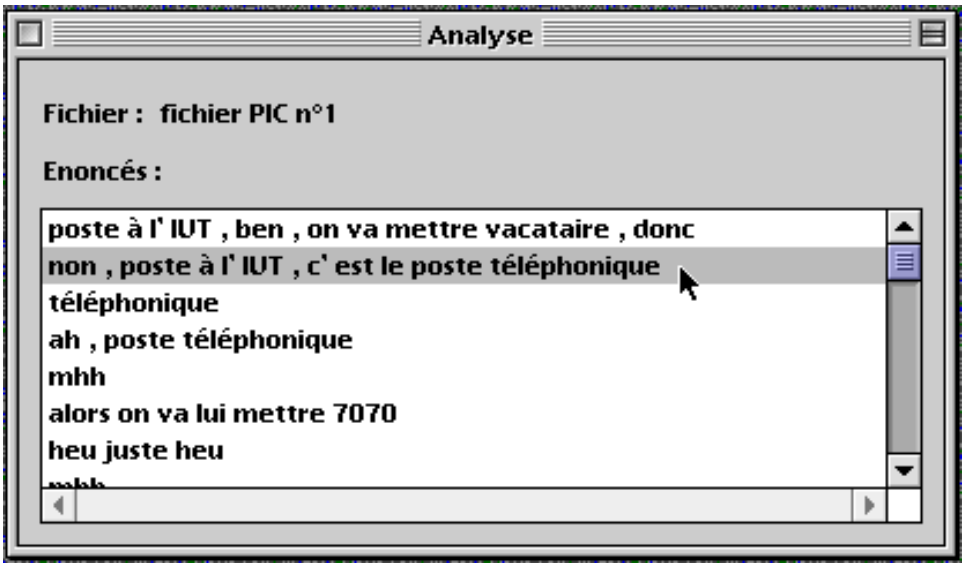


Figure 5.5 : Fenêtre d'analyse des énoncés d'un dialogue

La sélection de l'énoncé entraîne le début de son analyse interprétative. Elle produit l'affichage d'une fenêtre qui lui est propre (cf. Figure 5.6). Cette fenêtre permet de lancer le calcul des contraintes linguistiques de l'énoncé (par le bouton "Potentiel de sens ...").

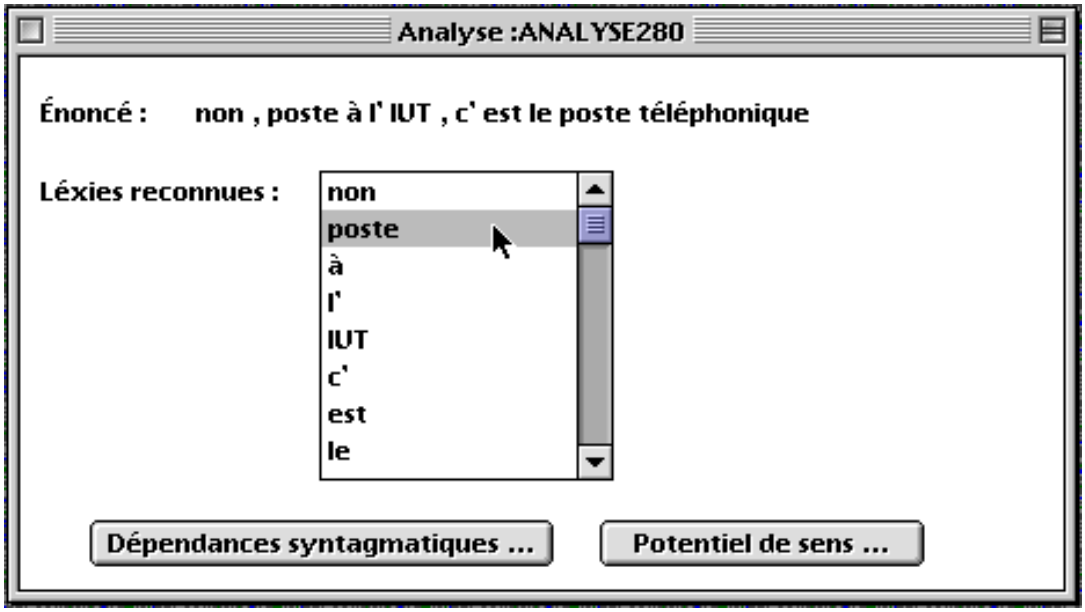


Figure 5.6 : Fenêtre d'analyse d'un énoncé

2.1. Étiquetage morpho-syntaxique de l'énoncé

Le processus préalable à l'analyse interprétative d'un énoncé consiste en une analyse morpho-syntaxique. Ainsi, les fichiers où figurent les séquences de corpus à traiter (ceux que l'on charge grâce au menu de l'agent) contiennent les énoncés de la séquence ainsi que les

résultats de leur étiquetage morpho-syntaxique. Cet étiquetage a pour objectif de produire des informations sur le découpage en mots de l'énoncé et sur la constitution des syntagmes où il convient de chercher des dépendances sémantiques, notamment les groupes nominaux et les groupes verbaux. Ces informations proviennent de l'analyseur syntaxique développé par J. Vergne[76] ([Vergne 94], [Giguet & al. 97a], [Giguet & al. 97b]). Le résultat de l'étiquetage morpho-syntaxique de l'énoncé *Non, poste à l'IUT, c'est le poste téléphonique* fournit les informations suivantes :

Lexème	**Catégorie syntaxique**	**Propriétés lexicales**
Non	adverbe de bloc	
poste	nom	singulier
à	préposition	
l'	déterminant	3ème personne du singulier
IUT	nom	nom propre ou sigle, 3ème personne du singulier
c'	pronom personnel sujet	masculin, 3ème personne du singulier
est	verbe conjugué transitif	masculin, 3ème personne du singulier
le	déterminant	masculin, 3ème personne du singulier
poste	nom	masculin, 3ème personne du singulier
téléphonique	adjectif épithète	masculin, 3ème personne du singulier
Groupes nominaux		*l'IUT* *le poste téléphonique*
Relations actancielles sujet - verbe		*c' - est*
Relations actancielles verbe - objet		*est - poste*

Dans la fenêtre d'analyse d'un énoncé (cf. Figure 5.6), le bouton "Dépendances syntagmatiques ..." permet l'affichage des groupes nominaux et des relations actancielles, et la liste des lexies reconnues permet l'examen des connaissances syntaxico-sémantiques sur chacune d'elles (il suffit pour cela de cliquer dessus dans la liste). Cet examen des connaissances sur une lexie amène l'affichage de deux types d'informations :

• Les informations lexicales sur cette lexie (cf. Figure 5.7). Elles proviennent de la fonction de la lexie dans l'énoncé, et sont obtenues par l'analyseur syntaxique.

[76] Des démonstrations de cet analyseur sur quelques sélections de corpus sont disponibles sur le WEB à l'adresse *http://www.info.unicaen.fr/~giguet/syntaxique.html*

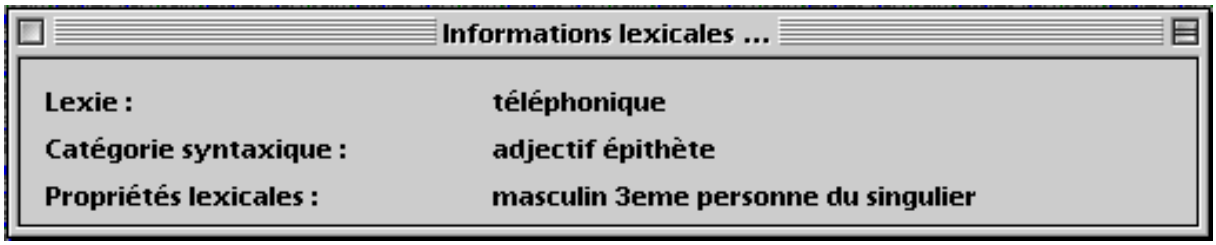


Figure 5.7 : Informations lexicales sur la lexie *téléphonique* de l'énoncé *"Non, poste à l'IUT, c'est poste téléphonique"*

• Le contenu sémique de la lexie (cf. Figure 5.8). On va montrer dans la suite comment les contenus sémiques sont construits.

Figure 5.8 : Informations sémantique sur la lexie *téléphonique.*

La lexie est associée à un sémème dont le nom apparaît dans la liste (quand un nom d'objet est formé du nom de sa classe et d'un numéro comme c'est ici le cas, c'est que le nom a été attribué par le système parce qu'il n'avait pas été précisé par l'utilisateur)

2.2. Recherche et mise en forme des représentations paradigmatiques

Dans la réalisation de l'outil logiciel, l'analyse syntagmatique des énoncés est construite comme un module qui vient se greffer sur Anadia. La description de l'implémentation permet d'expliquer comment cet outil exploite les fonctionnalités de catégorisation à des fins de représentation paradigmatique de systèmes de significations.

2.2.1. L'implémentation d'Anadia

Le chapitre n°3 a montré comment construire des catégories qui aient de bonnes propriétés cognitives, conceptuelles et algébriques (en ce qui concerne la connexité des topiques). Pour

tester une catégorisation des objets d'un domaine relativement à ses besoins, pour concrétiser un accord avec d'autres dans des interactions conversationnelles (ou des interactions de travail comme c'est le cas dans l'expérience mise en place dans le projet PIC), ou pour rendre manifestes des désaccords, les tables et les topiques sont très utiles. La formulation du point de vue de l'utilisateur résulte d'une activité mentale ou sociale de réflexion sur les conséquences de la perception, de l'action ou de la négociation. Anadia en fait des objets sociaux en les rendant partageables. Comme il y a beaucoup d'aller-retours dans la modélisation, un logiciel interactif d'aide à la construction des tables est nécessaire. Il présente le point de vue considéré sous des aspects dont l'usager n'avait pas conscience, en montrant les tables et surtout en montrant les topiques. Dans cette interaction, l'utilisateur est amené à ne pas considérer de critères superflus (car il sont révélés par l'examen de la table), ce qui est un moyen d'éviter l'explosion combinatoire.

Le logiciel Anadia existe en plusieurs versions : une en Pascal [GIL 95], une en Java [Delépine & al. 96] et celle que nous présentons ici, réalisée dans le paradigme de la programmation orientée objet à l'aide du langage MCL™ version 3.0. La version en CommonLisp est un module qui vient enrichir l'environnement de programmation pour s'intégrer dans d'autres projets[77], en réifiant comme objets les attributs, les registres, les tables et les topiques. Ces objets sont utilisables par l'intermédiaire d'une interface graphique accessible grâce à un menu qui vient s'ajouter au menu de l'application MCL 3.0.

L'utilisateur peut créer ses attributs (cf. Figure 5.9) en donnant pour chacun ses valeurs et éventuellement un nom (si ce n'est pas le cas, le système en donne un d'office). Il peut créer des registres en choisissant les attributs adéquats (cf. Figure 5.10), c'est-à-dire les attributs qui, selon son point de vue, sont de même niveau pour son projet de catégorisation.

[77] Notamment, cette version a été réutilisée par Y. Jullien dans un but de modélisation de groupes d'oppositions ([Jullien & al. 97]).

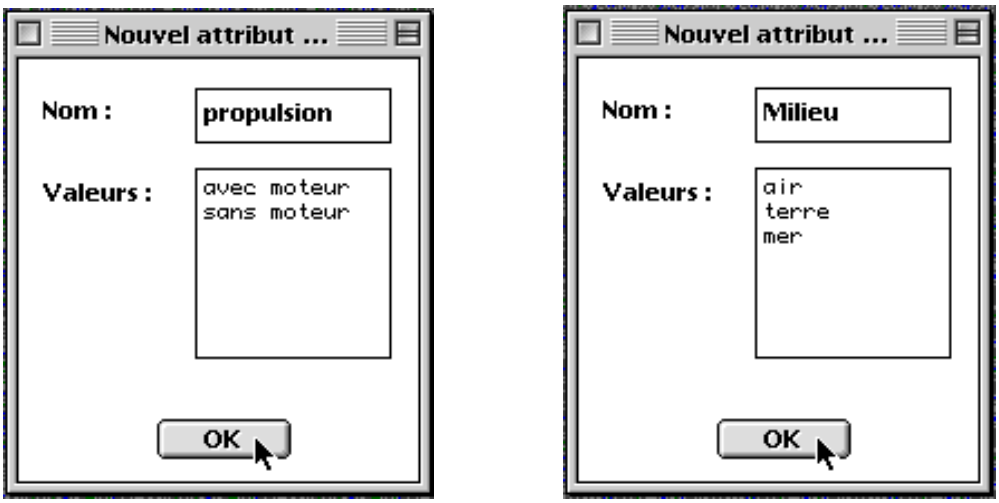


Figure 5.9 : Créations de nouveaux attributs

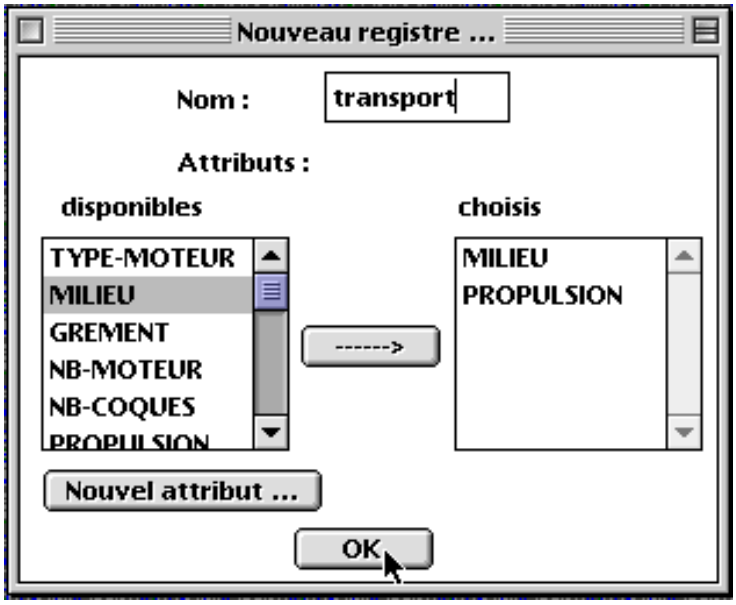


Figure 5.10 : Création d'un nouveau registre

Une fois un registre constitué, le logiciel calcule la table qui résulte de la combinatoire des valeurs des attributs du registre. Il affiche cette table dans une fenêtre qui forme alors le principal lieu d'interaction entre l'utilisateur et le système (cf. Figure 5.11). Du point de vue de la machine, les tables forment ainsi une représentation mémoire dynamique dans l'interaction. Elle opérationnalisent les tableaux issus de la méthode de description onomasiologique.

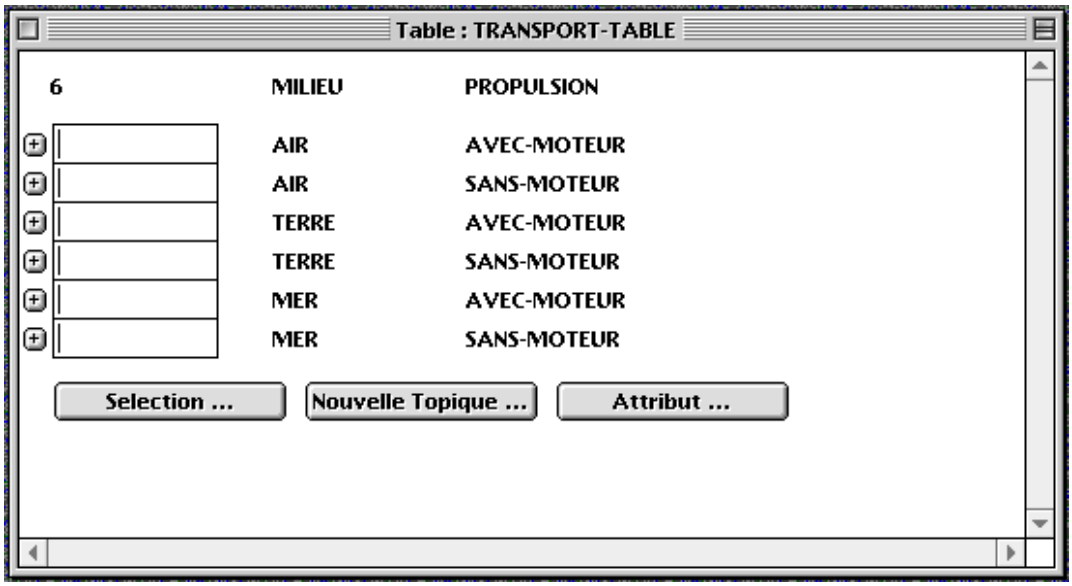


Figure 5.11 : Fenêtre de la table formée par les attributs MILIEU et PROPULSION.

Dans cette fenêtre, l'utilisateur examine chaque ligne de la table afin de trouver celles qui décrivent des jeux de valeurs représentant des sous-catégories effectives (i.e. dont il peut donner des exemples). Quand c'est le cas, il a la possibilité de donner un nom à cette ligne et c'est ainsi qu'il construit la sélection à partir de la combinatoire. Dans cette fenêtre de la table, l'utilisateur peut, à tout moment, faire plusieurs actions (cf. Figure 5.12) :

• il peut demander à ce que seule la sélection soit affichée (et non toute la combinatoire),

• il peut demander le calcul de la topique correspondant à l'état actuel de la sélection (elle s'affiche alors dans une fenêtre comme le montre la figure 5.13),

• il peut créer un attribut ayant pour valeurs les noms des sous-catégories de la sélection[78],

• il peut demander à recatégoriser une sous-catégorie de la sélection. Pour cela, il clique sur un bouton situé à gauche du nom de la ligne correspondant à la sous-catégorie, et le système redemande alors de constituer un nouveau registre donnant lieu à une nouvelle table au nom de la sous-catégorie en cause. Selon différents points de vue, on peut vouloir sous-catégoriser une catégorie de plusieurs façons. Anadia offre cette possibilité en assurant qu'une nouvelle sous-catégorisation n'efface pas une sous-catégorisation précédente de la même catégorie. C'est pour cela que le système attribue un numéro propre à chaque table issue d'une sous catégorisation. Ce numéro est indiqué dans le nom de la nouvelle table. Par exemple, une première sous-catégorisation de la catégorie *avion*

[78] Par exemple, dans une table représentant une catégorisation des *ordinateurs*, si l'utilisateur a pu nommer trois lignes de la façon suivante *Mac, PC, Station Sun*, il peut demander à ce que soit formé un attribut ayant pour valeurs *Mac, PC, Station Sun* dans le but de faire de nouvelles différences dans d'autres catégorisations.

permet de former la table AVION-TABLE1 et une seconde sous-catégorisation de *avion* sous un autre point de vue permet de former la table AVION-TABLE2.

