Un analogue de la mémoire pour un agent logiciel interactif

Anne Nicolle, Pierre Beust, Vincent Perlerin

GREYC CNRS UMR 6072 & Pôle Modescos de la MRSH
Université de Caen
Boulevard Maréchal Juin
14032 Caen Cedex, France
{anne.nicolle, pierre.beust, vincent.perlerin}@info.unicaen.fr

Résumé

À partir d'une synthèse sur les modèles de la mémoire d'une part et d'un modèle différentiel et componentiel de la sémantique des langues d'autre part, nous proposons un dispositif de tables de catégorisation comme analogue de la mémoire sémantique pour un agent logiciel. Après avoir décrit ce modèle, nous présentons l'état actuel de son implantation en machine et quelques applications expérimentales qui en ont été faites.

Abstract

From a synthesis on memory models and a differential and decomponential model of natural languages semantics, we suggest a categorization tables device used as a semantic memory equivalent for a software agent. In a first part, we describe this model before, in a second part presenting its actual implementation and some of its experimental applications.

1. Introduction

Cet article présente un modèle de la mémoire fondé sur l'observation des interactions sociales. Ce modèle permet de concevoir un analogue mémoire pour des agents logiciels, pour leur donner la capacité qu'ont les êtres vivants de distinguer les choses entre elles à partir de leur expérience sensible, alors même qu'ils n'ont pas eux-mêmes cette expérience. Cette capacité va se construire à partir des interactions langagières. Elle permet une co-référenciation des choses présentées dans le discours, analogue à ce qui peut être mis dans le terrain commun au cours d'un dialogue. Ce modèle n'est pas l'analogue d'une représentation mentale, il n'est pas non plus le support d'une représentation scientifique ou ontologique du monde, il vise à opérationnaliser la notion linguistique de *valeur* chez Saussure (Saussure, 1915). Fondée sur ce modèle, la mémoire d'un agent cognitif forme un système où chaque entité conceptuelle connue, au lieu d'avoir une représentation en propre, a une place qui la distingue des autres, et où chaque entité inconnue peut trouver place, moyennant une transformation plus ou moins profonde du système. Ce modèle de la mémoire s'insère dans le développement d'un modèle d'agent logiciel pour le dialogue humains/machines et pour l'interaction multi-agents qui a été initialement présenté dans (Nicolle et Saint-Dizier, 1998).

Pour introduire le modèle et le distinguer des conceptions courantes, nous rappellerons dans une première partie quelles conceptions de la mémoire ont été proposées en Intelligence Artificielle (IA) (Section 2.), et plus précisément comment ces conceptions ont évolué à partir de modèles biologiques et psychologiques, puis à partir de modèles sociaux. Nous détaillerons dans la partie suivante (Section 3.), les fondements d'un analogue machine de la mémoire organisée par le langage. Le modèle sera décrit en détail dans la 4ème section. Nous exposerons enfin une utilisation de ce modèle dans le cadre d'une modélisation de la sémantique des langues pour un agent logiciel. En conclusion, nous verrons en quoi ce modèle, profondément inscrit dans l'interaction personnes-machines, ouvre de nouvelles perspectives de recherche pour les agents logiciels.

2. Les conceptions de la mémoire en intelligence artificielle

Des modèles de la mémoire sont utilisés, le plus souvent implicitement, dans la manière de définir les bases de connaissances ou les bases de cas. Leur présence dans l'ergonomie des interfaces personnes-machines à travers la structure des présentations et des interprétations, ou des modélisations des utilisateurs, montre l'importance du travail d'explicitation des caractéristiques de la mémoire pour l'ingénierie du logiciel.

2.1. Il y a mémoire et mémoire

Le mot *mémoire* est utilisé dans de nombreux sens suivant les êtres auxquels il s'applique et suivant les contextes. Il a une dimension de stockage de l'information et une dimension d'activité de

représentation, de mémorisation et de remémoration (Tiberghien, 1997). Les êtres vivants, les humains qui vivent en société et les machines ont une mémoire dont les caractéristiques sont différentes.