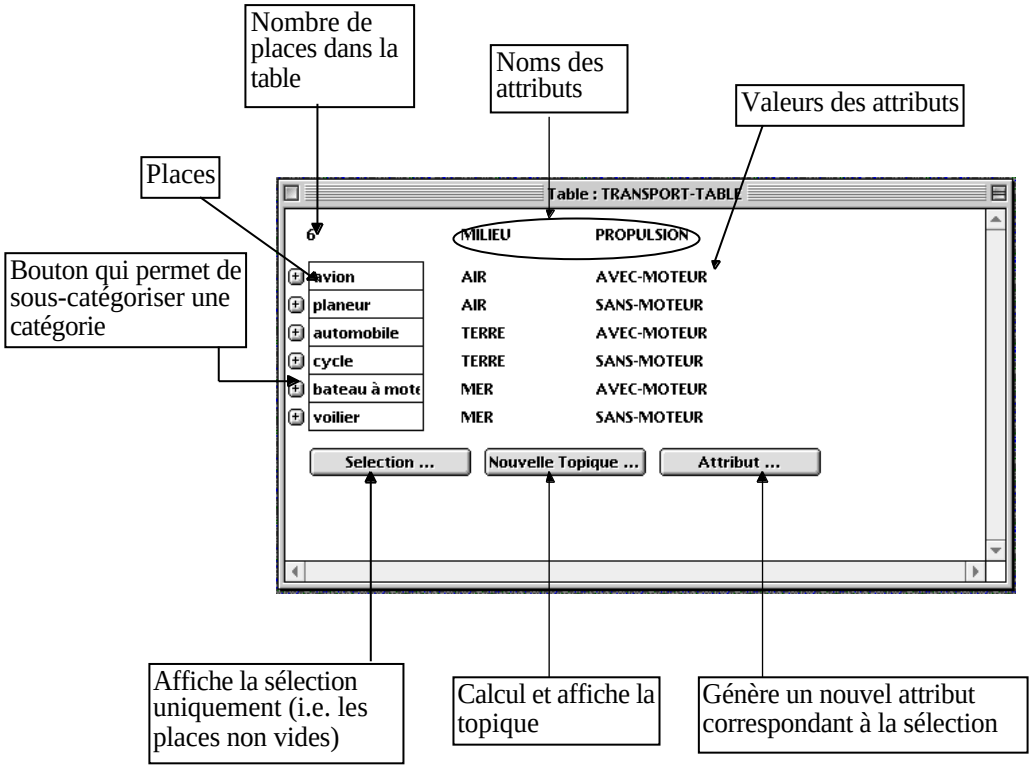


Figure 5.12 : Composants des fenêtres où s'affichent les tables

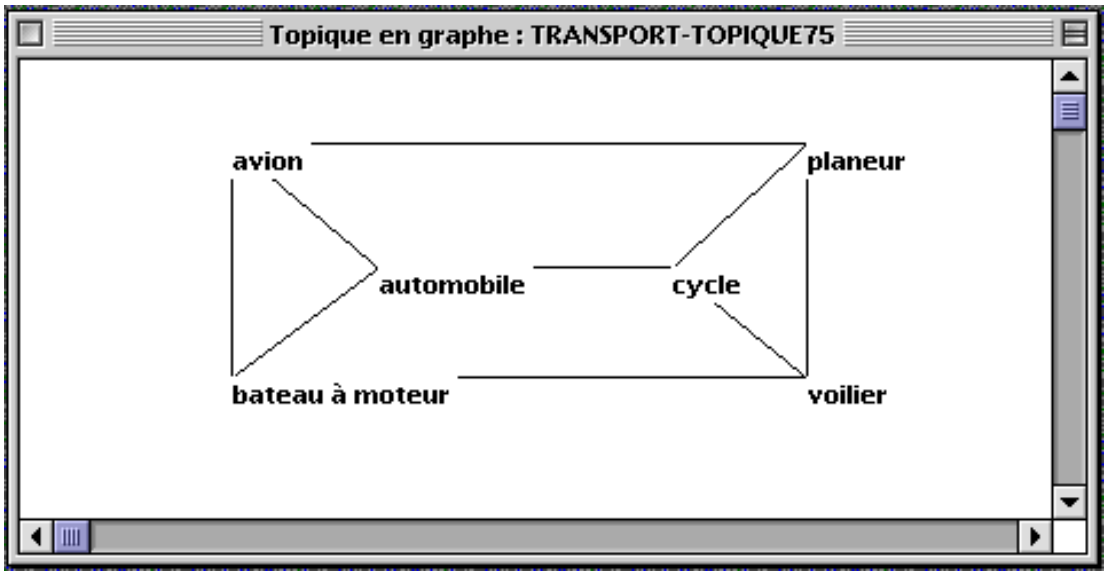


Figure 5.13 : Fenêtre de topique

Grâce à ces fonctionnalités, l'utilisateur réalise des sessions de travail qu'il peut sauvegarder dans un fichier qui contiendra les déclarations de ses attributs, de ses registres, de

ses tables et de ses topiques. Il peut bien sûr à tout moment demander le chargement d'une session déjà sauvegardée en sélectionnant l'option *Charger* du menu Anadia.

Il y a plusieurs choix de configuration du système (cf. Figure 5.14). Ainsi, il peut choisir le mode d'affichage des topiques (en graphe ou sous forme de listes de successeurs) mais surtout, il peut paramètrer la relation servant à construire les topiques. La plus utilisée est la relation de différence à un trait près (c'est celle qui est choisie par défaut), mais l'utilisateur peut demander à ce que les topiques soient calculées avec une relation à *n* traits près, exactement *n* ou bien au plus *n* traits.

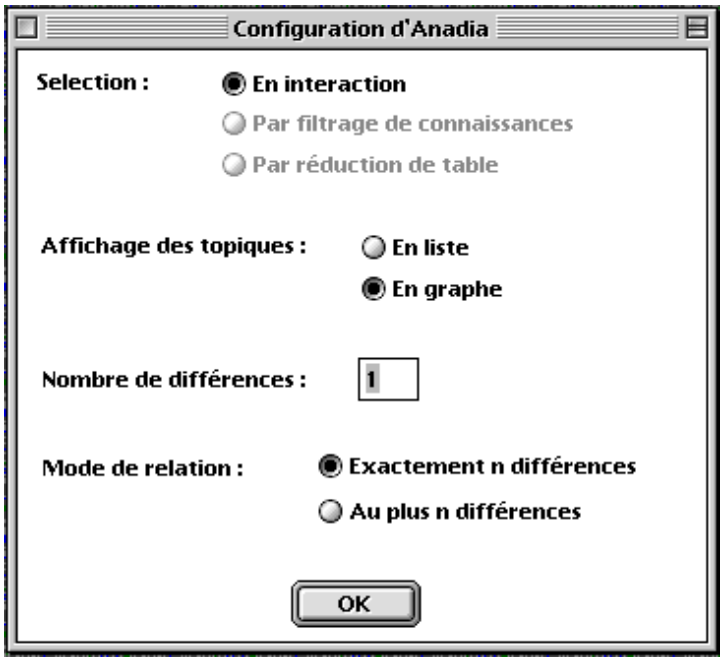


Figure 5.14 : Fenêtre de configuration

D'autres fonctionnalités sont possibles comme, par exemple, d'autres modes de sélection des places de la table : rechercher dans une base de connaissances les catégories effectives définies par un certain jeu de valeurs, ou encore rechercher si certains jeux de valeurs n'auraient pas déjà été envisagés dans d'autres tables déjà réalisées (ceci explique les deux choix grisés du mode de sélection dans le fenêtre de configuration que montre la figure 5.14). Une autre fonctionnalité envisagée concerne le mode de sélection en interaction (i.e. celui qu'on a décrit), pour offrir à l'utilisateur l'aide d'assistants logiciels dans le processus interactif, assistants qui auraient pour tâche d'examiner les propriétés algébriques de la représentation en cours. Par exemple, un assistant peut avoir comme tâche de surveiller la proportion dans la table de places vides par rapport au places effectives et de renseigner l'utilisateur sur les possibles raisons d'une disproportion (par exemple, voir des attributs de même niveau qui seraient sémantiquement dépendants). De même, on peut imaginer un assistant qui inspecterait la

connexité des topiques produites ou les utilisations de mêmes attributs dans différents niveaux de représentation. Enfin, nous prévoyons de préciser l'affichage des topiques sous forme de graphes en étiquetant les arcs par le (ou les) différence(s) entre les sommets du graphe.

2.2.2. Représenter les contenus sémantiques avec Anadia

L'outil logiciel proposé exploite le processus interactif mis en place par Anadia comme la modalité de co-construction de taxèmes entre l'usager et la machine. Les connaissances sémantiques sont des descriptions componentielles telles qu'on les a détaillées au chapitre n°3. Les significations des lexies sont décrites par un ou plusieurs sémèmes qui représentent leurs emplois possibles et ces sémèmes (structurés en classèmes et sémantèmes) sont constitués de sèmes oppositionnels actualisés (i.e. précisant une valeur au sein d'une opposition). Ainsi, les sèmes spécifiques d'un sémème sont ceux qui ont permis de former la table dans laquelle apparaît la lexie en tant qu'élément de la sélection et les sèmes génériques sont ceux qui proviennent des tables de niveau supérieur dans la hiérarchie mise en place. Les connaissances sémantiques de la machine sont donc inscrites dans un ensemble de tables (où certaines sont hiérarchiquement subordonnées à d'autres) et la recherche des descriptions componentielles d'une lexie de l'énoncé consiste alors à la trouver comme élément de la sélection d'une ou plusieurs tables et à calculer les sémèmes issus de la place de ces tables dans la hiérarchie. Avec la hiérarchie de tables proposée au chapitre n°3 pour décrire les outils d'aide à l'utilisation d'un logiciel, on forme ainsi un sémème pour la lexie *assistant* :

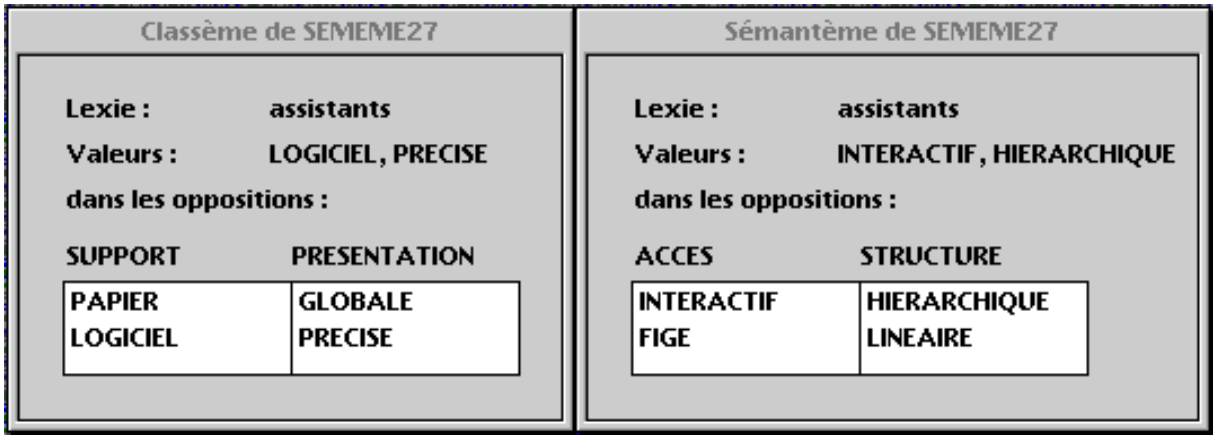


Figure 5.15 : Sémème de la lexie *assistants* (cf. chapitre n°3)

Si les contenus sémiques des lexies de la chaîne ne sont pas connus avant l'analyse, ils doivent être acquis à ce moment. C'est pour cette raison qu'à la phase de recherche parmi les contenus sémiques déjà décrits, succède une phase de description de nouveaux contenus, ou de reformulation de contenus déjà connus dans le cas où une catégorisation est remise en cause.

Dans cette phase, former un nouveau sémème pour une lexie consiste à définir ce sémème au sein d'un taxème. Ce taxème est formé à partir d'un ensemble de sèmes. Décrire un sémème d'une lexie revient alors à faire figurer la lexie comme élément de la sélection d'une table qui définit le taxème. Cette table est construite par la combinatoire d'attributs qui proviennent des sèmes retenus pour construire le taxème, car chaque sème oppositionnel donne lieu à un attribut (cf. Figure 5.16). La sélection de la table fournit les éléments du taxème (la topique produite en montre la structure oppositionnelle) et les jeux de valeurs d'attributs déterminant les éléments de la sélection permettent de savoir, pour chaque sémème du taxème, la façon dont les sèmes ayant permis de constituer la table sont actualisés.

La reformulation de contenus sémiques peut conduire à des modifications plus ou moins importantes dans la hiérarchie de tables. S'il s'agit de rajouter un sème spécifique à un sémème, les modifications seront localisées à une feuille de l'arbre des tables. Il s'agira de revoir la structure d'un seul taxème en examinant comment ses éléments actualisent le nouveau sème spécifique et en vérifiant si la nouvelle combinatoire met en évidence des nouveaux sémèmes dans le taxème. À l'inverse, s'il s'agit de rajouter un sème générique à un sémème, les modifications vont toucher plusieurs niveaux de la hiérarchie. Il convient, premièrement, de trouver quelle table T de l'arbre il faut redévelopper avec ce nouveau sème (i.e. trouver le taxème dans lequel ce sème est spécifique) et, deuxièmement, de vérifier que le sème est également générique pour tous les sémèmes des tables qui sont issues de T. Si ce n'est pas le cas, il faut revoir toutes les sous-catégorisations effectuées à partir de T. Par exemple, si l'on veut rajouter un sème générique au sémème 'assistant', il faut d'abord repréciser le sémème 'aide en ligne' (d'où provient 'assistant' par sous-catégorisation) en redéveloppant la table des outils d'aide (cf. chapitre n°3). Il faut ensuite vérifier que le sème est bien générique pour les autres sémèmes issus de la sous-catégorisation de 'aide en ligne' (i.e. 'sommaire', 'full text' et 'index'). C'est pour aider l'utilisateur dans ce genre de remise en cause de la hiérarchie de tables, que sont envisagées les fonctionnalités d'assistance d'Anadia (par examen des tables et des topiques dans le processus interactif).

Reprenons le cours de l'analyse de l'énoncé *Non, poste à l'IUT, c'est le poste téléphonique* et choisissons de formuler les contenus de *poste* et de *téléphonique*. On met en forme ces contenus en lançant la boucle interactive où l'on va définir des sèmes oppositionnels, constituer des registres avec les attributs qui en sont issus, développer les tables qui proviennent des registres, former des sélections dans ces tables et ainsi constituer des topiques.

Pour décrire le contenu sémique de *téléphonique*, on considère le sème [Communication à distance : téléphone *vs*. courrier] (cf. chapitre n°4) et on définit un registre formé par l'unique attribut représentant l'opposition téléphone *vs*. courrier.

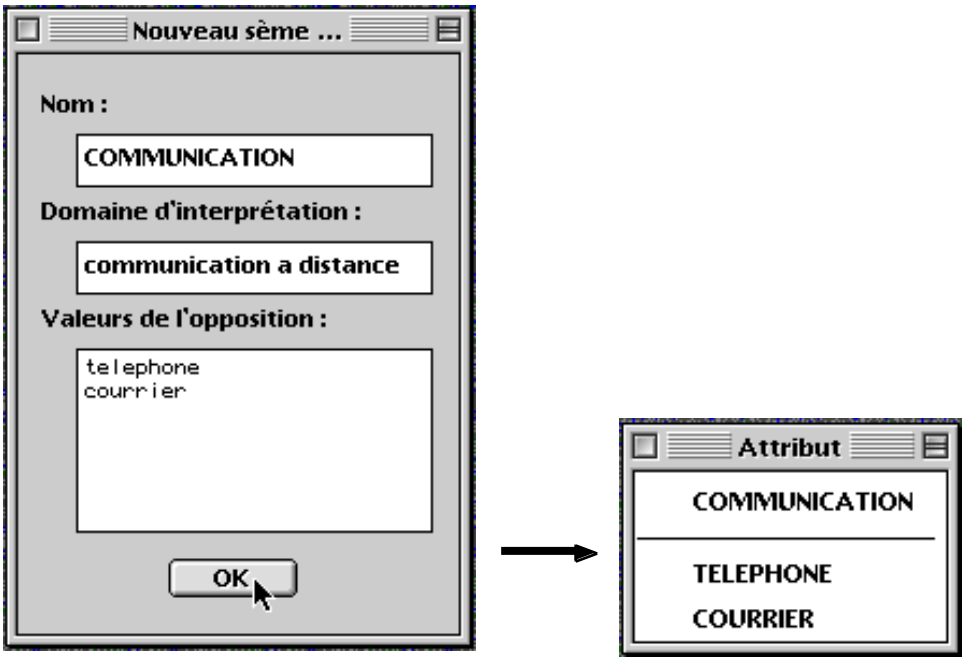


Figure 5.16 : Instanciation d'un attribut à partir d'un sème.

On obtient alors une table à deux places. La première place, celle correspondant à la valeur téléphone, permet de définir l'actualisation de [Communication à distance : téléphone *vs*. courrier] pour le sémème 'téléphonique' par opposition avec le sémème 'manuscrite' qui est issu de la deuxième place, celle correspondant à la valeur courrier (ceci met en place une topique à deux éléments 'téléphonique' et 'manuscrite'). Pour exprimer avec Anadia la signification de *poste* en terme de poste téléphonique, on peut choisir de sous-catégoriser la ligne de la table précédente correspondant à la valeur téléphone en utilisant l'attribut provenant du sème [Matériel téléphonique : appareil *vs*. ligne]. Ceci produit une seconde table à deux places, où la première (appareil) permet de définir le sémème 'poste' représentant la signification de poste téléphonique, et la seconde (ligne) permet de définir le sémème 'réseau' représentant la signification de réseau téléphonique (de même que précédemment ceci met en place une topique à deux éléments : 'poste' et 'réseau', cf. figure 5.17).

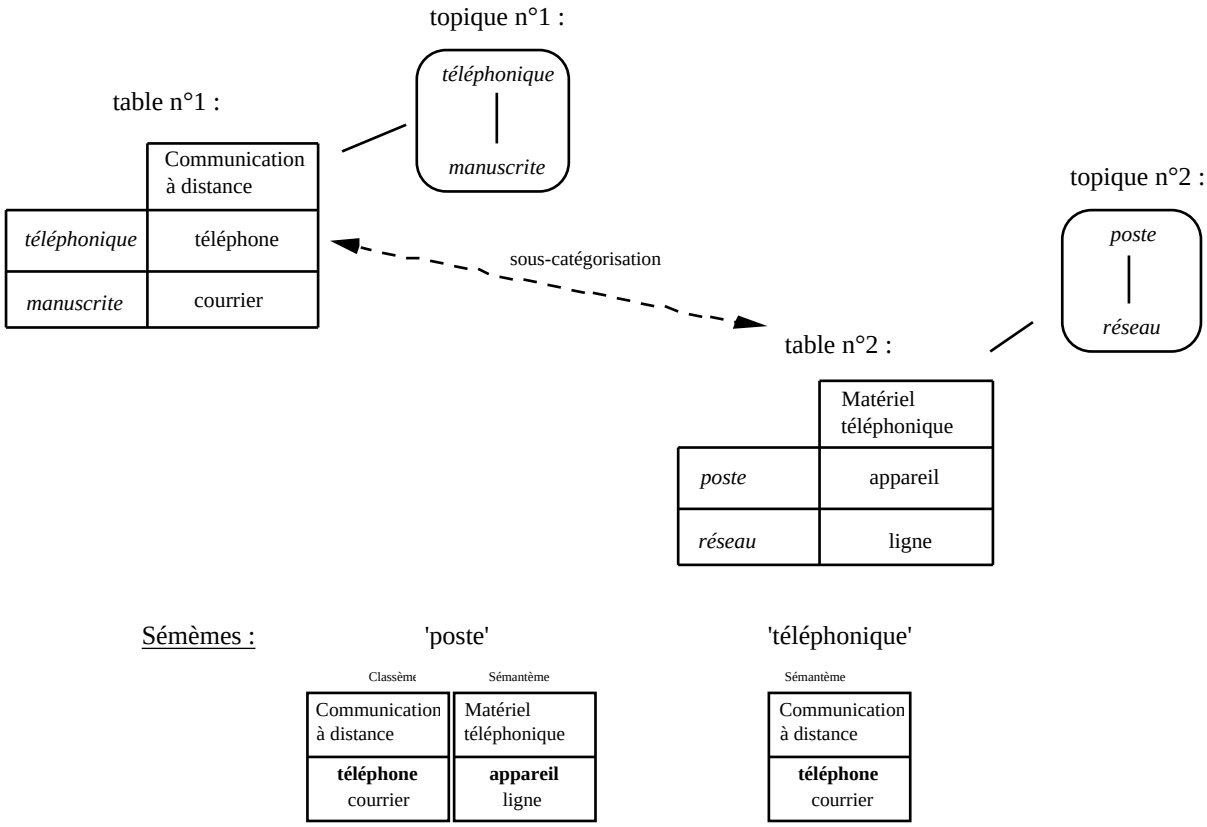


Figure 5.17 : Une session Anadia

Dans le contexte de dialogue dont est extrait l'énoncé, *poste* est polysémique car il peut aussi signifier "statut professionnel". On définit donc un autre sémème 'poste' représentant cette signification en constituant une table à partir du sème [Activité professionnelle : statut *vs*. fonction]. Le sémème 'poste' recherché actualise ce sème dans la valeur statut. Il est différencié du sémème 'profil de poste' qui actualise la valeur fonction. La lexie *poste* est alors associée à deux sémèmes, représentant chacun un des sens contextuellement possible (cf. figure 5.18 et 5.19).

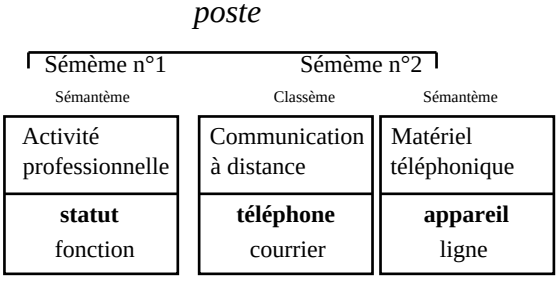


Figure 5.18 : Le contenu sémique de *poste*

Figure 5.19 : Informations sémantiques sur la lexie *poste*.

DE-TRAVAIL est le nom du sémème décrivant la signification de *poste* en terme de statut professionnel (cf. Chapitre n°4) et TELEPHONIQUE, le nom du sémème décrivant une signification en terme de matériel téléphonique.

Dans la fenêtre "Info. sémantiques ..." d'une lexie, en cliquant dans la liste sur le nom d'un sémème, on peut faire afficher les deux composantes du sémème : le classème et le sémantème. On obtient ainsi (comme le montrent les figures 5.15 et 5.20) les actualisations des sèmes oppositionnels dans la définition du sémème de la lexie. Dans cette fenêtre, on peut aussi lancer la description d'un nouveau sémème (par le bouton "Nouveau Sémème …"). Ceci conduit à décrire un taxème en engageant à nouveau le processus interactif de construction de tables. Lorsque la lexie en question apparaît comme sélection dans une table, le sémème qui en résulte est calculé et son nom vient figurer dans la liste des sémèmes connus pour la lexie. Du point de vue du système, un lexème est considéré comme polysémique dès lors que la liste des sémèmes rattachés à sa lexie contient au moins deux items.

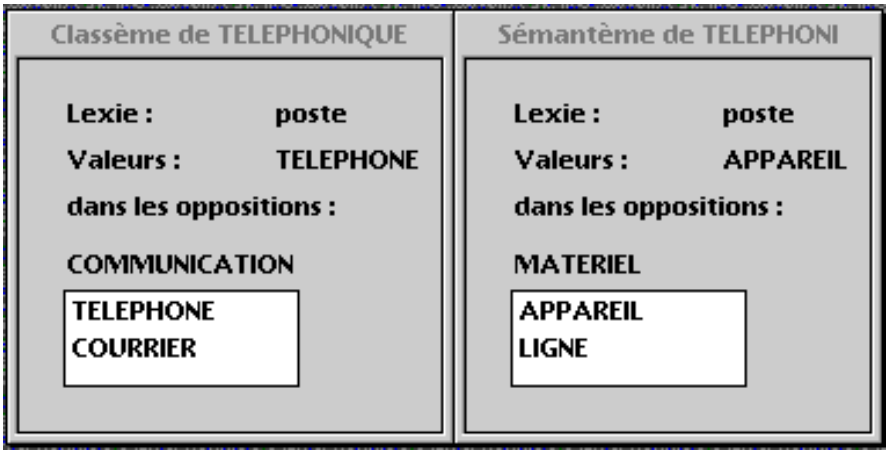


Figure 5.20 : Le sémème TELEPHONIQUE de la lexie *poste*.

2.3. Calcul interprétatif

Une fois formés les contenus des lexies de l'énoncé, on passe à la phase de calcul des isotopies dans la chaîne syntagmatique. On lance ce calcul en actionnant le bouton "Potentiel de sens ..." de la fenêtre d'analyse d'un énoncé (cf. Figure 5.6).

Le calcul des isotopies de l'énoncé est réalisé en quatre étapes successives :

1ère étape : Calcul des isotopies dans les lexies formant des groupes nominaux,
2ème étape : Calcul des isotopies dans les relations actancielle sujet - verbe,
3ème étape : Calcul des isotopies dans les relations actancielle verbe - objet,
4ème étape : Récapitulation et prolongement des isotopies sur l'ensemble de la chaîne.

Ces quatre étapes permettent d'appliquer la règle de réduction de la polysémie lexicale et de calculer son effet sur les isotopies de la chaîne. En effet, si un mot polysémique apparaît dans un groupe nominal et que l'un de ses sémèmes ne donne lieu à aucune isotopie dans le groupe nominal, alors ce sémème n'est pas retenu dans la contextualisation du mot. Il s'ensuit que les sèmes qui le composent ne pourront renforcer des isotopies dans le reste de l'énoncé. Ces quatre étapes sont un moyen de rendre compte des effets de la détermination locale sur le global. Les trois premières étapes du calcul correspondent à la recherche des dépendances sémantiques ; la dernière correspond à la recherche d'influences sémantiques.

Par exemple, dans le groupe nominal *le poste téléphonique* (le GN étant issu de l'analyse syntaxique), le processus calculatoire met en évidence la récurrence du sème [Communication à distance : téléphone *vs*. courrier], et construit ainsi une isotopie. Seul le sémème n°2 de *poste* prend part à cette isotopie. La règle de réduction dans le co-texte de la polysémie lexicale s'applique et consiste à ne garder que le sémème n°2 pour la deuxième occurrence de *poste* dans l'énoncé. Dès lors, dans les phases de recherche d'isotopies (relation actancielles et ensuite ensemble de la chaîne), le sème [Activité professionnelle : statut *vs*. fonction] porté par la première occurrence de *poste* ne donne lieu à aucune isotopie dans l'énoncé (étant donné la levée de l'ambiguïté sémantique de la lexie *poste* dans le groupe nominal *le poste téléphonique*). Une fois de plus, par application de la règle de réduction de la polysémie lexicale, seul le sémème n°2 est gardé pour la première occurrence de *poste*. En procédant successivement à la recherche des isotopies à différents paliers, on a réussi à contextualiser correctement le lexème *poste*. Ceci explique la non-récurrence de [Activité professionnelle : statut *vs*. fonction], alors que ce sème aurait été récurrent si l'on avait cherché les isotopies sur l'ensemble de la chaîne avant d'examiner l'effet des dépendances sémantiques "locales".

Au terme des quatre phases de calcul, la polysémie lexicale est levée et les isotopies trouvées sont le plus possible prolongées dans l'énoncé. Ainsi, dans le traitement sémantique de *Non, poste à l'IUT, c'est le poste téléphonique*, deux isotopies ont été trouvées:

• une mixte, celle du sème [Communication à distance : téléphone *vs*. courrier] (actualisé dans la valeur téléphone) dans les sémèmes 'poste' (première occurrence), 'poste' (seconde occurrence), et 'téléphonique',

• une spécifique, celle du sème [Matériel téléphonique : appareil *vs*. ligne] (actualisé dans la valeur appareil) dans les sémèmes 'poste' (première occurrence) et 'poste' (seconde occurrence).

2.4. Mise en forme des résultats de l'interprétation

Chacune des isotopies permet de mettre en forme une contrainte du potentiel de sens de l'énoncé. Une telle contrainte est exprimée par la récurrence d'une valeur (l'actualisation du sème dans l'isotopie) attachée à un domaine (le domaine d'interprétation du sème dans l'isotopie) dans un rapport d'opposition à une ou plusieurs autres valeurs de ce domaine d'interprétation (le jeu d'opposition est formé par les valeurs actualisables du sème dans l'isotopie). En fonction de la nature du sème formant l'isotopie dans les sémèmes de la chaîne, le degré de généricité de l'isotopie (générique, mixte ou spécifique) est trouvé. Ce degré permet de classer entre elles les contraintes issues des isotopies de la chaîne selon un critère d'importance.

Les deux isotopies précédentes permettent de présenter, par ordre d'importance, les deux contraintes suivantes :

• Il est question de la valeur	*téléphone*
dans le domaine d'interprétation	*Communication à distance*
et dans l'opposition	*téléphone vs. courrier*.

• Il est question de la valeur	*appareil*
dans le domaine d'interprétation	*Matériel téléphonique*
et dans l'opposition	*appareil vs. ligne.*

L'interface du système affiche (cf. Figure 5.21) la liste, triée de la plus générique à la plus spécifique, des contraintes linguistiques de l'énoncé. La sélection dans cette liste d'une contrainte permet d'en afficher les caractéristiques (cf. Figure 5.22).

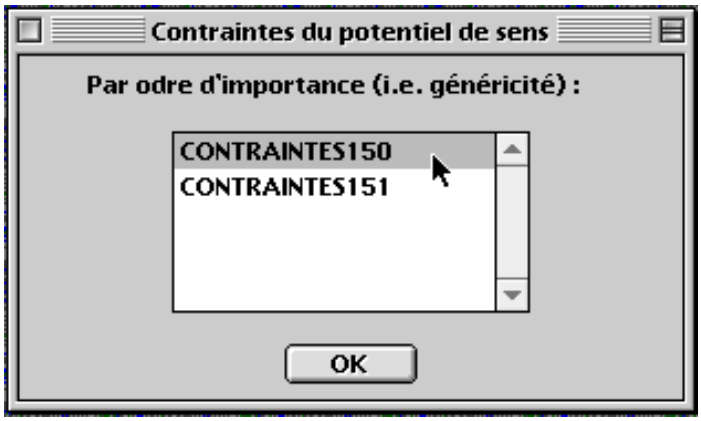


Figure 5.21 : Contraintes du potentiel de sens

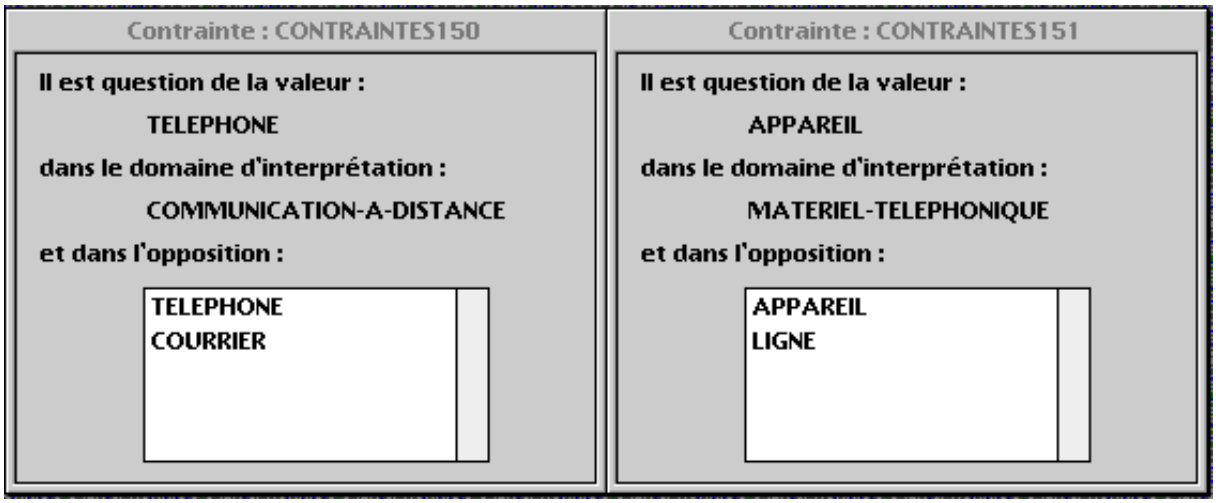


Figure 5.22 : Caractéristiques des contraintes du potentiel de sens

Une fois réalisé le calcul des contraintes, on peut revenir sur l'analyse qui vient d'être menée en redécrivant des connaissances paradigmatiques par un nouvel examen des contenus sémiques des lexies de l'énoncé (par l'intermédiaire de la liste des lexies de la fenêtre d'analyse d'un énoncé et en réactivant ensuite le calcul du potentiel de sens, cf. Figure 5.6). Comme le montrent [Trognon & al. 92], l'interprétation d'un énoncé dans un dialogue peut aussi conduire à revoir l'interprétation des énoncés précédents. Dans ce cas, on pourra relancer l'analyse d'un énoncé précédent en cliquant dessus dans la liste de la fenêtre d'analyse, (cf. Figure 5.5). Par exemple, dans l'interprétation de *Non, poste à l'IUT, c'est le poste téléphonique* est apparue une deuxième signification de *poste* (celle en termes de "poste téléphonique") et on peut se demander si cela change l'interprétation de l'énoncé d'avant *poste à l'IUT, ben, on va mettre vacataire, donc.* On relance alors une deuxième analyse de l'énoncé, où, à la différence de la première, il sera question de lever la polysémie lexicale de *poste* (les résultats de l'interprétation sont identiques car la signification retenue pour *poste* après la levée de la polysémie lexicale est toujours celle en terme de statut professionnel étant donnée l'isotopie avec le lexème *vacataire*).

Lorsque l'analyse interprétative réalisé par l'outil logiciel est jugée satisfaisante, on peut passer à l'analyse de l'énoncé suivant. Les énoncés d'une même séquence ne sont pas analysés de manière indépendante car les isotopies calculées dans l'analyse de l'énoncé précédent sont maintenues en mémoire. Ces isotopies forment les attentes de monstration (i.e. des attentes thématiques) de la machine dans l'interprétation de l'énoncé suivant, assurant ainsi la prise en compte de la continuité des effets de sens des énoncés d'une même séquence. Elle pourront ainsi être renforcées dans l'analyse de la suite de la séquence du corpus (ou remise en cause s'il y avait malentendu), renforçant ainsi les contraintes linguistiques des énoncés.

3. Une expérimentation de l'outil logiciel

L'implémentation que nous venons de détailler permet une expérimentation logicielle de nos principales hypothèses, comme par exemple celle qui consiste à formuler les descriptions des contenus sémantiques en termes de sèmes oppositionnels. Cette expérimentation sur des séquences du corpus PIC permet également l'acquisition, dans et par l'interaction homme - machine, de premiers systèmes hiérarchiques et différentiels de significations qui sont validés par les analyses des énoncés[79]. On met ainsi en place, grâce à ces représentations, une amorce d'une compétence interprétative. Dans cette partie, nous allons décrire un système hiérarchique différentiel de significations produit avec l'outil logiciel proposé.

Reprenons les tables construites dans le chapitre 3 à partir d'analyses du corpus PIC pour décrire les outils d'aide à l'utilisation d'un logiciel. Ces tables étaient au nombre de deux : la première décrivant le taxème de la documentation technique dans lequel sont définis différentiellement les lexies *manuel, référentiel* et *aides en ligne* ; la seconde (subordonnée à la précédente) décrivant le taxème des aides en ligne où sont définis différentiellement les lexies *assistant, full text, sommaire* et *index*.

//documentation//

	Support de l'information	Présentation de l'information
manuel	papier	globale
référentiel	papier	précise
	logiciel	globale
aides en ligne	logiciel	précise

[79] L'analyse du corpus PIC est toujours en cours pour expérimenter le modèle. Pour cela, nous comptons reprendre les résultats des travaux de Karine Taillebois, étudiante en thèse de psychologie développementale, dont une partie du travail consiste en une analyse des ambiguïtés dans le corpus.

//aides en ligne//		
	Accès à l'information	Structure de l'information
assistant	interactif	hiérarchique
full text	interactif	linéaire
sommaire	figé	hiérarchique
index	figé	linéaire

Nous allons construire maintenant un système hiérarchique plus conséquent en mettant en place un arbre dont ces deux tables sont des noeuds et où la table //aides en ligne// est une feuille. C'est-à-dire qu'en construisant cet arbre, nous allons décrire des niveaux sémantiques plus généraux que les aides en ligne et la documentation. Il en résultera que les classèmes des lexies décrites dans ces deux tables seront plus détaillés, car de ces niveaux plus généraux seront issus des sèmes génériques. Pour cela, nous avons à former une table dans laquelle soit définie la lexie *documentation* de telle sorte que la table //documentation// en soit une sous-catégorisation. Considérons le point de vue suivant : la documentation est ce qui permet d'apprendre à utiliser un logiciel. Selon ce point de vue, on a deux types de rapport avec un logiciel conditionnant deux oppositions : soit apprendre quelque chose *versus* pratiquer quelque chose, soit utiliser quelque chose *versus* concevoir quelque chose. Ainsi, un usager est dans une pratique d'utilisation d'un logiciel. De même, un programmeur est dans une pratique de conception de logiciel, par opposition aux étudiants en informatique qui sont dans une pratique d'apprentissage de conception de logiciels. On définit différentiellement de cette façon les lexies *application, documentation, programmation* et *formation* dans une table qui est un certain point de vue sur les logiciels.

//logiciels//		
	Rapport au logiciel	Pratique du logiciel
application	utilisation	pratiquer
documentation	utilisation	apprendre
programmation	conception	pratiquer
formation	conception	apprendre

De même que de cette table provient par sous-catégorisation la table //documentation//, on peut construire les tables suivantes :

• la table des formations permettant de définir différentiellement les lexies *technicien, ingénieur* et *docteur*.

//formation//	
	Durée de la formation
technicien	de 2 à 4 ans
ingénieur	en 5 ans
docteur	en plus de 5 ans

• la table des langages de programmation permettant de définir les lexies *Common Lisp, Caml, C++,*et *Pascal* dans la combinatoire des différences : programmation fonctionnelle *vs*. programmation impérative et programmation structurée *vs*. programmation par objet.

//programmation//		
	Concept de programmation	Type de programmation
Common Lisp	fonctionnelle	objet
Caml	fonctionnelle	structurée
C++	impérative	objet
Pascal	impérative	structurée

• une table exprimant un premier point de vue sur les applications en fonction de leur utilité, c'est-à-dire par exemple gérer, communiquer, simuler ou jouer.

//application-1//	
	Utilité de l'application
SGBD	gérer
e-mail	communiquer
système expert	simuler
jeu	jouer

• une table exprimant un second point de vue sur les applications, à savoir que ce que l'on voit d'une application c'est son interface, par opposition aux traitements qu'elle réalise qui ne sont pas forcément visibles à l'écran. Selon ce point de vue, on peut également définir par sous-catégorisation les éléments d'interface d'une application qui sont verbalisés dans le corpus PIC (info-bulle, onglet, icône, menu) en dressant la combinatoire des deux oppositions suivantes : soit l'élément d'interface amène une information, soit il permet de réaliser une action et soit l'élément concerne une chose bien précise, soit il en concerne plusieurs.

//application-2//	
	Affichage à l'écran
interface	visible
traitements	caché

//interface//		
	Fonction de l'élément	Objet de l'élément
info-bulle	informationnelle	une chose précise
onglet	informationnelle	plusieurs choses
icône	actionnelle	une chose précise
menu	actionnelle	plusieurs choses

Avec la table des logiciels, nous avons construit un niveau générique supplémentaire par rapport aux deux tables de départ. On peut répéter cette opération en formant une table où est définie la lexie *logiciels*. On forme, par exemple, la table suivante exprimant un point de vue sur les composants des systèmes informatiques (*logiciels, cartes mères, systèmes d'exploitation* et *périphériques*) par la combinatoire des différences : le type du composant est matériel ou pas et il essentiel ou accessoire.

//informatique//		
	Type du composant	Rôle du composant
systèmes d'exploitation	non matériel	essentiel
logiciels	non matériel	accessoire
cartes mère	matériel	essentiel
périphériques	matériel	accessoire

Comme on a détaillé deux points de vue sur les applications, on peut construire à partir de cette table un nouveau point de vue sur les logiciels pour représenter les rôles des participants lors de l'expérience PIC : *rédacteur*, *utilisateur*, *développeur*. On les définit dans la table suivante en opposant les fonctions possibles d'une personne par rapport à un logiciel.

//logiciels-2//	
	Fonction de la personne
rédacteur	faire la documentation
utilisateur	utiliser
développeur	programmer
chef de projet	gérer la conception
agent commercial	vendre

Enfin, on peut encore une fois répéter le processus de généralisation en définissant la table informatique comme une sous-catégorisation de la lexie *informatique*, définie par opposition avec les *télécommunications* dans le taxème des technologies de l'information.

//technologies de l'information//	
	Utilisation de l'information
informatique	manipuler
télécommunications	diffuser

Le système hiérarchique différentiel ainsi créé (cf. Figure 5.23) consiste donc en un arbre de 11 tables, c'est-à-dire 11 taxèmes, donnant lieu à la mise en forme des contenus sémantiques de 39 lexies.