1) La mémoire des êtres vivants dans leur ensemble permet leur adaptation au milieu. Elle est basée sur des mécanismes de catégorisation des situations rencontrées, en fonction de leur impact sensoriel et des variations qui résultent de l'interaction de l'être avec son milieu dans les situations semblables. La caractéristique principale de cette mémoire est d'être un processus de catégorisation-remémoration plutôt qu'un dispositif matériel de conservation des données. Il n'y a pas de moyen d'accès à un lieu de la mémoire en dehors de situations de rappel, contrairement à ce qui se passe avec les machines. La remémoration peut être décrite comme la reconstruction d'une expérience antérieure à partir de traits discriminants du contexte. La mémoire d'un individu se construit d'abord par l'expérience sensible, mais aussi, chez les animaux évolués, par les communications entre individus : observation et imitation.

2) La mémoire humaine est plus complexe. Elle a une composante individuelle (de même nature que la mémoire des autres êtres vivants) et une composante collective. La mémoire individuelle est plongée dans la mémoire collective par le partage de mécanismes de représentation comme le langage, les dessins, les cartes de géographie, l'écriture. Les discours ou les dessins permettent de rendre compte de la mémoire hors situation. Ces mécanismes permettent la transmission par le dialogue de mythes, de récits, de règles de conduite qui fondent les sociétés humaines. De ce fait, la mémoire des individus humains est plus élaborée que celle des autres êtres vivants. Elle met en jeu des processus de catégorisation, mais aussi des processus de généralisation et des processus d'abstraction qui lui donnent une dimension supplémentaire.

3) Pour les ordinateurs, le mot mémoire a d'abord été utilisé au sens de dispositif matériel permettant de stocker des données, de manière permanente ou non, et de les retrouver par leur adresse (disques, bandes, mémoire centrale...). La « mémoire » ainsi conçue est passive et elle est opposée aux programmes qui exploitent ces données, comme les systèmes de gestion de bases de données. L'aspect procédural de ces programmes fait qu'ils sont plus proches de la mémoire des êtres vivants que ce qu'on appelle la mémoire des machines, mais ce sont seulement des processus de stockage et de restitution des données à partir d'une demande extérieure. Il n'y a aucune dimension de catégorisation ou de reconstruction à partir de besoins propres comme chez les êtres vivants.

Actuellement, on a tendance à voir les machines informatiques comme un nouveau support de la mémoire partagée, avec les bases de données, les bases de connaissances et les hypertextes, donc à avoir un point de vue social plutôt qu'un point de vue individuel sur la mémoire des machines. On commence tout juste à imaginer les utilisations possibles de ces mémoires externes à l'homme, partageables et actives (Lévy, 1990) (Bourdon, 1992).

Nous allons décrire maintenant les modèles informatiques de la mémoire. Nous verrons à travers la présentation de l'évolution des modèles statiques de représentation (Section 2.2.) aux modèles dynamiques de mémorisation-remémoration (Section 2.3.) jusqu'aux modèles de mémoire sociale partagée (Section 2.4.), comment nous avons été amenés à construire un modèle interactionnel de la mémoire.

2.2. Les modèles de la mémoire proposés en IA classique

En IA classique, on ne s'intéresse pas spécialement à bien gérer la mémoire des machines au sens premier d'aire de stockage, mais à comprendre les processus de mémorisation et de remémoration. On propose des méthodes de représentation des connaissances, par des graphes conceptuels par exemple (Sowa, 1984), pour conserver et évoquer les choses et leurs relations. Ces processus sont basés sur le contenu et le sens des données plutôt que sur leur syntaxe ou leur adresse. La compréhension des phénomènes de mémoire aide à concevoir des modèles de la mémoire pour les mettre en œuvre dans des logiciels utilisant des connaissances. Ces logiciels vont simuler - ou plus souvent aider - l'activité des hommes dans des tâches complexes comme le diagnostic, la classification, la conception, la décision et autres tâches intelligentes.

Dans les modèles computationnels de la mémoire individuelle plusieurs caractéristiques opposées peuvent être mises en évidence, nous les examinerons à tour de rôle en montrant l'influence que ces notions ont eues dans les modèles.