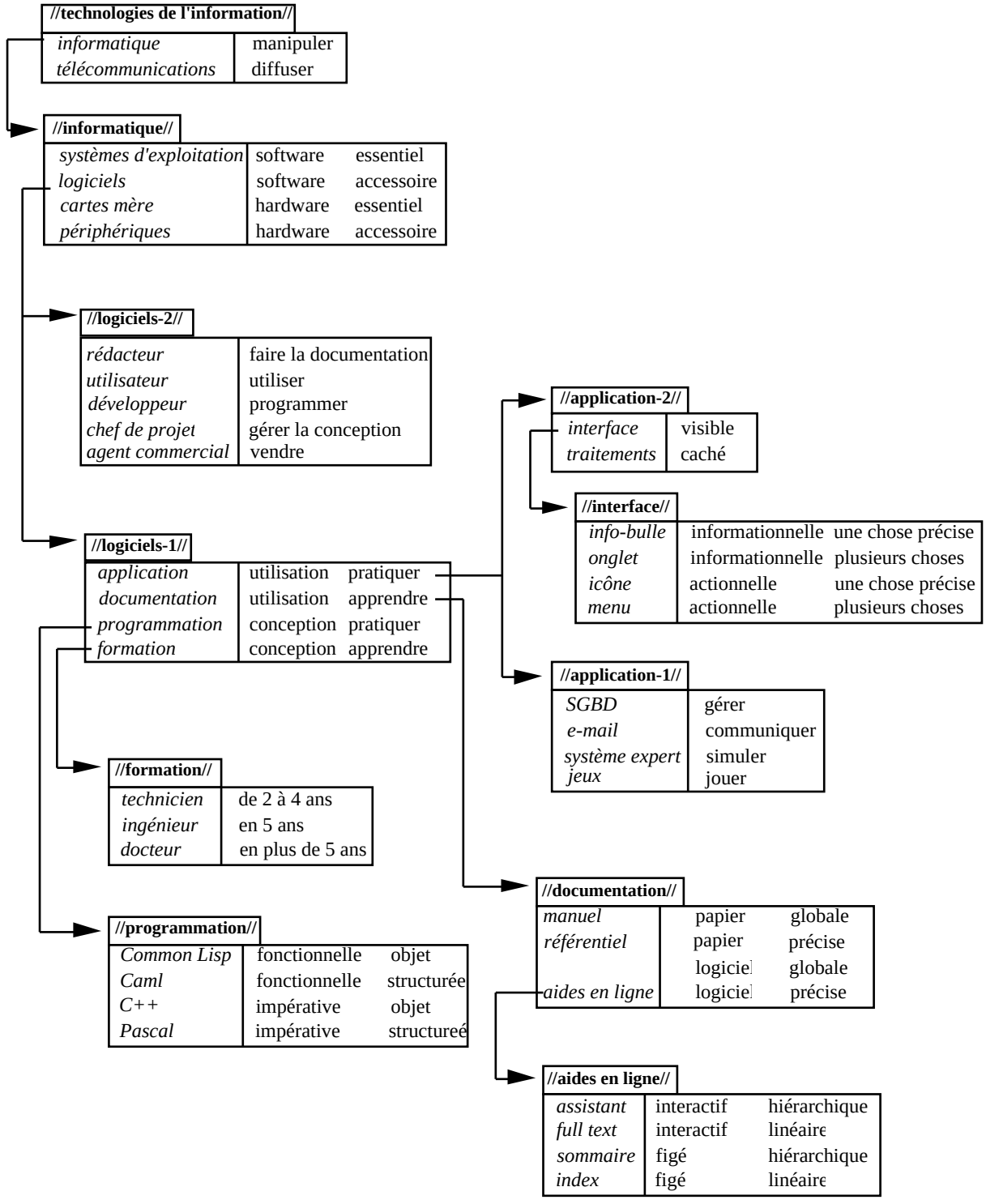


Figure 5.23 : Un système hiérarchique différentiel de significations

Les taxèmes sont indiqués entre doubles // et les lexies sont indiquées en italique.

Dans l'activité de construction d'arbres de tables, l'outil logiciel permet à un agent humain de préciser un point de vue sur ce qui est discuté dans le dialogue, mais l'intérêt principal de cette l'activité réside dans l'acquisition de connaissances pour la machine. Cette acquisition

consiste en la construction d'une mémoire sémantique lui permettant d'inférer des représentations componentielles pour la signification des mots. Par exemple, du système hiérarchique de signification présenté ici, la machine peut extraire un nouveau sémème pour *assistant* comportant maintenant 9 sèmes (cf. Figure 5.24).

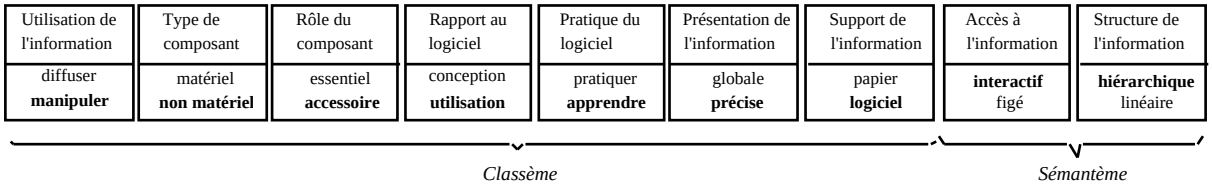


Figure 5.24 : Le sémème 'assistant'

Comme on l'a expliqué précédemment, l'acquisition des contenus sémantiques conditionne les résultats opératoires de l'analyse interprétative. Ainsi, alors qu'en considérant uniquement les deux tables //documentation// et //aides en ligne// on obtenait 2 contraintes sémantiques (cf. chapitre 4, § 3.), avec le système hiérarchique différentiel décrit le calcul interprétatif de l'énoncé

> *... vous connaissez les outils dans les **aides en ligne**, les **assistants** …*
> (énoncé du rédacteur à l'intention de l'utilisateur, à la 20ème minute)

mène maintenant à la mise en forme de 7 contraintes sémantiques, dues à la répétition des sèmes du sémème 'aides en ligne' dans le classème de 'assistant'.

4. Conclusions

Dans ce cinquième chapitre, nous avons proposé une articulation des modèles sémantiques de l'axe paradigmatique et de l'axe syntagmatique présentés dans les chapitres précédents. La réalisation logicielle proposée opérationnalise cette articulation au sein d'une interaction entre l'homme et la machine. L'enjeu de cette interaction est l'explicitation de l'interprétation de l'humain dans un but d'acquisition de connaissances sémantiques pour la machine. Au fur et à mesure que l'on utilise l'outil logiciel, sa base de connaissances (les systèmes de tables différentielles) augmente, il met ainsi en place un niveau embryonnaire de compétence interprétative. Dans la construction informatique de cette compétence, nous sommes encore loin d'un analogue car notamment les dimensions du sens autres que la monstration (la modalité, la prédication et l'argumentation) restent à modéliser de façon opératoire.

La première expérimentation a permis d'identifier quelques défauts de l'outil logiciel qui amènent à repréciser certains aspects calculatoires, notamment en ce qui concerne les rapports entre la syntaxe et la sémantique. Pour l'identification des lexies de l'énoncé, nous sommes jusqu'ici tributaires du découpage morphologique réalisé par l'analyseur syntaxique que nous avons utilisé. Du fait de ce découpage, le syntagme *"pomme de terre"*, par exemple, donnera lieu à une chaîne de trois lexies. On peut effectivement considérer que c'est le cas dans l'énoncé *Il a sorti cette pomme de terre* mais, dans la majeure partie des énoncés, le syntagme ne représente qu'une unique lexie (comme dans *À midi, il a mangé de la pomme de terre*). Cet exemple très connu montre bien que le découpage morphologique, qui conditionne la reconnaissance des lexies, n'est ni un problème purement sémantique, ni un problème purement syntaxique. Du point de vue de l'implémentation d'une analyse strictement sémantique, la réification d'une lexie dont la chaîne de caractères comporte plusieurs mots (séparés par des espaces) ne conduit à aucun problème calculatoire (et les éléments de la sélection des tables Anadia peuvent sans problèmes comporter des espaces). Cependant, si l'on reconnaît les lexies au niveau sémantique (soit manuellement par sélection d'une chaîne de caractère dans l'énoncé, soit automatiquement par interrogation des sélections des tables), cela veut dire que l'on ne peut plus tenir compte des dépendances syntagmatiques issues de l'analyseur syntaxique (constitution des groupes nominaux et relations actancielles), étant donné que l'énoncé d'un point de vue syntaxique et d'un point de vue sémantique, n'est plus formé des mêmes constituants. La solution consiste donc à faire appel à des connaissances sémantiques dans le découpage lexical de l'analyse syntaxique, c'est-à-dire concevoir un nouveau module d'étiquetage syntaxique interfacé avec la base de connaissances sémantiques. Ceci sort du cadre de cette thèse et c'est pour cette raison que nous avons préféré faire appel à un analyseur déjà existant, que nous n'aurions pas réussi à égaler qualitativement dans le temps dont nous disposions. Nous en acceptons donc ici les contraintes, tout en reconnaissant un manque pour ce qui concerne une validation de l'outil logiciel.

L'expérimentation reste à poursuivre car les systèmes différentiels de significations construits sont propres au corpus étudié, c'est-à-dire à un certain type de conversation (en l'occurrence l'utilisation d'un logiciel dans un contexte professionnel). Il convient donc d'expérimenter l'outil logiciel avec d'autres corpus de dialogue où l'enjeu des conversations n'est plus le même pour examiner comment les connaissances du système augmentent et comment celles qui sont déjà acquises peuvent être réutilisées dans d'autres contextes de manière fructueuse.

CHAPITRE 6 :

VALIDATION QUALITATIVE ET APPORTS À L'INFORMATIQUE

L'objectif de ce dernier chapitre est une valorisation informatique du modèle proposé dans cette thèse. Il s'agit de donner des éléments de validation du modèle et de la réalisation logicielle qui en est issue. Étant donné que ce modèle est construit sur un principe d'apprentissage interactif par amorce lié à une situation et à un certain type de corpus, la validation ne peut être quantitative en appliquant l'outil réalisé à d'autres corpus de type différents. Nous proposerons donc ici une validation qualitative en montrant comment les connaissances sémantiques acquises constituent des systèmes pertinents et efficaces en termes de contraintes opératoires.

Dans la première partie du chapitre, nous discuterons les aspects qualitatifs de la représentation des connaissances sémantiques sous la forme de systèmes hiérarchiques et différentiels de significations. Dans une seconde partie, nous ferons le point sur l'apport de notre modèle de l'interprétation pour la réalisation d'agents logiciels dialoguant en langue naturelle. Dans une troisième partie, nous aborderons l'apport de notre travail à d'autres domaines du traitement automatique des langues que le dialogue homme-machine. En cela, nous montrons la généricité du modèle informatique proposé.

1. Aspects qualitatifs des systèmes hiérarchiques différentiels

Un système hiérarchique différentiel (SHD) n'a pas la prétention de fixer de la meilleure façon qu'il soit une connaissance objective partagée par le plus grand nombre. Au contraire, il convient de rappeler que les représentations mises en place avec l'outil logiciel proposé sont propres à un certain type de dialogue étudié, qu'elles sont situées dans une interaction entre un partenaire humain et une machine et qu'elles ne sont justifiées que par le point de vue de l'usager relativement à son projet (le projet étant ici de rendre compte du dialogue observé). Il s'agit de donner à la machine des représentations liées à une dynamique d'apprentissage et d'interaction, et non des représentations fixant des significations de façon universelle. Ceci a pour conséquence que, dans un SHD, tout peut être discuté à tout moment par chacun. Par exemple, dans le SHD détaillé dans la dernière partie du chapitre précédent, on pourrait très bien contester le fait d'opposer dans une même table (le table du taxème //logiciels-1//) les significations de *application* et de *programmation* en disant que l'une est le moyen de l'autre (et vice versa) et donc qu'elles ne relèvent pas d'un même niveau. En discutant les aspects qualitatifs de ce genre de représentations, nous ne chercherons donc pas à défendre le point de vue sous-jacent mais plutôt à montrer ce que celles-ci apportent à la personne qui les met en place et surtout, dans une démarche informatique, à mettre en évidence leurs propriétés systémiques et opératoires. Celles-ci permettent d'assouplir les interactions homme - machine en les adaptant au projet de l'usager.

Si l'outil informatique réalisé est simple conceptuellement, la construction avec cet outil d'un système hiérarchique différentiel de signification est le résultat d'une activité complexe de la part d'un agent humain. Cette activité résulte d'une alternance entre trois processus : un processus d'observation par l'interprétation, un processus de généralisation et un processus de spécialisation.

• Le processus d'observation par l'interprétation : c'est le processus qui consiste à appréhender une séquence de dialogue comme un tout témoignant d'une cohésion et d'une cohérence. En observant les énoncés du dialogue, l'agent humain est amené à rendre explicites des parcours interprétatifs en isolant des signes produits dans l'interaction, en repérant leur justification (notamment en recherchant si des oppositions entre ces signes sont verbalisées dans le corpus) et en cherchant à construire des classes sémantiques où ces signes se définissent mutuellement. Ainsi, en rapport à notre exemple, l'observation par l'interprétation de la séquence étudiée à permis de repérer un certain nombre de lexies importantes (*manuel, référentiel, aides en ligne, assistant, full text, sommaire, documentation, info-bulle, onglet, icône* et *menu*) ainsi que certaines oppositions entre elles (celles ayant permis la construction des tables //documentation// et //aides en ligne//).

À l'issue de ce processus, les premières tables sont formées sans forcément être liées les unes aux autres (par exemple, la table du taxème //documentation// est liée à celle de //aides en ligne// mais pas à celle où sont définis les éléments d'interface *info-bulle, onglet, icône* et *menu*). Ce sont ensuite les processus de généralisation et de spécialisation qui vont permettre de connecter ces tables dans une même structure hiérarchique formant une base de connaissances conséquente pour la machine.

• Le processus de généralisation : c'est le processus qui consiste à former, à partir d'un taxème, la table dans laquelle ce taxème apparaît en tant que lexie ; c'est-à-dire former, à partir d'une table A, une table B dont A est un fils dans l'arbre de tables. Dans le processus de généralisation, on construit l'arbre en montant vers sa racine. C'est, par exemple, par un processus de généralisation que l'on a formé la table //logiciels-1// en partant de la table //documentation//.

• Le processus de spécialisation : c'est le processus dual de la généralisation. C'est-à-dire qu'à partir d'une table A, on va former une table B dont A est le père dans l'arbre. La spécialisation consiste à donner à une lexie figurant dans une table le statut de taxème en construisant une nouvelle table par sous-catégorisation. Dans le processus de spécialisation, on construit l'arbre en descendant vers ses feuilles. C'est, par exemple, par un processus de spécialisation qu'on a sous-catégorisé la lexie *programmation* dans la table //logiciels-1// pour former la table //programmation//. Ainsi, en ce qui concerne la mise en forme des contenus sémantiques, du processus de généralisation provient la construction les classèmes, tandis que du processus de spécialisation provient principalement la construction des sémantèmes.

Outre le retour du SHD sur les résultats de l'analyse interprétative (cf. chapitre 5), la valeur qualitative de ce mode de représentation sémantique tient à la simplicité et à l'efficacité de sa mise en œuvre informatique. À la différence des réseaux sémantiques massivement utilisés dans les développements d'intelligence artificielle (par exemple dans le dictionnaire *Dicologique*™ de la société Mémodata[80] ou encore dans le système Caramel [Sabah & al. 93]), les systèmes hiérarchiques différentiels sont des représentations "légères" au niveau algorithmique et opérationnel. Cette légèreté vient du fait qu'en nous rattachant aux principes de la sémantique structurale, nous avons pris le parti, dans notre modèle, de séparer la question du sens de celle de l'ontologie. Il ne s'agit donc pas, à l'inverse des réseaux sémantiques, de constituer des classifications ontologiques exhaustives. À titre d'exemple des problèmes opératoires posés par les réseaux sémantiques (problèmes d'occupation en mémoire et/ou problèmes de temps de traitements par exemple), l'utilisation de *Dicologique*™ comme base de

[80] Pour de plus amples détails voir *http://www.memodata.com*.

connaissances demande une compilation du dictionnaire sous forme d'un graphe par un algorithme de fermeture transitive du graphe, et cette compilation dure plusieurs heures. C'est parce que les systèmes hiérarchiques différentiels n'ont pas pour objectif d'être des représentations de connaissances objectives complètes qu'il ne sont pas tenus d'être préalablement fondés sur un modèle ontologique tacitement partagé. Ainsi, les connaissances que notre outil logiciel permet de représenter sont des connaissances suffisantes pour l'analyse interprétative d'une situation de dialogue et les questions consistant à savoir si ces connaissances sont complètes et nécessaires ne se posent pas.

Pour la même raison, la question d'une cohérence logique du système de significations par rapport à des représentations conceptuelles du monde n'est pas non plus posée. Par exemple, dans le système hiérarchique différentiel présenté au chapitre précédent, la signification de *e-mail* dans la table //application-1// est déterminée par la valeur *communiquer* dans le domaine d'interprétation Utilité de l'application alors que cette table est subordonnée à la table //informatique// provenant de la définition de *informatique* par opposition à *télécommunications* dans la table //technologies de l'information//. On pourrait donc voir ici une incohérence logique entre quelque chose qui est défini pour communiquer et qui n'est pas un moyen de télécommunication. De ce point de vue, le système serait effectivement logiquement incohérent s'il était entendu au préalable que, lorsqu'une table A est subordonnée à une table B, cela doit signifier que les entrées de A sont des éléments de même type que les entrées de B. Ce n'est pas le cas car la relation entre deux tables subordonnées l'une à l'autre, c'est-à-dire en définitive la relation d'appartenance d'une lexie à un taxème, n'est pas une relation conceptuelle du type EST-UN comme dans les réseaux sémantiques. C'est là une différence de fond entre les deux systèmes de représentation sémantique. La relation représentée par les branches d'un arbre de tables différentielles indique qu'une table B, qui est subordonnée à une table A, exprime un certain point de vue sur A. Par exemple, le taxème //formation//, définissant des niveaux de formation en informatique, est subordonné au taxème //logiciel-1// parce qu'il provient d'un certain point de vue sur les logiciels, mais il ne s'agit pas de représenter ici que les formations sont des types de logiciels.

Un dernier aspect qualitatif des systèmes hiérarchiques différentiels que nous mettons en évidence concerne leur caractère évolutif intrinsèque. Dans notre modèle, les connaissances de la machine sont partielles car elles sont liées à un projet et constituées de façon minimales comme base d'un processus interactif incrémental de complémentation et de restructuration des tables. C'est parce que les systèmes hiérarchiques différentiels sont construits sur le modèle Anadia, dont l'intérêt tient plus au processus mis en place qu'aux résultats de ce processus (cf. chapitre 3), qu'ils ne peuvent être considérés comme des bases de connaissances stabilisées et objectivées à l'égal des réseaux sémantiques. Plutôt que d'envisager cet aspect comme un défaut

pour la construction en informatique d'une mémoire sémantique conséquente, nous considérons, à l'inverse, que la mémoire n'est pas une structure statique de stockage mais un système dynamique d'apprentissage. Ainsi, nous défendons qu'Anadia constitue un modèle informatique pertinent pour la modélisation en intelligence artificielle de systèmes mémoriels dynamiques intégrant une structure et un protocole de révision perpétuelle de cette structure. Chercher à automatiser, au moins en partie, ce protocole de révision demande la modélisation de processus de recatégorisation, de réorganisation des tables et également d'oublis lorsque des tables ne sont plus utilisées. C'est ainsi, en automatisant des processus de méta-contrôle, qu'un outil logiciel développé avec Anadia pourra acquérir un réel statut d'agent (un agent aide-mémoire par exemple) ayant un comportement autonome du point de vue de ses représentations internes. Ceci constitue une autre thématique de recherche que la nôtre, dans laquelle Delépine ([Delépine & al. 98]), par exemple, utilise Anadia comme modèle de représentation interne pour un agent d'aide à la remémoration de documents dans un contexte d'assistance à la navigation dans un hypertexte. Dans la réalisation informatique proposée ici, le protocole de révision est resté à la charge de l'humain par la mise en place d'un processus interactif initié par une situation de recherche de consensus entre son interprétation des dialogues étudiés et les résultats obtenus par la machine.

2. Apports dans le domaine du dialogue homme - machine

Avec cette thèse, nous voulions contribuer à la construction d'un modèle opératoire de l'interprétation dans une problématique de conception d'agents logiciels capables de dialoguer avec des humains en langue naturelle. Nous allons ici préciser l'apport du modèle à la réalisation d'un tel agent.

L'apport principal de notre travail dans le domaine du dialogue homme-machine est de proposer un module d'interprétation des énoncés. Ce module coopère avec un modèle de gestion du dialogue dans la tâche en cours entre la machine et l'humain. Le module de l'interprétation prend en charge le partage des significations dans l'interaction et il fournit au modèle de dialogue des contraintes sémantiques issues de l'analyse des énoncés. C'est alors dans la gestion de la tâche que les résultats issus de l'interprétation prendront réellement leur statut de contraintes en tant qu'elles produiront des connaissances qui doivent être compatibles avec l'avancement de cette tâche. C'est donc dans et par la tâche que sont opérationnalisées les contraintes. Par exemple, dans le cadre d'une recherche documentaire, l'interprétation de l'énoncé suivant

Qui a écrit le livre intitulé "Les fleurs du mal" ?

signale

• qu'il est question d'une activité humaine (ce qui est marqué par la co-détermination du lexème *qui* et du verbe *écrire*),

• qu'il est question de livre (ce qui marqué par le verbe *écrire*, le lexème *livre* et le lexème *intitulé*),

• qu'il est probable que *"Les fleurs du mal"* soit un titre (étant donné que c'est une lexie postposée à *intitulé*),

• et qu'il s'agit d'une requête (ce qui est marquée par la forme interrogative).