2.2.1. Mémoire de travail - Mémoire à court terme - Mémoire à long terme

Cette distinction a été posée par les psychologues pour expliquer les variations dans les compétences des individus. La mémoire de travail correspond aux choses présentes simultanément à l'esprit au cours d'une tâche. Le nombre de choses qui peuvent être examinées en même temps est très petit (5 ou 7 éléments) et cette mémoire est très volatile. Les humains suppléent souvent à ces insuffisances par une organisation des perceptions en blocs caractéristiques repérables comme un tout (faire une

phrase avec tous les mots dont on doit se rappeler) ou en utilisant des mémoires externes, comme une feuille de papier pour écrire ou dessiner. Ce qui est perçu comme **un** élément dans une situation varie beaucoup avec la familiarité de la situation. Des études sur les joueurs d'échecs ont bien montré cette différence interindividuelle entre débutants, joueurs de clubs ou experts du point de vue de la mémorisation des situations. La mémoire à court terme nous permet de gérer le temps présent : on se souvient de ce qu'on a fait et de ce qu'on a à faire pour la période en cours. Seules les choses importantes seront mémorisées plus longtemps et seulement si elles sont répétées. La mémoire a long terme correspond à ce qu'on retient : elle a une capacité pratiquement illimitée, car plus on apprend et plus on peut apprendre.

La distinction entre mémoire à court terme et mémoire à long terme a été introduite en IA avec les systèmes de productions (base de faits ≠ base de règles) dont CLIPS[1] est un exemple moderne. Les langages logiques comme Prolog ne font pas cette distinction entre les énoncés car faits et règles sont tous sous la même forme de clause et l'usage est de les mettre indifféremment dans les mêmes paquets. Un système de productions comporte 3 éléments :

• une base de faits qui représente la situation courante (mémoire à court terme)

• une base de règles (ou de règles de production) qui décrivent les connaissances permettant de traiter une famille de problèmes (mémoire à long terme). Elles sont de la forme :

prémisses -> conclusions
ou
situation -> actions

• un mécanisme d'inférence pour apparier la situation examinée avec les règles et déclencher les règles instanciées choisies.

La base de faits sert à décrire les problèmes, les situations intermédiaires et les solutions. Elle n'est pas limitée à quelques éléments, comme la mémoire de travail des êtres humains, car la simulation n'est pas l'objet de ces systèmes. La base de règles ou base de connaissances sert à conserver les méthodes et les règles d'inférence utiles pour résoudre les problèmes. Les règles peuvent être déclenchées quand leur partie droite s'apparie avec l'état de la mémoire de travail. La mémoire de travail est alors modifiée par la partie gauche de la règle : l'activité mémoire est donc vue comme une pure activité d'appariement entre des modèles et des situations. Ce modèle de raisonnement avec une représentation des connaissances par des règles a permis de construire des systèmes experts qui sont des logiciels d'aide à des tâches complexes, comme le diagnostic, la conception, l'interprétation de résultats expérimentaux. Ces tâches complexes nécessitent de grandes bases de connaissances, parfois plusieurs milliers de règles et nécessitent un mécanisme d'inférence non déterministe, car il faut choisir à tout moment la règle la plus intéressante à appliquer si on ne veut pas crouler sous les faits nouveaux. Anderson (Anderson, 1983) a défendu la thèse de l'analogie entre les systèmes de productions et la mémoire sémantique humaine, mais cette analogie a été remise en cause par l'observation des multiples liens qui se tissent entre nos connaissances, qui n'ont pas d'analogue dans les systèmes de productions. On peut aussi remarquer que la résolution d'un problème par un système de productions n'a pas de conséquences sur la mémoire à long terme du système alors que lorsque nous avons résolu un problème, nous nous souvenons de la démarche qui a réussi et nous tentons de l'utiliser dans des cas analogues.

2.2.2. Mémoire épisodique - Mémoire sémantique

Ce sont deux aspects de la mémoire à long terme. La mémoire épisodique est la mémoire des événements, avec les sensations, les faits, les dates, les lieux qui leur sont associés. On l'appelle aussi la mémoire autobiographique. Elle contient les souvenirs. Elle est souvent peu fiable chez l'homme comme le montrent les contradictions dans les témoignages de personnes de bonne foi. La mémoire sémantique est la mémoire de la structure des choses, de leurs relations, de leurs fonctions et de leur genèse. Elle contient le langage, les codes et les règles, et les connaissances. Les principaux modèles de la mémoire sémantique : systèmes de productions et réseaux sémantiques, ont été proposés vers la fin des années 70 avec l'augmentation de la capacité de stockage (ou de mémoire) des machines. À la même époque, les scripts et les scénarios ont tenté de modéliser l'organisation de la mémoire épisodique.

Les réseaux sémantiques ont eu d'abord pour objet de représenter la mémoire sémantique des êtres humains sous forme de graphes dont les nœuds représentent des concepts et dont les arcs représentent les relations entre ces concepts (Figure 1). Les principales relations sont les relations de généralisation-spécialisation entre les concepts, les relations entre un objet complexe et ses parties, et les relations structurelles entre certains concepts comme les jours de la semaine ou les couleurs de l'arc-en-ciel. Les algorithmes de parcours dans ces graphes, à partir d'instances de concepts (comme "aide en ligne de Anadia v1.0" et "Anadia.java" sur l'exemple – Figure 1), sont l'analogue des activations de la mémoire (Falhman, 1979).

[1] www.ghgcorp.com/clips/CLIPS.html

Les réseaux sémantiques ont été utilisés pour améliorer les systèmes experts en associant des règles aux concepts au lieu de les mettre en vrac. Il est alors possible de ne s'intéresser qu'aux règles qui ont quelque chose à voir avec le problème en cours. Les hiérarchies de généralisation entre les concepts, qui représentent des connaissances ontologiques, sont le mode d'organisation le plus souvent utilisé. Elles permettent de factoriser les connaissances, en les associant au concept le plus général pour lesquelles elles sont pertinentes et en les évoquant pour les concepts plus spécialisés par héritage. Elles permettent ainsi de retrouver plus rapidement les règles intéressantes car au lieu d'examiner toutes les règles pour voir si elles s'appliquent à un objet, on peut n'examiner que celles qui sont associées à la hiérarchie d'héritage de l'objet considéré. Par exemple, "Anadia.java", qui est un programme, a aussi les propriétés des fichiers texte et des objets informatiques (Figure 1). Par contre, il n'a pas les propriétés des documents ou des événements, qui ne sont même pas évoquées. Dans les systèmes de productions purs où la base de connaissances n'est pas structurée, on ne peut pas présélectionner ainsi les propriétés pertinentes, il faut toutes les examiner à chaque cycle d'inférence.

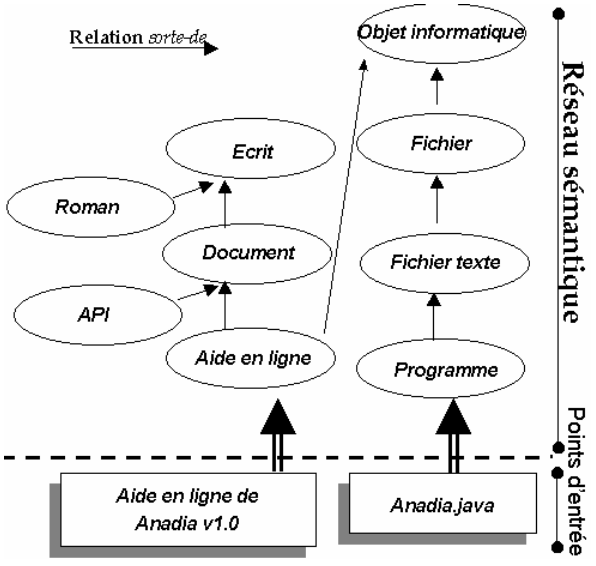

Figure 1— *Exemple de réseau sémantique*

Les scripts ou schémas (Minsky, 1975) ont pour objet la représentation synthétique de la mémoire épisodique. L'hypothèse sous-jacente est que les situations ne sont pas stockées de manière complète et à la suite les unes des autres mais qu'elles sont organisées en une partie générique : les schémas de situations fréquemment rencontrées, et une partie spécifique à une situation donnée. Une situation particulière est donc appariée à un ou plusieurs schémas connus et mémorisée en terme de schémas + particularités. En particulier, il existe des schémas plus ou moins généraux organisés les uns par rapport aux autres comme dans l'exemple ci-dessous (Figure 2) où une réunion, une expérience ou un enregistrement sont des événements particuliers. Les schémas mémorisés peuvent alors être évoqués pour résoudre des problèmes, pour comprendre des histoires, en complétant les descriptions données explicitement par référence aux situations les plus courantes.