C'est au modèle de gestion du dialogue qu'il incombe d'actualiser, dans le contexte de la tâche, les résultats de l'interprétation ; c'est-à-dire d'utiliser ces informations pour inférer qu'une requête sur l'activité humaine liée à l'écriture d'un livre implique une recherche sur le champ "Auteur" de la base de données et de lancer la recherche adéquate dans la base pour fournir la réponse *Baudelaire*. Dans la construction d'une compétence dialogique artificielle, le modèle du dialogue prend donc en compte l'actualisation pragmatique des contraintes issues de l'interprétation ainsi que la gestion de la tâche. La modélisation de cette actualisation pour un système de dialogue homme - machine est l'objectif principal de la thèse de Gérard ([Gérard & al. 98]).

Comme on l'a vu (cf. Chapitre 1), cette actualisation pragmatique du discours dans une tâche amène à modéliser les spécificités du contexte et des interlocuteurs et à poser le problème de la référence. Il faut alors donner au modèle de gestion du dialogue un modèle extralinguistique de l'objet référent (dans [Beust & al. 97a], nous en avons proposé les bases dans un modèle fondé sur l'identité de l'objet), par lequel il pourra exprimer les contraintes sémantiques (principalement celles de monstration) de l'énoncé en tant que connaissances conceptuelles sur de possibles objets qu'il conviendrait de rechercher dans une représentation du contexte et de la situation d'énonciation. Une telle démarche, selon Bachimont, passe par une normalisation des signifiés sans laquelle une analyse sémantique reste uniquement descriptive et ne constitue pas de connaissances.

> Pour déterminer quelles connaissances sont exprimées par des énoncés linguistiques, il est donc nécessaire de leur associer des prescriptions interprétatives permettant de déterminer le contenu conceptuel qu'ils véhiculent, c'est-à-dire les connaissances qu'ils expriment. Il faut donc normaliser la signification des expressions pour qu'il soit possible de leur associer un contenu conceptuel même hors contexte. La normalisation repose sur le constat que le signifié linguistique n'est pas conceptuel et qu'il faut comprendre le concept comme un *signifié normé*. ([Bachimont 96, p. 138])

Le but d'une modélisation de l'actualisation du sens, à partir d'un modèle sémantique de l'interprétation, consistera finalement à dépasser le statut descriptif des contraintes sémantiques. Cette actualisation aura pour fin de préciser les aspects référentiels du sens ainsi que les aspects

actionnels (performatifs) de la compréhension par le choix dans l'enchaînement conversationnel d'un pattern d'action approprié ([St Dizier De Almeida 96]). La façon dont il conviendra de normer le signifié dépendra de la tâche en jeu dans le dialogue. Ce n'est que dans le rapport à une tâche qu'une contrainte sémantique pourra permettre d'exprimer des connaissances ; ce que [Lehuen 97] (qui envisage un dialogue finalisé pour une tâche de recherche dans une médiathèque) appelle des *granules de compréhension.*

L'outil logiciel réalisé dans cette thèse met en place un modèle opératoire où l'interprétation d'énoncés de dialogue est supervisée par un agent humain dans un processus interactif. L'interactivité est un premier pas dans la modélisation informatique d'une activité dialogique. Dans un dialogue entre deux humains, lorsque l'un des partenaires ne connaît pas la signification d'un mot et qu'il ne peut la déduire du co-texte par des opérations interprétatives, il engage un dialogue incident pour que son interlocuteur lui apporte les contenus sémantiques qui lui manquent. C'est alors que les reformulations, les paraphrases ou encore les métaphores sont d'une grande utilité dans la construction du terrain commun, et c'est ainsi que, selon [Vivier 93], le dialogue construit le dialogue. Du point de vue de l'apprentissage d'un agent logiciel, cela pose problème parce que le dialogue homme-machine est alors autant un but à atteindre qu'un moyen pour atteindre ce but. Le processus mis en place a permis d'éviter cette circularité, car il prend en charge, de façon interactive, les situations d'explicitation qui, dans un dialogue homme-homme, consisteraient en des dialogues incidents. De ce fait, un modèle de l'interprétation interactif est une première approche dans la modélisation de l'activité dialogique.

Dans la modélisation informatique du dialogue homme-machine, la gestion de l'interaction en termes de dialogues régissants ou incidents est primordiale, comme le montre le modèle dynamique de Luzzati ([Luzzati 92]). Cette gestion est étroitement liée à l'interprétation car détecter un changement d'axe dans le dialogue (un passage de l'axe régissant à l'axe incident ou le contraire), ou bien une poursuite de l'axe en cours, est un problème qui doit prendre en compte le rapport entre ce qui est dit dans l'interaction et l'avancement de la tâche. De plus, avec un modèle de dialogue conçu sur un principe d'amorce où l'agent logiciel entre dans le dialogue avec presque aucune connaissance, la dissymétrie entre les connaissances de l'agent et celles de son partenaire humain rend les dialogues incidents très fréquents et mêmes s'imbriquant souvent les uns dans les autres. C'est bien ce que montrent les expérimentations du système Coala de Lehuen ([Lehuen 97]). Plus le processus d'interprétation est fin, meilleure est donc la gestion de ce problème omniprésent dans le dialogue. La contribution de cette thèse à la problématique du dialogue homme-machine est d'apporter un modèle opératoire de l'interprétation plus complet (et linguistiquement fondé) que les modèles où l'interprétation d'un énoncé résulte d'un jeu d'unifications avec des filtres syntaxico-sémantiques représentés par des phrases à trous (c'est le cas notamment dans le système Coala). Par exemple, Coala déclenche un dialogue incident

lorsque l'unification de l'énoncé traité avec le jeu de filtres a échoué, mais lorsqu'une unification est réussie, il conserve un schéma régissant. Pourtant, si la thématique du discours a changé du tout au tout (dans le cas d'un discours qui saute du coq à l'âne) bien que l'unification ait réussi, il convient d'engager un dialogue incident pour discuter ce changement de thématique.

Considérons, par exemple, un système de consultation de bibliothèque du type Coala qui disposerait des filtres suivants :

Filtre n°1 : *je voudrais le livre intitulé* <TITRE>
Filtre n°2 : *je voudrais le livre* <TITRE>
Filtre n°3 : *je voudrais* <TITRE> *de* <AUTEUR>
Filtre n°4 : *je voudrais* <TITRE> *d'* <AUTEUR>

L'interprétation de l'énoncé

je voudrais le livre "Les fleurs du mal"

va déclencher, par l'unification du filtre n°2 dans l'axe régissant, la recherche d'un livre en posant l'hypothèse que "Les fleurs du mal" est un titre. De même, si l'énoncé qui suit est

je voudrais aussi boire un verre d'eau

par application du filtre n°4 dans l'axe régissant, une recherche va être déclenchée en supposant que *aussi boire un verre* est un titre et que *eau* est un auteur. Avec le modèle de l'interprétation que nous proposons, la gestion du dialogue pourrait être différente. Le premier énoncé conduit à extraire la contrainte sémantique

• Il est question d'un livre (c'est une contrainte sémantique marquée par l'utilisation conjointe du déterminant *un*, et des lexèmes *livre* et *intitulé*).

Le deuxième énoncé quant à lui permet de produire la contrainte sémantique

• Il est question de consommer un liquide (cette contrainte vient de l'utilisation conjointe du verbe *boire* et des lexèmes *verre* et *eau*)

Cette deuxième contrainte exprime notamment qu'il est sémantiquement inconcevable qu'il soit question dans le deuxième énoncé de la recherche d'un livre ; ce qui vient en contradiction avec la tâche en cours. Le changement de thématique entre les deux énoncés est donc un indice pour engager un dialogue incident et permettre à l'agent de faire une réponse du type *je suis désolé, je ne peux vous aider que dans le cadre d'une recherche de livre*. De plus, l'application dans cet exemple du filtre n°4 pour le deuxième énoncé traduit la non prise en compte d'une contrainte sémantique de modalité. En effet, *vouloir* utilisé avec un autre verbe (en l'occurrence *boire*) a une valeur modale, ce qui n'est pas forcément le cas s'il est utilisé seul (en l'occurrence *vouloir* dans le premier énoncé ne marque pas une modalité mais un désir). Dans sa valeur modale, *vouloir* s'oppose à *pouvoir*, tandis que dans sa valeur qui marque un désir, *vouloir* s'oppose à

avoir ; les deux cas ne doivent donc pas être traités de façon similaire. Ceci montre bien les limites d'un calcul interprétatif réalisé à l'aide d'un jeu de filtres syntaxico-sémantiques. Avec le modèle opératoire de l'interprétation que nous proposons, nous défendons que pour mettre en place une meilleure gestion du dialogue, il convient de rechercher à exploiter des indices syntaxiques, sémantiques et thématiques plutôt qu'uniquement des indices liés à la forme syntaxique des énoncés.

Un dernier apport possible d'une meilleure modélisation de l'interprétation dans la problématique du dialogue homme-machine concerne le problème dual de la génération des énoncés de la machine. Coursil ([Coursil 92]) considère que la parole, dans sa réalisation, n'est pas préméditée, car ce qui est prémédité, c'est ce que l'on veut dire, mais ce n'est pas la façon dont le dit en situation. Ceci traduit la raison pour laquelle le fait de s'entendre parler lorsque que l'on profère un énoncé modifie la façon dont on le construit en temps réel. C'est une singularité de la parole car, à l'écrit, une préméditation de la forme de phrase à produire est nécessaire afin de l'écrire. Cette construction de l'énoncé au moment où il est proféré fait intervenir le processus d'interprétation dans la mesure où, en s'entendant parler, on interprète ce que l'on dit. La forme de la suite de la production verbale provient alors d'un équilibre entre le but de l'énonciation (ce que l'on veut dire a priori) et l'interprétation en direct de ce qui vient d'être dit. Ainsi, on pourrait faire l'hypothèse que cette auto-interprétation conditionne la suite de la production verbale en cherchant notamment à prolonger les isotopies déjà initiées dans le début de l'énoncé pour appuyer sa cohésion et sa cohérence et renforcer ainsi son intelligibilité dans le respect des maximes conversationnelles de Grice. Ceci n'est pas en soi une modélisation de la génération, mais c'est une hypothèse opératoire qu'il conviendrait d'expérimenter dans une problématique de conception d'un modèle computationnel de la production verbale.

Notre travail sur la représentation en machine de systèmes de significations et sur le calcul interprétatif des énoncés en langue naturelle nous amène à repenser le modèle d'agent pour le dialogue homme-machine proposé par Lehuen. Lehuen distingue 3 sous-modèles : un modèle du dialogue, un modèle de la tâche et un modèle de la langue. En ce qui concerne la place de l'interprétation, nous argumentons pour ne plus la considérer comme une compétence uniquement invoquée pour l'analyse des énoncés du partenaire dans le dialogue. Elle est, par ses relations avec la représentation lexicale, la génération des énoncés et l'enchaînement conversationnel, l'activité centrale de la compétence linguistique de l'agent, c'est-à-dire de son modèle de la langue. Nous proposons également de considérer Anadia, non plus comme étant uniquement lié au modèle de la langue, mais comme un principe de structuration de la base de connaissances de l'agent artificiel. Dans cet objectif, le modèle de la langue utilise Anadia pour représenter des systèmes de significations de même que le modèle de la tâche et de la situation l'utilise pour la représentation de connaissances pragmatiques (par exemple, la représentation

dans une table de différences des personnes présentes dans la situation et de leurs fonctions relatives est une connaissance pragmatique situationnelle). Anadia devient ainsi pour l'agent un modèle de mémoire, mais c'est ici d'une mémoire atemporelle dont il s'agit car, bien que cette mémoire soit dynamique, le temps en tant que tel n'y est pas représenté. Les aspects qui concernent le temps de l'interaction langagière sont liés à la dynamique du dialogue dans l'espace formé des axes régissant et incident. C'est donc le modèle du dialogue qui doit comprendre un principe de mémoire épisodique, d'autant plus que les aspects temporels de l'interaction peuvent être des indices pertinents pour l'actualisation du sens et l'enchaînement conversationnel dont on a signalé précédemment qu'ils incombent au modèle de dialogue.

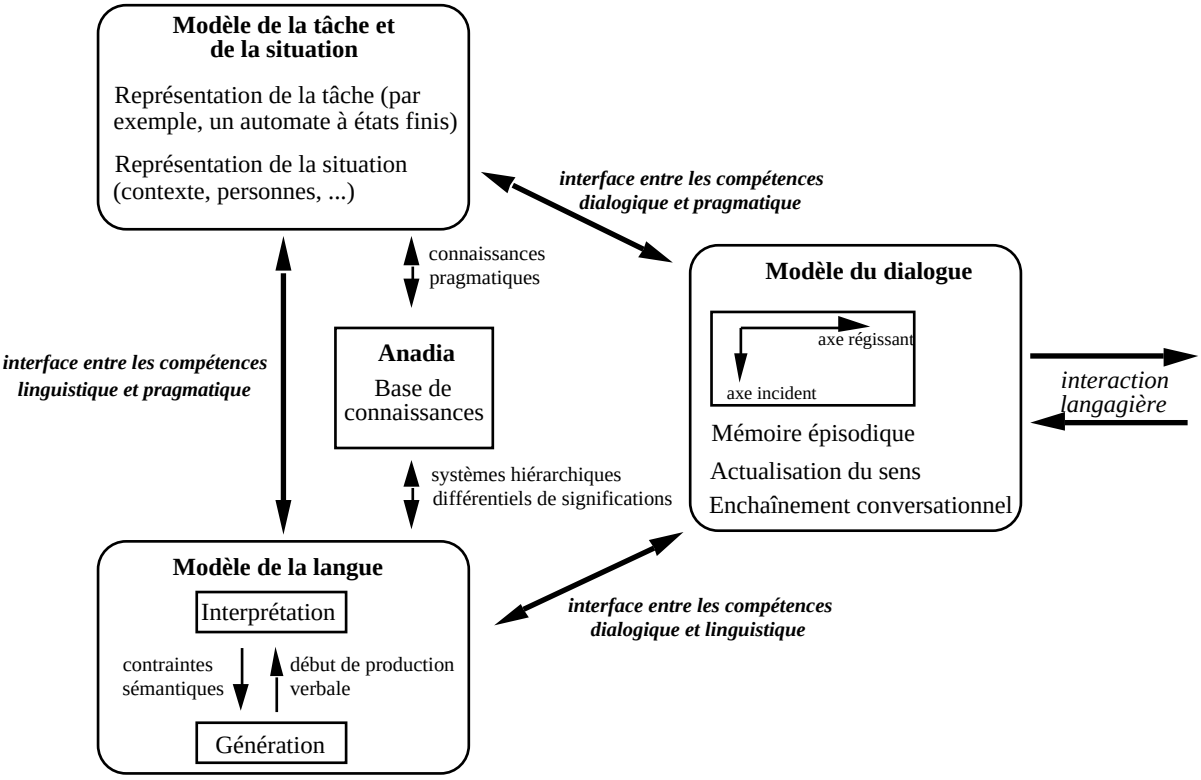


Figure 6.1 : Un modèle conceptuel pour agent logiciel dialoguant en langue naturelle

3. Apports pour le Traitement Automatique des Langues

Nous allons tenter de mettre en évidence la généricité du modèle proposé en montrant son apport à d'autres problématiques de l'informatique, et plus précisément, du traitement automatique des langues. Nous allons montrer l'intérêt du modèle pour deux problèmes différents. Le premier est celui du calcul automatique des références anaphoriques pour lequel il s'agit d'un apport théorique. Le second est un problème de veille technologique consistant en

une analyse sémantique de surface de documents textuels pour lequel nous allons montrer un apport applicatif de notre réalisation informatique.

3.1. Eclairer le problème du calcul des anaphores

Le problème du calcul des anaphores (cf. chapitre 2) est bien connu en linguistique ainsi qu'en traitement automatique des langues. Étant donné que les anaphoriques traduisent des reprises référentielles dans le discours, il s'agit de savoir quel syntagme du co-texte (ou bien selon l'approche choisie, quel objet du monde) l'anaphorique reprend. Par exemple, dans la phrase :

*La femme est montée dans la voiture, **elle** était très belle.*

la question posée est de savoir (quand c'est possible, ce qui n'est pas le cas ici étant donné que la phrase est donnée de façon isolée) à qui, la femme ou la voiture, le pronom anaphorique *elle* fait référence.

Le problème du calcul des anaphores a déjà fait l'objet d'un bon nombre de recherches ([Ariel 90], [Dupont 95] notamment). Des résultats assez satisfaisants ont pu être obtenus, par exemple, par le système RAP ("Resolution for Anaphora Procedure" [Lappin & al. 95]). RAP a pu identifier correctement 85% des anaphores pronominales sur un corpus en langue anglaise contenant jusqu'à 560 pronoms. Cette recherche sur le calcul des antécédents anaphoriques se place dans une approche cognitive ([Kleiber 94]) du calcul du sens et elle exprime le rattachement anaphorique en termes de saillance mémorielle. Ceci revient à classer d'entrée par ordre de préférence les antécédents possibles d'un anaphorique dans son co-texte et à retenir celui qui a le degré de saillance requis par l'élément anaphorique. À la suite de notre travail, ce qui nous semble pertinent pour le calcul de la dépendance sémantique de nature anaphorique, c'est de pouvoir proposer, dans un premier temps, les candidats possibles pour le rattachement anaphorique, sans se poser le problème de les classer sur des critères syntaxiques ou morphologiques. Le calcul de l'ensemble des candidats prend en compte les réalités syntaxiques sur lesquelles l'anaphore s'appuie et c'est en fonction de ces réalités que des lexèmes peuvent être exclus d'une possibilité de co-référence avec un anaphorique donné. La théorie du gouvernement et du liage de Chomsky ([Chomsky 91]) offre des principes qui vont dans ce sens que nous allons très rapidement expliquer.

La théorie du gouvernement et du liage propose un ensemble de règles pour contraindre le rattachement des pronoms anaphoriques que peuvent contenir les propositions d'une phrase. Dans la phrase, le liage permet de déterminer le domaine syntaxique où l'antécédent d'un anaphorique doit se trouver et, en fonction de ce résultat, on peut exclure d'une possibilité de rattachement les entités n'appartenant pas à ce domaine. Le domaine d'un anaphorique, c'est sa

proposition, et le problème consiste alors à savoir en fonction de la nature même de l'anaphorique s'il est lié dans son domaine ou pas. Dans cette perspective, Chomsky distingue deux formes d'anaphoriques : d'une part, les "*anaphors* "(expressions réciproques et pronoms réfléchis, ex. *se, l'un, l'autre,* etc.) et, d'autre part, les "*pronouns*" (pronoms personnels et définis, ex. *il, elle, le,* etc.). À chacune de ces deux catégories d'anaphoriques correspond un axiome de liage :

Axiome A : Les *anaphors* doivent être liés dans leur domaine.

Axiome B : Les *pronouns* doivent être libres dans leur domaine.

Ainsi, dans la phrase

Jean dit que ***Pierre se*** *lave.*

l'antécédent de *se* est *Pierre* car *se* satisfait l'axiome A de la théorie du liage et qu'il n'y a pas d'autre antécédent possible dans son domaine (en l'occurrence, la proposition *que Pierre se lave*). De même, dans la phrase

Jean *dit que Pierre* ***le*** *lave.*

l'antécédent de *le* est *Jean* car, d'après la théorie du liage, il doit satisfaire l'axiome B, et le seul substantif en dehors de son domaine (la proposition *que Pierre se lave*) est le nom *Jean.*

Les principes de la théorie du liage sont formulés pour une étude syntaxique de la phrase dans le cadre théorique de la grammaire générative. Notamment, un autre principe du liage énonce qu'un élément qui lie un autre doit le c-commander ([Reinhart 76]) et la relation de c-commande est directement liée à une écriture de la syntaxe de la phrase dans un formalisme de grammaire transformationnelle (A c-commande B si le premier noeud qui domine A dans l'arbre syntaxique domine B sans que A domine directement B). Cependant, les principes de Chomsky sont des principes liés à la langue et pas uniquement à un formalisme théorique. Il faut les réexprimer dans le formalisme des grammaires de dépendance et surtout étendre les principes concernant les domaines à des séquences plus étendues que la phrase, la période de [Rastier & al. 94] par exemple. Mais ceci suppose un travail de modélisation syntaxique qui sort du cadre de cette thèse.