Dans la conception des schémas, la principale difficulté est le choix des informations pertinentes à mémoriser. Chaque situation peut donner lieu à une infinité de descriptions dont la pertinence et le niveau de détail dépendent du contexte et il faut bien choisir celles qui seront considérées comme assez générales pour faire partie d'un schéma. Lors de l'utilisation des schémas, le principal problème est l'évocation du ou des schémas pertinents. Contrairement à l'appariement entre les règles de production et la base de faits qui est toujours stricte, la comparaison entre un schéma et une situation doit accepter des différences si elles sont moins significatives que les ressemblances.

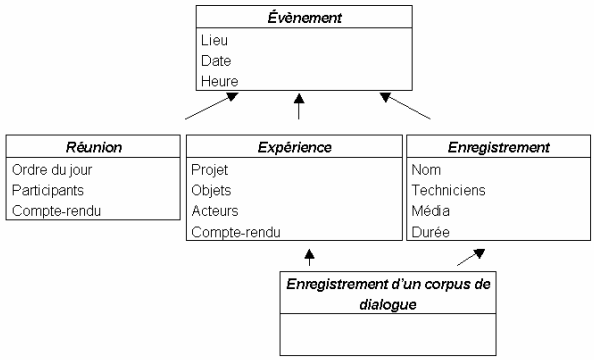

Figure 2— *Exemple de réseau de schémas*

2.2.3. Mémoire déclarative - Mémoire procédurale

Les souvenirs et les connaissances sont connus explicitement par le sujet, il peut en parler et les acquérir par le discours. Ils correspondent à la mémoire déclarative dont l'équivalent pour les machines utilise les représentations présentées ci-dessus. Les habitudes ou les savoir-faire sont largement implicites, le sujet n'en a pas toujours conscience et ils sont acquis par imitation plutôt que consciemment. Cette mémoire procédurale a comme équivalent pour les machines le logiciel système, les compilateurs et les programmes qui vont exécuter des algorithmes ou interpréter des connaissances.

2.3 La mémoire comme activité dans la « nouvelle IA »

La mémoire peut être vue comme un dispositif matériel où seraient stockés les souvenirs et les connaissances. On a cherché ainsi à isoler dans le cerveau l'endroit où seraient stockées de telles

informations sans obtenir de résultats probants. Cette hypothèse impliquerait que la capacité de mémorisation soit limitée alors que plus on apprend et plus on est capable d'apprendre. Elle considère la mémoire de manière passive et elle ne permet pas d'interpréter les difficultés à se souvenir de manière fiable. Elle ne rend pas compte des aptitudes à la catégorisation. Les conceptions plus récentes de la mémoire, comme la théorie d'Edelman (Rosenfield, 1989) la décrivent comme une activité. Cette activité se déploie dans deux directions, la mémorisation et la remémoration. La mémorisation met en jeu une activité de catégorisation des perceptions, des sensations et les interprétations en terme de relations qui en découlent. La remémoration est une activité de reconstruction à partir de déclencheurs trouvés dans le contexte. Ce modèle est compatible avec les imprécisions et les confusions observées dans les souvenirs.

À la suite des théories d'Edelman sur la mémoire comme activité plutôt que comme support de données, il y a eu déplacement du focus de l'IA des représentations vers les processus. Les processus de catégorisation et de classement des situations rencontrées sont à la base de la mémorisation. Les processus de remémoration permettent la reconstruction des souvenirs à partir d'évocations. Ils ont été étudiés de manière systématique pour en faire des modèles computationnels. Les réseaux connexionnistes sont une des voies possibles pour implanter des mémoires procédurales actives, mais qui ne prennent pas en compte les aspects d'explicitation et d'interaction. Alors que les systèmes symboliques s'intéressent plutôt à la mémoire déclarative, les modèles à objets, l'apprentissage par la découverte et le raisonnement par cas sont des exemples de modélisation et d'utilisation de l'activité mémoire.