Dans un second temps, après le calcul des candidats possibles, le choix de l'un d'eux pour le rattachement anaphorique constitue la dimension purement sémantique du problème de l'anaphore. C'est ce problème du choix sémantique que le modèle que nous proposons permet d'éclairer, car il peut être ramené à un problème de conservation de la cohérence interne de l'énoncé (cf. chapitre 4). L'anaphorique entretient des relations syntagmatiques dans sa proposition et ces relations contraignent son contenu sémique afférent en termes de cohésion et de cohérence interne. Par exemple, dans le cas d'une dépendance sémantique due à une relation

actancielle entre un anaphorique et un verbe, il doit y avoir isotopie entre le contenu de l'anaphorique et celui du verbe pour étayer la cohésion de la proposition. L'actualisation de son contenu sémique doit aussi satisfaire les prescriptions syntagmatiques dues à la relation actancielle (i.e. les actualisations entre les contenus de l'anaphorique et du verbe doivent être compatibles) pour assurer la cohérence interne de la proposition. Ce sont des contraintes syntagmatiques de ce genre qui permettent de rattacher différemment l'anaphorique *ils* dans les deux exemples (1) et (2) suivants[81], alors que le co-texte est très peu différent d'un énoncé à l'autre :

(1) ***Les conseillers municipaux*** *interdisent aux manifestants de défiler parce qu'**ils** craignent des troubles*

(2) *Les conseillers municipaux interdisent aux **manifestants** de défiler parce qu'**ils** veulent faire la révolution.*

Par exemple, on peut définir la lexie *révolution* par l'actualisation de la valeur *changer* du sème oppositionnel [Structures sociales : conserver *vs*. changer]. De ce fait, dans la phrase (2), le maintien de la cohérence interne de la proposition *ils veulent faire la révolution* demande l'affférence à l'anaphorique *ils* du même sème, actualisé de la même façon. En considérant que *manifestants* actualise aussi la valeur *changer* du sème [Structures sociales : conserver *vs*. changer] alors que *conseillers municipaux* actualise la valeur *conserver*, le rattachement de l'anaphorique à son syntagme n'est plus problématique. Pour maintenir la cohérence sémantique de l'énoncé, il doit être lié au syntagme qui actualise le sème de la même façon, c'est-à-dire qu'il doit être lié à *manifestants*.

Au niveau opératoire, envisager de cette façon le calcul du rattachement anaphorique consiste finalement à éclairer la phase sémantique de ce calcul par la recherche opératoire des isotopies. L'apport de notre modèle au problème du calcul des anaphorique tient en cette recherche d'isotopies.

Enfin, dans un dernier temps du calcul des antécédents des anaphoriques, il s'agira de retenir l'entité la plus récemment évoquée s'il y a encore le choix entre plusieurs possibilités.

En résumé, nous proposons trois étapes à la recherche des antécédents anaphoriques :

[81] Ces exemples sont empruntés à J. Pitrat.

1) Sélection des candidats syntaxiquement possibles (application des principes du liage étendus)

2) Sélection des candidats sémantiquement possibles (en préservant la cohérence interne de l'énoncé)

3) Choix de l'entité évoquée la plus saillante (i.e. celle qui est l'objet de la monstration langagière)

Comme on vient de le voir, le modèle de l'interprétation proposé peut être envisagé comme un module spécifique intervenant dans un calcul des anaphores. En retour, un module du calcul des antécédents anaphoriques pourrait permettre d'affiner le modèle de l'interprétation. Dans la réalisation informatique du modèle que nous avons présentée, les co-adaptations des contenus des lexèmes en relation d'anaphore sont traités dans la quatrième phase de recherche d'isotopies. Le calcul des antécédents anaphoriques n'étant pas implémenté, ces co-adaptations sont envisagées uniquement par la co-présence des lexèmes dans la chaîne, donc comme une simple influence sémantique. Il s'agit d'une approximation et un calcul syntagmatique des anaphores donnerait à ce genre de co-adaptation co-textuelle un réel statut de dépendance sémantique, conformément à nos hypothèses sur la dépendance sémantique de nature anaphorique (cf. chapitre 2). Au niveau opératoire, une fois les anaphores calculées dans le co-texte, la recherche de la dépendance sémantique (plus précisément le calcul de l'influence sémantique portée par la dépendance anaphorique) pourra s'effectuer de la même manière que dans les cas de constitution de groupes nominaux et de relations actancielles. Elle donnera lieu à une phase qui précède le calcul des isotopies dues aux co-présences de lexèmes n'entretenant pas de relations syntagmatiques (i.e. la phase de recherche de simples influences sémantiques).

3.2. Une application de veille technologique assistée par ordinateur

Nous allons montrer dans cette partie un apport applicatif de la réalisation logicielle proposée en expliquant comment adapter l'outil initialement prévu pour l'analyse des énoncés d'un dialogue à un traitement interprétatif de documents textuels.

Le développement du courrier électronique et du Wold Wide Web rend l'information nécessaire et omniprésente. L'informatique s'adapte à cette réalité à tel point que les systèmes d'exploitation intègrent aujourd'hui des outils de navigation et de recherche sur l'Internet. La prolifération des sources d'information à gérer au jour le jour pose des problèmes d'accès à l'information pertinente et de filtrage de documents. L'information à traiter étant majoritairement textuelle, les recherches en traitement automatique des langues trouvent, à travers ces problèmes dits de veille technologique, un bon nombre de champs d'application.

Comme on l'a vu précédemment, l'outil logiciel proposé dans cette thèse est conçu pour la recherche des contraintes de monstration d'un énoncé. De ce point de vue, il réalise un traitement sémantique de surface des énoncés car les contraintes de modalité, de prédication et d'argumentation (cf. chapitre 1) ne sont pas l'objet de son analyse interprétative. Ce traitement sémantique de surface est suffisant dans un contexte de veille technologique pour amener une réponse au problème consistant à savoir si un document donné traite de tel ou tel sujet. L'application pratique ici proposée de l'outil logiciel est la réalisation d'une prothèse cognitive pour un usager devant savoir a priori quels documents sont susceptibles de l'intéresser parmi plusieurs.

En expliquant au chapitre précédent le fonctionnement de la réalisation logicielle, nous avons signalé que les isotopies calculées dans le processus d'interprétation d'un énoncé sont conservées en mémoire pour renforcer les contraintes de monstration dans l'analyse des énoncés suivants. En exécutant le processus interprétatif sur un texte entier, les contraintes obtenues à la fin de l'exécution du processus rendent compte ainsi de la monstration du texte entier, et non simplement de son dernier énoncé. Dans l'objectif de l'analyse d'un texte, la modification à apporter à la réalisation informatique présentée dans le chapitre précédent consiste donc à ne plus convoquer la validation du partenaire humain à la fin de chaque énoncé, mais à la fin du texte. Cette modification a été réalisée pour générer l'outil logiciel de veille technologique que nous présentons ici.

Cette prothèse cognitive d'interprétation de documents donne la possibilité à un usager de paramétrer un système de filtrage de textes par rapport à ses centres d'intérêt. En construisant ses propres systèmes hiérarchiques différentiels (en fonction de ce qui l'intéresse), un utilisateur particulier peut guider le processus interprétatif pour qu'il ne repère dans un texte uniquement ce que ses connaissances lui permettent de repérer. Les systèmes hiérarchiques et différentiels sont ainsi une alternative à la description des buts de l'usager par des formules logiques. L'application consiste alors à interpréter automatiquement les documents à traiter pour extraire ceux qui semblent présenter un intérêt pour l'utilisateur. Le résultat de l'analyse d'un texte par l'outil logiciel conduit à la mise en forme d'un rapport d'interprétation indiquant :

- Quelles lexies du texte ont un contenu sémantique dans la base de connaissances (c'est-à-dire dans les systèmes hiérarchiques différentiels construits par l'utilisateurs).

- Quels taxèmes de la base de connaissances sont évoqués dans le texte.

- Quels domaines d'interprétation sont évoqués dans les isotopies calculées lors de l'analyse du texte.

Avec les rapports interprétatifs associés à chacun des documents à traiter, l'utilisateur peut ainsi avoir un premier aperçu de ceux qui présenteront à ses yeux un intérêt plus particulier de manière plus pertinente qu'avec une utilisation statistique du lexique. La valeur ajoutée de l'outil logiciel par rapport aux méthodes statistiques est, qu'en plus des lexies repérées dans le texte, le rapport d'analyse fournit des informations (les taxème et les domaines d'interprétation) sur la thématique du document telle qu'elle est perçue à l'issue du processus interprétatif.

Pour donner un exemple de l'utilisation de l'outil d'interprétation, considérons un usager dont les centres d'intérêt sont la documentation technique de logiciels, les outils d'aides à l'utilisation d'un logiciels et les métiers et applications de l'informatique. Supposons que cette personne ait construit en utilisant l'outil proposé le système hiérarchique différentiel que nous avons détaillé à la fin du chapitre précédent. Envisageons que cette personne ait à traiter plusieurs documents dont le texte suivant extrait du site Web présentant le projet PIC (*http://doct.info.unicaen.fr/~fgerard/pic/partie-A/a1.html*) :

La conduite du projet rédactionnel

La réalisation d'une **documentation utilisateur** peut, suivant la complexité du **logiciel** à documenter, représenter une charge de travail relativement importante - il n'est pas rare qu'une **documentation** requière plus de six mois-hommes de travail. De nombreuses méthodes de conduite de projet ont été développées pour aider les **rédacteurs** à maîtriser le processus de développement de la **documentation**. On retrouve généralement dans ces méthodes les deux premières étapes suivantes : le transfert de compétence entre l'équipe de développement et l'équipe de rédaction, et la définition des besoins des lecteurs des **manuels** à rédiger.

Le transfert de compétence des équipes de développement vers les **rédacteurs** représente une charge de travail plus ou moins importante selon que l'équipe rédactionnelle est intégrée très en amont ou en aval du cycle de développement du **logiciel**. Plus les **rédacteurs** sont impliqués dans les phases d'expression des besoins et de conception, moins cette étape de transfert représente une charge importante.

Dans un premier temps, les **rédacteurs** procèdent à l'analyse de l'application en consultant les documents existants (cahier des charges, réponse à cahier des charges, spécifications fonctionnelles, etc.), en manipulant les maquettes ou les prototypes et en conduisant des interviews avec les **chefs de projets**, les **utilisateurs** cibles (lorsque cela est possible), les **développeurs** (éventuellement), etc. Les **rédacteurs** s'assurent de leur compréhension des grands principes de l'**application** en présentant celle-ci, comme s'ils en étaient les auteurs, à un public composé des **chefs de projets** et d'**utilisateurs** potentiels.

Dans un deuxième temps, et lorsqu'un accord est obtenu sur la compréhension générale de l'**application**, les **rédacteurs** se lancent dans la rédaction d'une première version de la **documentation** qui a encore principalement pour vocation de s'assurer que les **rédacteurs** ont une vision de l'**application** conforme à ce qu'en attendent les **utilisateurs** et à ce qu'en a compris l'équipe de développement. L'analyse des besoins des lecteurs constitue la deuxième étape du projet rédactionnel. Les **rédacteurs** doivent s'assurer qu'ils sont en contact avec des **utilisateurs** suffisamment représentatifs de la population qu'ils visent. Cette étape a pour but :

- de centrer sur les **utilisateurs** les décisions à prendre concernant le développement de la **documentation**, de demander aux **utilisateurs** quels sont leurs critères d'appréciation et leurs attentes vis à vis de la **documentation**, en s'appuyant le plus possible sur des exemples concrets

- de déterminer de façon concrète comment travaillent les **utilisateurs** et d'imaginer comment ils vivront leur nouvel environnement afin de se défaire d'idées a priori sur les lecteurs et sur le contexte d'utilisation des **manuels**
- d'observer les flux d'informations et d'actions constituant l'environnement de travail de l'**utilisateur** : l'organisation générale de la **documentation** et l'organisation interne à chacun des **manuels** doit se calquer sur l'organisation de travail de l'**utilisateur** et l'ordre d'exécution des tâches qui lui sont attribuées
- de recueillir la terminologie employée par les **utilisateurs** (en particulier le jargon, les sigles, les mots techniques, etc.)
- de repérer des exemples pertinents à proposer dans les futurs documents.

Dans la totalité des méthodes de réalisation de la **documentation**, (Galloüin & al., 1996) les deux étapes de transfert de compétence et de définition des lecteurs cibles sont menées de façon indépendante. Les problèmes d'organisation et de charge de travail qu'entraînent ces méthodes, la qualité généralement médiocre des **documentations** produites et l'éclairage apportée par des expériences de conception située et distribuée dans différents domaines conduisent à penser qu'il pourrait être intéressant de transposer cette approche au contexte de la rédaction de la **documentation utilisateur**. L'objectif est donc de modifier légèrement la conduite du projet rédactionnel en intégrant dans l'étape de transfert de compétence un futur lecteur/**utilisateur** du **logiciel** et de mettre le **logiciel** à la disposition de l'équipe ainsi formée pendant la phase préparatoire.

Figure 6.2 : Exemple d'un document à traiter : le fichier 'a1.html'

Toutes les lexies en gras dans ce texte sont celles qui sont définies dans la base de connaissance de l'outil logiciel (cf. chapitre 5 § 3.).

Le traitement interprétatif de ce texte conduit au résultat suivant. L'outil logiciel édite dans une fenêtre ce résultat sous la forme d'un rapport d'interprétation (cf. Figure 6.3).

• 7 lexies sont connues dans le texte :

utilisateur	(13 occurrences)
documentation	(10 occurrences)
rédacteur	(8 occurrences)
logiciel	(4 occurrences)
manuel	(3 occurrences)
application	(3 occurrences)
développeur	(1 occurrence)

• 5 taxèmes sont évoqués :

//technologie de l'information//
//informatique//
//logiciels-1//
//logiciels-2//
//documentation//

• 6 domaines d'interprétation ont été impliqués :

Utilisation de l'information (dans les 7 lexies repérées)
Type de composant informatique (dans les 7 lexies repérées)
Rôle du composant informatique (dans les 7 lexies repérées)
Fonction de la personne (dans 'rédacteur', 'utilisateur', 'chef de projet' et 'développeur')
Rapport au logiciel (dans 'application', 'documentation', 'manuel')
Pratique du logiciel (dans 'application', 'documentation', 'manuel')

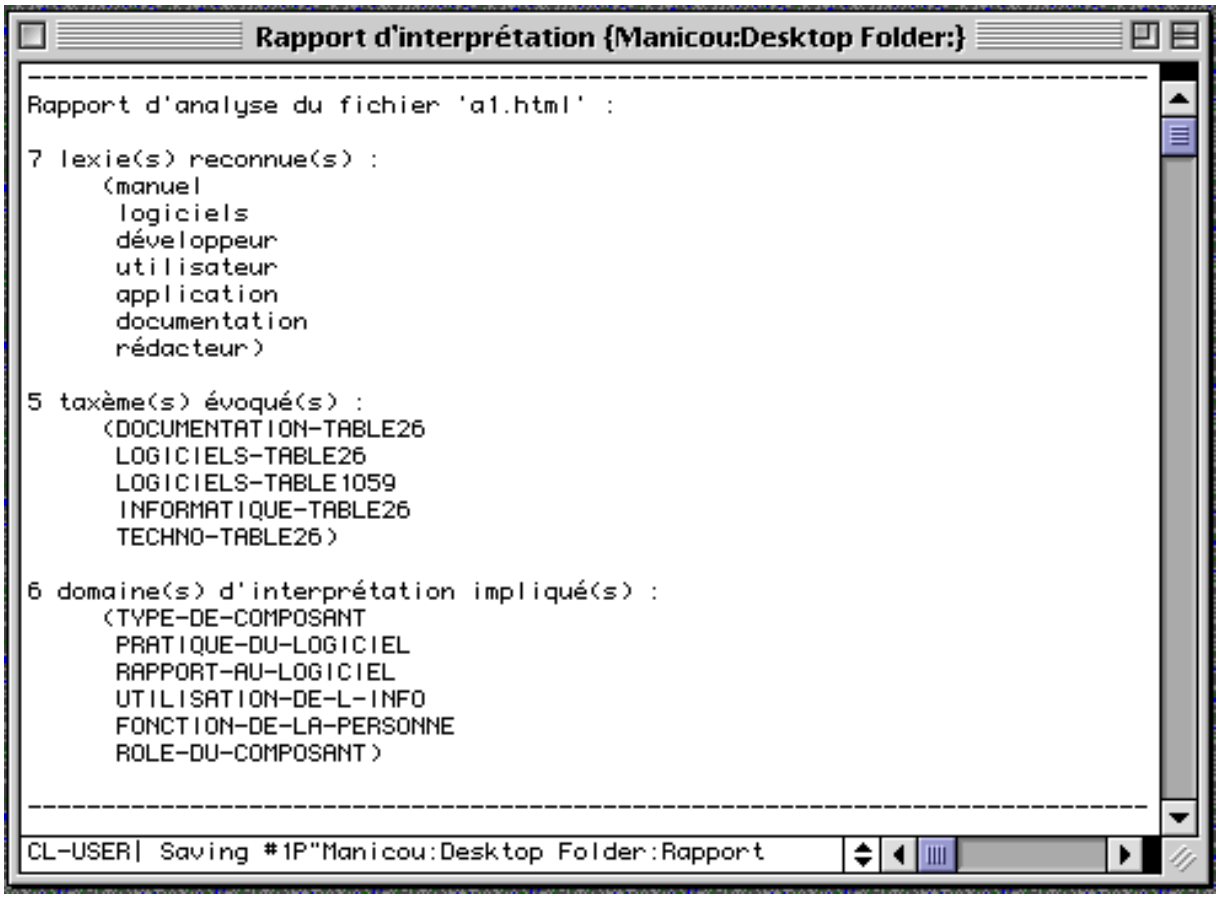


Figure 6.3 : Le rapport d'interprétation du fichier 'a1.html'

On peut remarquer deux choses à travers ce rapport d'interprétation. Premièrement, la lexie *chef de projet* présente dans le texte (2 occurrences) n'a pas été repérée et pourtant elle est définie dans le système hiérarchique différentiel (dans la table //logiciels-2//). Cette erreur est due au découpage morphologique réalisé par l'analyseur syntaxique qui en a fait trois mots au lieu d'un (c'est un problème que nous avons précédemment discuté en conclusion du chapitre 5). Deuxièmement, les domaines d'interprétation *Présentation de l'information* et *Support de l'information* n'ont pas été extraits alors qu'ils sont pointés par le contenu sémique de la lexie *manuel* dont 3 occurrences ont été repérées dans le texte. Cette fois, il ne s'agit pas d'une erreur. Ces domaines d'interprétation n'ont pas été trouvés parce qu'ils n'ont pas pris part à une isotopie bien qu'ils soient 3 fois indiqués dans le texte par la lexie *manuel*. Lorsque *manuel* apparaît dans une phrase du texte, seul son contenu générique s'inscrit dans une isotopie et son contenu spécifique (indiquant les domaines d'interprétations *Présentation de l'information* et *Support de l'information*) ne donne pas lieu à répétition. Pour que *Présentation de l'information* et *Support de l'information* figurent dans le rapport d'interprétation, il aurait fallu que deux des occurrences de *manuel* dans le texte soient dans une même phrase (et ce n'est pas le cas).

Malgré l'erreur concernant la lexie *chef de projet*, ce rapport est signe que l'interprétation du texte a été fructueuse et que le texte convoque une bonne part du système hiérarchique différentiel. C'est signe que le document est susceptible d'intéresser l'utilisateur. Au contraire, si aucune des lexies de la base de connaissances n'apparaît dans le texte analysé, le rapport sera le suivant :

- 0 lexies sont connues dans le texte
- 0 taxèmes sont évoqués
- 0 domaines d'interprétation ont été impliqués

Ceci indique a priori que le document en cause cette fois ne traite pas du domaine d'intérêt de l'utilisateur.

Dans une telle application, il est primordial que les représentations sémantiques du système ne soient pas complètes. Simplement, elle doivent être suffisantes pour mettre en évidence les documents susceptibles d'intéresser l'utilisateur. Dans la constitution de la base de connaissances sémantiques, la non connaissance d'une lexie n'est pas un défaut mais un indice important, synonyme pour l'utilisateur de non pertinence de cette lexie par rapport à l'information recherchée.

Le traitement sémantique de surface ainsi réalisé intègre l'utilisateur dans le processus d'extraction d'informations. Ce processus adapté à un utilisateur réalise une interprétation sélective et supervisée. En examinant les rapports d'interprétation fournis, et au fur et à mesure de son activité de veille technologique, l'utilisateur a toujours la possibilité de revoir la construction de ses systèmes hiérarchiques différentiels (en créant par exemple de nouvelles tables) pour affiner l'analyse sémantique et paramétrer de mieux en mieux l'outil logiciel par rapport à ses exigences. La méthode de conception par amorce est ici envisagée comme une méthode d'adaptation progressive de l'outil logiciel aux besoins de son utilisateur.

La réalisation logicielle que nous venons de présenter est une première maquette pour l'outil de vielle technologique visé. Certains de ses aspects restent à améliorer avant de pouvoir diffuser le logiciel à plusieurs personnes. On a, par exemple, évoqué le problème de découpage morphologique qu'il faut revoir en intégrant des connaissances syntaxiques et sémantiques. C'est aussi le cas du problème de la lemmatisation qui n'est pas ici traité en soi. Dans l'état actuel de l'implémentation, un contenu sémantique est attaché à une lexie si cette lexie apparaît (strictement) comme élément de la sélection d'une table. Ceci pose le problème des flexions, car si un document contient une lexie au pluriel alors que cette lexie apparaît au singulier dans une table, elle ne sera pas trouvée dans les connaissances paradigmatiques (étant donné la différence entre les chaînes de caractères de la lexie au pluriel et de la même lexie au singulier). Par conséquent, le contenu sémantique sera considéré comme vide. Pour représenter les diverses

formes morphologiques d'une lexie dans l'état actuel du système, il faut dédoubler les tables de différences (par exemple une table donne les formes des lexies au singulier et une autre table identique à la première donne les formes des lexies au pluriel). Pour apporter une solution réelle à ce problème, il faudra mettre en place, dans l'interface entre l'analyse syntaxique et l'analyse sémantique, un module de lemmatisation d'une lexie, et ne faire figurer que des radicaux dans les sélections des tables. Pour fournir une implémentation "solide" de l'outil de veille technologique, un projet d'étudiant d'informatique (maîtrise ou DEA) va être proposé cette année à l'université de Caen. L'objectif de ce stage sera de réaliser l'outil de sorte qu'il puisse être diffusé sur le World Wide Web, et donc utilisé par un bon nombre d'usagers. Une expérimentation de cet outil permettra alors de réaliser une validation quantitative du modèle présenté dans cette thèse.

4. Conclusions

Dans ce dernier chapitre, nous avons souligné les aspects importants de notre travail du point de vue de la recherche en informatique sur le traitement automatique des langues naturelles. Notamment, nous avons défendu l'intérêt d'un processus d'interprétation automatique guidé par une base de connaissances légère car non exhaustive. De cette légèreté provient principalement deux atouts. Le premier est d'ordre opératoire : moins la base de connaissance est vaste, moins les traitements informatiques qu'elle nécessite sont importants en terme de temps de traitement et d'espace mémoire. Le deuxième atout concerne le but même du processus d'interprétation : moins la base de connaissance est exhaustive, mieux elle peux être précisément adaptée à un but, que ce but soit de participer à la gestion d'une tache particulière dans un contexte de dialogue homme - machine, ou qu'il soit d'aider un utilisateur particulier dans un contexte de veille technologique.

En montrant les apports de notre modèle opératoire de recherche de contraintes sémantiques à d'autres problématiques que celle du dialogue homme-machine, nous avons montré que de la simplicité du modèle informatique découle sa réutilisabilité et sa généricité. Ceci constitue une élément de validation important du modèle. Le modèle étant fondé sur des bases théoriques linguistiques (plus précisément sur des bases théoriques de linguistique structurale), notre travail a également contribué à montrer l'apport de théories linguistiques à une problématique informatique d'ingénierie de traitement automatique des langues.

CONCLUSIONS ET PERSPECTIVES

Nous avons proposé dans cette thèse un modèle de compétence interprétative logicielle pour le dialogue homme-machine. Cette compétence se constitue au fur et mesure de son usage par interaction avec un partenaire humain selon un processus d'amorce.

Ce modèle prend place dans les recherches en informatique visant à améliorer les interactions entre l'homme et la machine. Les interfaces homme-machine à initiative variable sont le lieu d'une collaboration entre deux agents (un agent humain et un agent logiciel). Ces deux agents ont à gérer conjointement une tâche en associant à des actes de communication des actes informatiques tels les calculs, les affichages et les impressions, les accès aux banques de données ou encore la recherche d'information. Dans certains de ces domaines, les modèles grammatico-logiques ont apporté des solutions efficaces, comme en témoignent les moteurs de recherche sur Internet alliant une requête formelle à une recherche dans une base de documents indexés. Cependant, ces modèles n'offrent pas toujours la souplesse nécessaire à la gestion de l'interaction, et on voit leurs limites quand il convient d'exposer naturellement les problèmes des usagers, de définir conjointement la tâche à effectuer, d'indiquer des préférences et des contraintes ou encore d'utiliser les erreurs et incompréhensions mutuelles pour éviter les impasses et les bouclages. Pour tous ces problèmes, modéliser les compétences cognitives individuelles ne suffit pas, l'enjeu principal est de construire un terrain commun avec un partenaire. Le langage naturel est la modalité idéale pour la construction de ce terrain commun, mais pour cela, il a fallu modéliser les propriétés systémiques et sémiotiques des langues de manière pertinente pour une machine. C'est grâce à cela qu'une compétence interprétative

interactionniste peut permettre des interactions homme-machine plus naturelles, et donc plus efficaces.

La linguistique fournit des bases pour la mise en œuvre informatique de compétences langagières naturelles. Les apports de la linguistique à notre modèle sont importants, au niveau méthodologique et théorique. La démarche suivie en linguistique consiste à observer et à décrire les productions langagières. Suivant cette démarche dans une perspective informatique, nous avons mené une étude de corpus (comme notamment [Poirier 96], [Ferrari 97] ou encore [Déjean 98]) en introduisant des contraintes de modélisation (par exemple, la nécessité des contraintes de calculabilité et d'implémentation pour un modèle informatique). L'analyse de dialogues, et plus précisément l'analyse de l'enchaînement conversationnel et des négociations entre les différents interlocuteurs sur le sens en contexte des énoncés, nous a permis d'observer la langue *in situ*. C'est un moyen pour mettre au jour son fonctionnement avant d'en faire un modèle informatique. Au niveau théorique, l'apport de la linguistique (et plus spécifiquement de la sémantique différentielle) nous a fait poser deux hypothèses pour la réalisation d'une compétence interprétative pour un agent logiciel :

La première hypothèse est de considérer que le signifié linguistique n'est pas de nature ontologique et que la sémantique des langues ne se réduit pas à un inventaire référentiel mais forme un système où les signifiés se déterminent les uns et les autres. C'est bien ce que montrent les métaphores culturelles mises en évidence par Lakoff et Johnson, où les significations propres à un certain domaine s'expliquent dans un rapport à un autre domaine (c'est par exemple le cas dans la métaphore LA DISCUSSION, C' EST LA GUERRE). La sémantique que nous proposons est construite sur le paradigme différentiel des structuralistes et elle est conçue pour rendre compte de cet aspect systémique des langues. Ainsi, lorsqu'un domaine de signification entretient un rapport métaphorique avec un autre domaine de signification, c'est parce que l'on y retrouve les mêmes différences, conditionnant des oppositions similaires (par exemple, on utilisera les différences qui structurent le domaine de LA GUERRE pour dresser la combinatoire de significations du domaine de LA DISCUSSION). Nous retenons donc les notions de différence, d'opposition et de valeur comme moyen de décrire la langue en tant que système sémiotique, et de séparer la question du sens de celle de l'ontologie. Pour cela, nous avons proposé un modèle informatique basé sur des principes de catégorisation différentielle.

La seconde hypothèse est que, si la signification est déterminée par la différence, c'est-à-dire par des jeux d'oppositions, le sens d'un énoncé, quant à lui, est fondé sur la récurrence, c'est-à-dire par des faisceaux de répétitions. L'interprétation des énoncés opère donc une articulation entre deux ordres : l'ordre de la différence, le paradigmatique ; et l'ordre de la récurrence, le syntagmatique. La langue discrétise le continu de notre environnement par la

différence et la récurrence. La notion d'isotopie permet de rendre compte de la façon dont les différences paradigmatiques conditionnent les récurrences syntagmatiques, et réciproquement. C'est le fondement du modèle informatique proposé pour l'interprétation. Elle a attiré notre attention par la grande simplicité de sa définition et par la pertinence et l'omniprésence de son utilisation (que ce soit pour décrire des groupes nominaux, des énumérations, des relations actancielles ou encore des anaphores). Elle nous a permis de mettre au jour la co-détermination co-textuelle des constituants de la chaîne, et par là d'opérationnaliser une mise en forme holiste du sens qui articule la différence et la récurrence. Elle nous a également apporté un cadre de description pour les notions de cohésion et de cohérence des énoncés, permettant notamment d'envisager la part de la sémantique dans l'enchaînement conversationnel.

En nous focalisant sur les processus de co-référenciation, nous avons principalement modélisé la dimension de monstration de la sémantique des langues. Avant d'expérimenter une compétence dialogique artificielle dans une situation d'interaction homme-machine, il faut décrire sur les mêmes bases les autres dimensions de la sémantique. Ainsi, il faut aussi spécifier la mise en forme de contraintes sémantiques pour le calcul des antécédents anaphoriques, pour la dimension prédicative et propositionnelle du sens des énoncés, pour la négation et l'argumentation ainsi que pour la sémantique des modalités. Dans cet objectif, on pourra utiliser le modèle de [Gosselin 96a] pour la sémantique des procès, le modèle de [Nicolle & al. 98] pour les modalités et la force illocutoire des actes de langage ou encore le modèle de [Raccah 97] pour la sémantique de l'argumentation. Sur le plan pragmatique, il faut un modèle opératoire de l'actualisation en contexte des contraintes sémantiques (c'est l'objectif de [Gérard & al. 98] qui modélise avec Anadia les aspects pragmatiques d'une tâche, avec comme exemple la distribution de boissons). Pour cette actualisation, il convient de considérer, entre autres, les spécificités du contexte et des interlocuteurs. Comme on l'a vu (cf. Chapitre 1), cela amène à poser le problème de la référence. Il est donc nécessaire de construire un modèle extralinguistique de l'objet référent, dont nous avons proposé les bases dans un modèle fondé sur l'identité de l'objet ([Beust & al. 97a]), et d'exprimer les contraintes sémantiques (principalement celles de monstration) de l'énoncé en tant que connaissances conceptuelles sur de possibles objets qu'il conviendrait de rechercher dans une représentation du contexte et de la situation d'énonciation.

Une autre perspective de notre travail est l'acquisition de significations non supervisée par un partenaire humain. Pour faciliter les interactions homme-machine dans Anadia, il conviendrait de mettre en place une acquisition automatique de systèmes de significations pour donner à un agent logiciel des connaissances sémantiques en préalable à leur partage dans l'interaction avec son partenaire humain. Il ne s'agit pas ici de remettre en cause la nécessité de l'interaction homme-machine dans la co-construction de systèmes de signification, mais il est question d'examiner dans quelles conditions un agent logiciel construit sur Anadia pourrait à

partir de dictionnaires électroniques ou de bases de connaissances, inférer des systèmes différentiels de signes. Une perspective est de modéliser des processus de méta-contrôle au dessus d'Anadia (construction et examen automatique de tables par exemple) et de reprendre les résultats de [Ploux & al. 98] (issus d'une collaboration entre l'INaLF et l'ELSAP) qui, à partir de dictionnaires de synonymes (sept au total : le *Bailly*, le *Benac*, le *du Chazaud*, le *Guizot*, le *Lafaye*, le *Larousse* et le *Robert*), modélise l'espace synonymique des mots en terme de variation sémantique, d'homonymie, de prototypie et de généricité. Les systèmes de significations qui seraient ainsi construits devraient alors, par la suite, être attestés, reformulés ou réfutés dans l'interaction homme-machine quand leur pertinence est posée dans l'interprétation d'un énoncé.

Lorsque le système informatique aura à gérer un très grand nombre de différences dans ses représentations paradigmatiques, une évaluation sera nécessaire. Évaluer la compétence langagière de la machine ne sera pas un enjeu strictement informatique. Ce genre d'évaluation concerne toutes les dimensions des sciences du langage et des sciences cognitives. Nous avons proposé ici un modèle informatique de la sémantique des interactions langagières en nous appuyant sur des travaux pluridisciplinaires. De ce fait, l'évaluation du modèle ne peut être menée ni par une seule personne, ni par une seule discipline. C'est la raison pour laquelle cette thèse touche à sa fin. Par exemple, dans l'outil logiciel réalisé nous avons choisi de présenter les contraintes sémantiques par ordre d'importance allant de la plus générique à la plus spécifique ; mais il conviendrait de faire une étude pluridisciplinaire pour savoir si dans certains discours, en fonction de certaines tâches ou en fonction de certains sujets, les contraintes spécifiques ne sont pas plus importantes que les contraintes génériques. Il faut donc maintenant mettre en place diverses collaborations (notamment entre l'informatique, la linguistique et la psychologie) pour confronter l'apprentissage interactif de systèmes différentiels de significations et les résultats opératoires qui en sont issus, à différentes pratiques, à différentes situations d'interaction, et avec différents corpus.

La question d'une validation quantitative du modèle n'est pas simple. Il conviendrait de mettre au point, de façon pluridisciplinaire, des méthodes d'évaluation pour les systèmes d'interaction homme-machine pour pouvoir confronter différents modèles et, par conséquent, les affiner. C'est une question passionnante dont l'intérêt dépasse de loin le classement des modèles en fonction de leur taux de succès. Sur le problème d'un classement relatif des systèmes d'étiquetage morpho-syntaxique, l'action GRACE[82] a montré la difficulté de la mise au point de critères d'évaluation fiables alors qu'*a priori* attribuer une fonction syntaxique à un mot dans un énoncé apparaît simple et consensuel. De même, quand il s'agit d'extraire d'une

[82] http://m17.limsi.fr/TLP/grace/

dépêche journalistique une information déjà partiellement connue, les conférences MUC[83] ont montré que le problème du choix des critères d'évaluation n'est encore pas évident bien qu'il y ait déjà eu six conférences sur ce problème. La question d'évaluer l'aspect "naturel" d'une interaction avec une machine est de toute évidence une question difficile. La mise au point de critères d'évaluation pour les systèmes de dialogue homme-machine demandera donc certainement beaucoup d'investissements en temps et en ressources humaines. D'autant plus que le dialogue homme-machine est un champ d'expérimentation et d'intégration pour les modèles de la morpho-syntaxe, de la sémantique, de la pragmatique, de l'enchaînement conversationnel, de l'apprentissage personne-système ou encore de la représentation de l'autre dans l'interaction. À terme, c'est grâce aux apports de plusieurs disciplines que pourront être mises en place, comparées et évaluées des interactions langagières "naturelles" avec des machines.

[Victorri 98] signale que, depuis une trentaine d'années que l'informatique se penche sur les problèmes de modélisation des langues naturelles pour les machines, les recherches ont profondément changé de nature. Ainsi, les informaticiens ont appris à connaître les autres disciplines des sciences du langage (linguistique et psychologie par exemple), à travailler avec elles, à se méfier des annonces trop rapides et à envisager les problèmes langagiers dans des recherches à long terme. Il en résulte, selon l'auteur, une maturité nouvelle de l'informatique qui fait suite à l'enthousiasme et à l'ingénuité des débuts du traitement automatique des langues. Nous espérons, par notre travail, avoir contribué à cette maturation.

[83] http://cs.nyu.edu/cs/faculty/grishman/muc6.html

RÉFÉRENCES BIBLIOGRAPHIQUES

[Andreae & al. 95] : Andreae T., Nölle M., Schreiber G., February 1995, Embedding Cartesian Products of Graphs into de Bruijn Graphs, Technical Report, Technische Informatik I, TU-HH, Number TR-402-95-014.

[Andrès 95] : Andrès M., 1995, *Étude des stratégies de prise de décision d'une machine dans le contexte d'un dialogue enfant-machine, CEDRE : Compère En Dialogue de Référence avec un Enfant*, Thèse de l'Université de Caen.

[Appelt & al. 93] : Appelt D. E., Hobbs J. R., Bear J., Israel D., Tyson M., 1993, *FASTUS : A Finite-state Processor for Information Extraction from Real-world Text*, In proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI), Chambéry.

[Ariel 90] : Ariel M., 1990, *Accessing Noun Phrases Antecedents*, London, Routledge.

[Aristote] : Aristote, *Organon, Les catégories*, Ed. Jean Tricot, 1969.

[Asher & al. 94] : Asher N., Aurnague M., Bras M., Sablayrolles P., Vieu L., 1994, Computing the Spaciotemporal Structure of Discourse, In proceedings of the Computational Semantics Workshop, Tilburg (Netherlands).

[Austin 70] : Austin J. L., 1970, *Quand dire, c'est faire*, Paris, Le Seuil.

[Baader 97] : Baader F., 1997, *Logic-based knowledge Representation*, Reader of ESSLLI Summerscholl, Aix-en-Provence.

[Bachimont 96] : Bachimont B., 1996, *Herméneutique matérielle et artéfacture : des machines qui pensent aux machines qui donnent à penser. Critique du formalisme en intelligence artificielle*, Thèse d'épistémologie de l'Ecole Polytechnique.

[Bartning 95] : Bartning I., 20-22 Septembre 1995, *Les syntagmes nominaux complexes en "de". Typologie des interprétations et stratégies discursives*, Colloque Anaphore et Référence, Nancy, CRIN.

[Bassi Acuña 95] : Bassi Acuña A., 1995, *Un modèle dynamique de la compréhension de textes intégrant l'acquisition de connaissances*, Thèse de l'Université Paris XI, Orsay.

[Beust & al. 96] : Beust P., Delépine L., Nicolle A., Coursil J, Octobre 1996, *Anadia, a Relevance Examination Tool for Representations*, Third European Congress on System Science, AFCET-CSCI, Rome.

[Beust & al. 97a] : Beust P., Nicolle A., Février 1997, *L'identité dans un modèle de la dépendance nominale*, Sixièmes Journées de Rochebrune "Invariance, Interaction, Référence : L'identité en question", Megève.

[Beust & al. 97b] : Beust P., Nicolle A., Juin 1997, *La référence dans un modèle interactionniste de la signification*, 11ème colloque international du Cercle de Recherche Linguistique du Centre et de l'Ouest (CERLICO), Caen. In *La référence, statut et processus*, sous la direction de N. Le Querler et E. Gilbert, Presses Universitaires de Rennes, 1998, p. 269-289.

[Beust & al. 97c] : Beust P., Delépine L., Jacquet D., Septembre 1997, *Rédacteur, Utilisateur, Concepteur : Quelle référence commune?*, 01DESIGN'97, Théoule-sur-mer, In *Les objets en conception*, sous la direction de B. Trousse et K. Zreik, Europia, 1998, p. 173-180.

[Beust & al. 98] : Beust P., Jacquet D., Nicolle A., 1998, *La co-référence dans la conception*, Revue des Sciences et Techniques de la Conception, Vol. 6, n° 1, p. 25-37.

[Beust 98] : Beust P., Septembre 1998, *Un modèle oppositionnel de la signification*, RECITAL'98, Le Mans.

[Beust & al. 99] : Beust P., Nicolle A., 13-15 Janvier 1999, *A bootstrap of interpretation comeptence for a software agent*, In Third Internationnal Worshop on Computational Semantics (IWCS-3), Tilburg, The Netherlands

[Blanché 66] : Blanché R., 1966, *Les structures intellectuelles*, Paris.

[Brachman 78] : Brachman R. J., 1978, *Structured Inheritance Networks*, in W. A. Woods and R. J. Brachman : *Research in Natural Language Understanding*, Quarterly Progress Report No. 1, BBN Report No. 3742, Cambridge, Bolt, Beranek and Newman Inc., p. 36-78.

[Brassac & al. 96] : Brassac C., Stewart J., Février 1996, *Le sens dans les processus interlocutoires, un observé ou un co-construit ?*, Cinquièmes Journées de Rochebrune "Du collectif au social", Megève.

[Bréal 1897] : Bréal M., 1897, *Essai de sémantique (science des significations)*, Paris, Hachette (ré-édité chez Brionne, Gérard Monfort, 1982).

[Bricon-Souf 94] : Bricon-Souf N., 1994, *ARCICE : Architecture Réflexive d'un Compèrobot Interactif, Compréhension des Énoncés*, Thèse de l'Université de Caen.

[Bruner 83] : Bruner J. S., 1983, *Le dévelopement de l'enfant. Savoir faire, savoir dire*, Paris, Presses Universitaires de France.

[Caron 89] : Caron J., 1989, *Précis de psycholinguistique*, Paris, Presses Universitaires de France.

[Charolles 91] : Charolles M., Juillet 1995, *L'anaphore. Définition et classification des formes anaphoriques*, École d'été de l'Association pour la Recherche Cognitive (ARC), Bonas.

[Charolles & al. 93] : Charolles M., Schnedeker C., 1993, *Coréférence et identité : le problème des référents évolutifs*, Langages 112, p. 106-126.

[Charolles 94] : Charolles M., 1994, *Reprise pronominale et prédications transformatrices : l'interprétation de la référence pronominale à la suite du verbe "changer"*, In actes du colloque "Relations anaphoriques et (in)cohérence" (1-2 Décembre 1994), Anvers.

[Chinchor & al. 93] : Chinchor N., Hirschman L., Lewis D. D., 1993, *Evaluating Message Understanding Systems : An analysis of the Third Message Understanding Conference (MUC-3)*, Computational Linguistics, vol. 19, n° 3, p. 409-451.

[Chomsky 69] : Chomsky N., 1969, *Structures Syntaxiques*, Paris, Le Seuil.

[Chomsky 91] : Chomsky N., 1991, *Théorie du gouvernement et du liage. Les conférences de Pise*, Paris, Seuil.

[Clark & al. 86] : Clark H. H., Wilkes-Gibbs D., 1986, *Referring as a Collaborative Process*, Cognition, n° 22, p. 1-39.

[Corbin 88] : Corbin D., 1988, *Pour un composant lexical associatif et stratifié*, DRLAV, 38, p. 63-92.

[Coseriu 76] : Coseriu E., 1976, *L'étude fonctionnelle du vocabulaire : précis de lexématique*, Cahiers de lexicologie, n° 29, p. 5-23.

[Coursil 92] : Coursil J., 1992, *Grammaire analytique du français contemporain. Essai d'intelligence artificielle et de linguistique générale*, Thèse de l'Université de Caen.

[Coursil 93a] : Coursil J., 1993, *Rapport en vue de l'obtention du diplôme d'habilitation à diriger des recherches*, Université de Caen.

[Coursil 93b] : Coursil J., 1993, *Dialogue : the semiology of transfert*, Second Congress on System Science, AFCET-CSCI, Prague.

[Coursil 97] : Coursil J., 1997, *Sémantique terministe*, Seminaire de sciences du langage, Département d'Etudes Pluridisciplinaires Appliquées, laboratoire GIL, Université des Antilles et de la Guyane.

[Coursil 98] : Coursil J., Mars 1998, *Le syllabaire saussurien. Introduction à la phonologie des groupes*, Langages, n° 129, Larousse, p. 76-88.

[Davis & al. 93] : Davis R., Shrobe H., Szolovits P., 1993, *What is a knowledge representation*, AI Magazine.

[Delépine & al. 95] : Delépine L., Nicolle A., Coursil J., 1995, *Transfert frameworks for modelling human-machine dialogue*, International symposium in Varna : Language and consciousness, Varna (Bulgaria).

[Delépine & al. 96] : Delépine L., Nicolle A., Juillet 1996, *Un assitant interactif de recherche d'information : Conception d'un agent aide-mémoire*, Conférence internationale sur l'Apprentissage Personne Système (CAPS), Caen.

[Delépine & al. 98] : Delépine L., Nicolle A., Lemaréchal R., Aout 1998, *ARIH - A Personal Interactive Assistant for Document Retrieval when Navigating the WWW*, Artificial Intelligence and Cognitive Science (AICS'98), Dublin, Irelande.

[Déjean 98] : Déjean H., Juillet 1998, *Approche distributionnelle pour la découverte de règles de syntaxe*, In *Deuxième Conférence Personne Système* (CAPS'98), p. 173, Caen, France.

[Dubois 93] : sous la direction de Dubois D., 1993, *Sémantique et cognition. Catégories, prototypes, typicalité*, Paris, CNRS Éditions.

[Dupont 95] : Dupont M., 1995, *En traitement automatique du langage, le calcul de la référence nécessite la gestion d'un modèle des attentes du lecteur*, Rapport technique du GREYC - CNRS URA 1526, Universté de Caen.

[Ducrot & al. 72] : Ducrot O., Todorov T., 1972, *Dictionnaire encyclopédique des sciences du langage*, Paris, Le Seuil.

[Ducrot & al. 95] : Ducrot O., Schaeffer J. M., 1995, *Nouveau dictionnaire encyclopédique des sciences du langage*, Paris, Le Seuil.

[Eco 88] : Eco U., 1988, *Sémiotique et philosophie du langage*, Paris, Presses Universitaires de France.

[Ermine 96] : Ermine J. L., 1996, *Les systèmes de connaissances*, Paris, Hermes.

[Fauconnier 84] : Fauconnier G., 1984, *Espaces mentaux*, Paris, Éditions de Minuit.

[Ferrari 97] : Ferrari S., 1997, *Méthode et outils informatiques pour le traitement des métaphores dans les documents écrits*, Thèse de l'Université Paris XI, Orsay.

[Fillmore 68] : Fillmore C., 1968, *The Case for Case*, In *Universals in Linguistic Theory*, Bach E. and Hams R. T. (eds.), New York, Holt, Rinehart and Winston.

[Fodor 86] : Fodor J., 1986, *La modularité de l'esprit*, Paris, Éditions de Minuit.

[François 80] : sous la direction de François F., 1980, *Linguistique*, Vendôme, Presses Universitaires de France.

[Fuchs 95] : Fuchs C., 1995, *Encore … des paraphrases : approches linguistique de la signification et mises en perspective cognitives*, in Bouscaren J. (ed.), *Langue et langage ; problèmes et raisonnement en linguistique*, Paris, Presses Universitaires de France, p. 279-300.

[Gérard & al. 98] : Gérard F., Nicolle A., Septembre 1998, *Bistro : un modèle de dialogue intégrant la manipulation de concepts*, RECITAL'98, Le Mans.

[Giguet & al. 97a] : Giguet E., Vergne J., 11-13 September 1997, *Syntactic analysis of unrestricted French*, In proceedings of the International Conference on Recent Advances in Natural Languages Processing (RANLP'97), Tzigov Chark, p. 276-281.

[Giguet & al. 97b] : Giguet E., Vergne J., 17-20 September 1997, *From Part-of-Speech Tagging to Memory-based Deep Syntactic Analysis*. In proceedings of the International Workshop on Parsing Technologies (IWPT'97), Boston, Massachussets, MIT.

[GIL 95] : Coursil J., Montlouis Calixte D., Delépine L., Beust P., 1995, *Anadia 1, Manuel de l'utilisateur*, Université des Antilles et de la Guyane.

[Gosselin 96a] : Gosselin L., 1996, *Sémantique de la temporalité en français, un modèle calculatoire et cognitif*, Louvain-la-neuve, Duculot.

[Gosselin 96b] : Gosselin L., 1996, *Le traitement de la polysémie contextuelle dans le calcul sémantique*, Intellectica, n° 22, p. 93-117, Association pour la Recherche Cognitive (ARC).

[Gosselin 98] : Gosselin L., 1998, *Le "paradoxe imperfectif", ou la disjonction entre assertion et prédication*, In actes du colloque "Prédication, Assertion, Information" (Juin 96), ed. M. Forsgren, K. Jonasson, H. Kronning, Acta Universitatis Uppsaliensis, Upsala, Sous Presses.

[Greimas 66] : Greimas A. J., 1966, *Sémantique structurale*, Paris, Larousse (ré-édité aux PUF en 1986).

[Greimas 83] : Greimas A. J., 1983, *Du sens II, Essais sémiotiques*, Paris, Editions du Seuil.

[Grice 67] : Grice H. P., 1967, *Logic and Conversation*, The William James Lectures, Université de Harvard.

[Groenendijk & al. 96] : Groenendijk J., Stokhof M., Veltman F., 1996, *Coreference and Modality*, In *Handbook of Contemporary Semantic Theory*, S. Lappin (ed.), Oxford, Blackwell.

[Guiraud 65] : Guiraud P., 1965, *Les structures élémentaires de la signification,* Bulletin de la société linguistique de Paris, p. 97-114.

[Harris 54] : Harris Z. S., 1954, *Distributional structure,* Word, p. 146-162.

[Hjelmslev 43] : Hjelmslev L., 1943, *Prolégomènes à une théorie du langage,* trad. française, 1968, Paris, Ed. de Minuit.

[Hofstadter 88] : Hofstadter D., 1988, *Ma Thèmagie,* Paris, InterEditions.

[Huls & al 95] : Huls C., Bos E., Claassen W., 1995, *Automatic Referent Resolution of Deictic and Anaphoric Expressions,* Computational Linguistics, vol. 21, n° 1, p. 59-81.

[Jackendoff 72] : Jackendoff R., 1972, *Semantic Interpretation in Generative Grammar,* MIT Press, Cambridge.

[Jacquet 95] : Jacquet D., 1995, *Processus d'adaptation dans un dialogue enfant-machine : "Utilisation d'une situation de dialogue enfant-machine simulé pour une étude développementale de production de consignes"*, Thèse de l'Université de Caen.

[Jakobson 63] : Jakobson R., 1963, *Essai de linguistique générale,* trad. Ruwet, Éditions de Minuit, Paris.

[Johnson-Laird 83] : Johnson-Laird P.N., 1983, *Mental models,* Cambridge University Press.

[Jullien 95] : Jullien Y., 1995, *Communication expérimentateur-machine sur une expérience psychologique : application de l'argumentation à la communication homme-machine,* In actes de 01DESIGN'95 : Aspects communicatifs en conception, Autrans.

[Jullien & al. 97] : Jullien Y., Nicolle A., 1997, *Magica, une plateforme pour des agents cognitifs contruisant leurs connaissances dans l'interaction.* Cahiers du GREYC (Université de Caen), n°1/97.

[Kamp & al. 93] : Kamp H., Reyle U., 1993, *From Discourse to Logic*, Kluwer Academic Publishers.

[Kerbrat-Orecchioni 96] : Kerbrat-Orecchioni C., 1996, *Texte et contexte,* Scolia, n°6.

[Kleiber 90] : Kleiber G., 1990, *Quand "il" n'a pas d'antécédent,* Langages 97, p. 24-50.

[Kleiber 93] : Kleiber G., 1993, *Anaphores pronominales et référents évolutifs ou comment faire recette avec un pronom,* Colloque "Relations anaphoriques et (in)cohérence", Anvers.

[Kleiber 94] : Kleiber G., 1994, *Contexte, interprétation et mémoire: approche standard vs approche cognitive,* Langue française 103, p. 9-22.

[Kleiber 95] : Kleiber G., 20-22 Septembre 1995, *Des anaphores associatives méronymiques aux anaphores associatives locatives,* Colloque Anaphore et Référence, Nancy, CRIN.

[Labat 98] : Labat J.M., 1998, *Résolution de problèmes : interactions entre l'homme et la machine*, Mémoire pour l'habilitation à diriger des recherches, Université de Paris VI.

[Lakoff & al. 85] : Lakoff G., Johnson M., 1985, *Les métaphores dans la vie quotidienne*, Paris, Éditions de Minuit.

[Langacker 87] : Langacker R., 1987, *Fondation of cognitive grammar*, Standford University Press.

[Lappin & al. 95] : Lappin S., Leass H. J., 1995, *An algorithm for Pronominal Anaphora Resolution*, Computational Linguistic 20(4), p. 535 - 561.

[Latraverse 87] : Latraverse F., 1987, *La pragmatique*, Bruxelles, Mardaga.

[Lehuen 97] : Lehuen J., 1997, *Un modèle de dialogue dynamique et générique intégrant l'acquisition de la compétence linguistique*, Thèse de l'Université de Caen.

[Lenat & al. 90] : Lenat D., Guha R., 1990, *Builing large knowledge based systems*, Reading, Addison Welsey.

[Levrat 93] : Levrat B., 1993, *Le problème du sens dans les systèmes de traitement automatique du langage naturel : une approche alternative au travers de la paraphrase*, Thèse d'etat de l'Université de Paris-Nord.

[Luzzati 92] : Luzzati D., 1992, *Un modèle dynamique pour le dialogue homme-machine*, In *Recherches sur la Philosophie et le Langage*, n° 14, p. 263-293.

[Martin 83] : Martin R., 1983, *Pour une logique du sens*, Paris, Presses Universitaires de France.

[Mauger 98] : Mauger S., 1998, *Réduction sémantique des énigmes, Essai sur le rôle de l'automate dans l'interprétation des messages obscurs*, Cahiers du GREYC, Université de Caen, à paraître.

[Meillet 52] : Meillet A., 1952, *Linguistique historique et linguistique générale*, Paris, Klincksieck.

[Mel'cuk 86] : Mel'cuk I., 1986, *Dictionnaire explicatif et combinatoire du français contemporain*, Presses de l'université de Montréal, Montréal, Québec.

[Mettinger 94] : Mettinger A., 1994, *Aspects of Semantic Opposition in English*, Oxford, Clarendon Press.

[Michaux 66] : Michaux H., 1966, *L'espace du dedans*, Paris, Gallimard.

[Milner 82] : Milner J. C., 1982, *Ordres et raisons de langue*, Paris, Editions du Seuil.

[Milner 89] : Milner J. C., 1989, *Introduction à une science du langage*, Paris, Editions du Seuil.

[Minsky 75] : Minsky M., 1975, *A framework for representing knowledge,* In P. Winston ed., *The psychology of Computer Vision*, New-York, McGraw-Hill.

[Montague 70] : Montague R., 1970, *English As a Formal Langage,* in R. Thomason, *Formal Philosophy, Selected Papers of Richard Montague*, New-Haven, Yale University Press.

[Morris 39] : Morris C. W., 1939, *Foundations of the theory of signs,* In *Encyclopaedia of unified science*, Neurath O., Carnap R., Morris C.W. (eds.), University of Chicago Press.

[Mugur-Schächter 89] : Mugur-Schächter M., 1989, *Esquisse d'une méthode générale de conceptualisation relativisée,* In "Arguments pour une méthode, autour de Edgar Morin", Paris, Editions du Seuil.

[Nicolle 96] : Nicolle A., 1996, *L'expérimentation et l'intelligence artificielle,* Intellectica, n° 22, p. 9-19, Association pour la Recherche Cognitive (ARC).

[Nicolle & al. 97] : Nicolle A., Beust P., 3-5 Septembre 1997, *Anadia, un analogue machine de la mémoire,* Workshop "Les modèles de représentation : quelles alternatives ?", Association Ferdinand Gonseth, Neuchâtel.

[Nicolle & al. 98] : Nicolle A., St Dizier De Almeida V., 1998, *Vers un modèle des interactions langagières,* In *Analyse et simulation de conversations : de la théorie des actes de discours aux systèmes multiagents*, S. Delisle, B. Chaïb-Draa et B. Moulin (eds.), InterÉditions, Lyon.

[Peirce 78] : Peirce C. S., 1978, *Écrits sur le signe*, rassemblés, traduits et commentés par Gérard Deledalle, Paris, Seuil.

[Piaget 23] : Piaget J., 1923, *Le langage et la pensée chez l'enfant*, Neuchâtel - Paris, Delachaux et Niestlé.

[Pierrel & al. 97] : Pierrel J. M., Romary L., 1997, *Quelles références dans le dialogue homme-machine?,* in *Machine, Langage et Dialogue,* Figures de l'interaction, L'Harmattan, Paris.

[Ploux & al. 98] : Ploux S., Victorri B., 1998, *Construction d'espaces semantiques à l'aide de dictionnaires de synonymes,* à paraître dans T.A.L.

[Poirier 96] : Poirier C., 1996, *Conception et réalisation d'un analyseur sémantique de documents composites (textes et schémas),* Thèse de l'Université de Caen.

[Pottier 73] : sous la direction de Pottier B., 1973, *Le langage,* Les encyclopédies du savoir moderne, Paris, RETZ.

[Pottier 74] : Pottier B., 1974, *Linguistique générale. Théorie et description,* Paris, Klincksieck.

[Pottier 92] : Pottier B., 1992, *Sémantique générale,* Paris, Presses Universitaires de France.

[Quillian 68] : Quilian M., 1968, *Semantic memory*, in M. Minsky : *Semantic Information Processing*, Cambridge, MIT Press, p. 216-270.

[Raccah 96] : Raccah P. Y., Octobre 1996, *De la sémantique théorique à la terminologie*, Third European Congress on System Science, Rome.

[Raccah 97] : Raccah P. Y., 1997, *L'argumentation sans la preuve : prendre son biais dans la langue*, Cognition et Interaction, volume 2, n° 1-2, Nancy.

[Rastier 87] : Rastier F., 1987, *Sémantique interprétative*, Paris, Presses Universitaires de France.

[Rastier 91] : Rastier F., 1991, *Sémantique et recherches cognitives*, Paris, Presses Universitaires de France.

[Rastier 92] : Rastier F., 1992, *La sémantique unifiée. Aide-mémoire*, notes et documents du LIMSI (Université de Paris X), n°92-3.

[Rastier & al. 94] : Rastier F., Cavazza M., Abeillé A., 1994, *Sémantique pour l'analyse*, Paris, Masson.

[Reinhart 76] : Reinhart T., 1976, *The syntactic domain of anaphora*, Thèse de Ph.D., Departement of foreign literature and linguistics, MIT, Cambridge, MA.

[Sabah 90] : Sabah G., 1990, *L'intelligence artificielle et le langage, représentation des connaissances*, Paris, Hermes.

[Sabah & al. 93] : Sabah G., Briffault X., 1993, *Caramel : a Step towards Reflexion in Natural Language Understanding systems*, In Actes IEEE International Conference on Tools with Artificial Intelligence, Boston, p. 258-265.

[St Dizier De Almeida 96] : St Dizier De Almeida V., 1996, *Un modèle théorico-empirique pour la conception d'un système informatique d'assistance interactif*, Thèse de psychologie de l'Université Nancy 2.

[St Dizier De Almeida &. al. 98] : St Dizier De Almeida V., Beust P., 19-24 Juillet 1998, *Catégorisation de connaissances sémantico-pragmatiques : le cas des verbes assertifs*, Communication affichée au 6ième Congrès International de Pragmatique, Reims.

[Sallantin 97] : Sallantin J., 1997, *Les agents intelligents. Essai sur la rationalité des calculs.* Paris, Hermes.

[Saussure 1915] : Saussure F. de, 1915, *Cours de linguistique générale*, Paris 1986, Ed. Mauro-Payot.

[Schank 72] : Schank R., 1972, *Conceptual dependency : a theory of natural language understanding*, Cognitive psychology, vol. 3, p. 552-631.

[Searle 72] : Searle J. R., 1972, *Les actes de langages,* Paris, Hermann.

[Shannon & al. 49] : Shannon C. E., Weaver D. W., 1949, *The mathematical theory of communication,* University of Illinois Press.

[Sowa 84] : Sowa J. F., 1984, *Conceptual Structures. Information Processing in Mind and Machine,* Addison Wesley, Reading, MA.

[Tanguy 97] : Tanguy L., 1997, *Traitement automatique de la langue naturelle et interprétation : contribution à l'élaboration d'un modèle informatique de la sémantique interprétative,* Thèse de l'Université de Rennes 1.

[Tesnière 59] : Tesnière L., 1959, *Elements de syntaxe struturale,* Paris, Ed. Klincksieck (dernière édition 1982).

[Thlivitis 98] : Thlivitis T., 1998, *Sémantique Interprétative Intertextuelle : assistance informatique anthropocentrée à la compréhension des textes*, Thèse de l'Université de Rennes 1.

[Trognon & al. 92] : Trognon A., Brassac C., 1992, *L'enchaînement conversationnel*, Cahiers de Linguistique Française, vol. 13, p. 77-107.

[Tyvaert 95] : Tyvaert J. E., 20-22 Septembre 1995, *Continuité substantielle et discretisation substantivale,* Colloque Anaphore et Référence, Nancy, CRIN.

[Van Dijk 77] : Van Dijk T.A., 1977, *Text and Context. Explorations in the Semantics and Pragmatics of Discourse,* Londre, Longman.

[Vergne 94] : Vergne J., 1994, *A non-recursive sentence segmentation applied to parsing of linear complexity in time,* International Conference on New Methods in Langage Processing (NeMLaP), Manchester.

[Vernant 98] : Vernant D., 1998, *Du discours à l'action,* Paris, Presses Universitaires de France.

[Victorri & al. 96] : Victorri B., Fuchs C., 1996, *La polysémie,* Paris, Hermes.

[Victorri 98] : Victorri B., 1998, *La construction dynamique du sens : un défi pour l'intelligence artificielle,* RFIA'98.

[Vivier 92] : Vivier J., 1992, *Faire et dire ce qu'il faut faire pour ... Physique naturelle et discours : Projet d'étude developpementale.* Diplôme d'habilitation à diriger des recherches, Université de Caen.

[Vivier 93] : Vivier J., 1993, *Comperobot, when Dialogue Enlarges Dialogue Capacity ?,* Second European Congress on System Science, Prague.

[Vivier & al. 94] : Vivier J., Jacquet D., 1994, *Étude comparative des stratégies d'explication d'une tâche de construction chez des enfants en classe de perfectionnement et en cycle normal,* In Deleau M. et Weil-Barais A. (Eds.), *Le développement de l'enfant, approches comparatives,* Paris, Presses Universitaires de France, p. 165-174.

[Wittgenstein 61] : Wittgenstein L., 1961, *Tractatus logico philosophicus,* Paris, Gallimard.

[Yagi 84] : Yagi T., 1984, *Le Mahabhasya ad Panini,* Paris.

INDEX DES AUTEURS

—D—

—E—

—F—

—G—

—H—

—I—

—J—

—K—

—L—

—M—

—N—

—P—

—Q—

—R—

—S—

—T—

—V—

—W—

—Y—

INDEX