2.3.1. Les modèles à objets

Les modèles informatiques de représentation par objets encapsulent dans un "objet" des données et les procédures pour les manipuler. Ils sont beaucoup plus proches du fonctionnement de la mémoire des êtres vivants que les modèles de la génération précédente, quand il y avait d'une part des modèles fonctionnels (où tout est calcul sans modification de données globales) et d'autre part des modèles logiques dits déclaratifs (où tout est données, y compris les règles de raisonnement, et où les processus de calcul sont systématiques et cachés).

À l'origine de ces nouveaux modèles à "objets", il y a eu la proposition de Minsky de décrire les objets, les concepts, les situations, les événements des scripts par des "frames" (Minsky, 1975) regroupant dans une même unité une description et des procédures pour remplir et utiliser cette description. Cette idée a été implantée ensuite sous des formes variées dans de très nombreux systèmes dont Airelle (Adam-Nicolle, 1990). Chaque unité est décrite par un nom et une liste d'attributs. Les attributs sont décrits par des facettes dont certaines sont des valeurs (description, type, format, valeur, valeur par défaut...) et d'autres sont des procédures réflexes qui se déclenchent lors des accès aux attributs pour tester les valeurs rentrées, pour chercher une valeur manquante, pour propager des mises à jour. On obtient avec ce principe des systèmes réagissant en chaîne à toute modification ou consultation de données, sans que ces modifications soient contrôlées de manière centrale. Certaines unités représentent des classes d'objets ayant même structure et même comportement, d'autres unités représentent des individus de ces classes. Les langages à objets de la famille Smalltalk (Golberg & Robson, 1984) reprennent certains principes des frames, comme l'héritage et l'instanciation, mais pas les réflexes, qu'on peut toutefois simuler en Java avec la gestion d'événements.

2.3.2. L'apprentissage par la découverte

Pour simuler l'apprentissage des êtres humains, une voie de recherche, explorée par Lenat dans les années 75-80 consiste à construire un modèle minimal extensible et le faire évoluer par apprentissage à partir des problèmes qu'il résout. Pour cela, le système doit savoir généraliser les renseignements qu'il tire de ses essais et de ses erreurs, quitte à les spécialiser plus tard si des contre-exemples sont rencontrés. Il doit savoir construire des catégories pertinentes et les raffiner. Le système AM (Lenat, 1982) découvre des concepts mathématiques à partir de concepts plus élémentaires et EURISKO (Lenat, 1983) découvre la bonne façon de jouer à des jeux de bataille navale. Ensuite le système SOAR (Laird et al., 1987) a repris ce principe d'apprentissage par l'expérience en résolution de problèmes. Il a mis en évidence la notion de "chunk", description partielle d'une situation qui résume les caractéristiques utiles pour l'apprentissage d'un comportement dans une situation donnée et il a obtenu des résultats intéressants sur le jeu de Taquin et d'autres problèmes standards. Au fur et à mesure de leur travail, ces systèmes modifient leur base de règles et d'heuristiques. Dans les deux cas, après une phase où les résultats de l'apprentissage sont indéniables, ils plafonnent, faute de savoir réorganiser leur mémoire. On peut donc penser que seule une des phases de l'apprentissage a été modélisée et qu'un apprentissage plus profond qui, pour les humains est peut-être révélé par le rêve, est encore à étudier. Pour cela, il faut concevoir des systèmes qui puissent prendre leurs représentations, leurs connaissances et leur propre fonctionnement comme objet d'étude. C'est le cas pour les systèmes réflexifs dont les bases ont

été posées par (Maes et Nardi, 1987) et par (Pitrat, 1990) avec l'étude de la métaconnaissance.

2.3.3. Le raisonnement à partir de cas

Les premiers modèles du raisonnement des experts ont été des modèles du raisonnement déductif, comme les systèmes de production présentés ci-dessus. Les difficultés de réalisation des systèmes experts, en particulier le recueil de l'expertise, ont montré que le raisonnement d'un expert n'est pas purement déductif, mais qu'il s'appuie sur les ressemblances globales de la situation du problème avec des situations connues (raisonnement par analogie). Ils ont aussi montré que l'expertise ne pouvait pas être décrite complètement dans l'abstrait mais qu'elle pouvait être révélée par la parole au cours de la résolution de problèmes (Brixhe et al., 1994)

Par exemple, les premiers systèmes de conception assistée par ordinateur étaient des logiciels de dessin industriel. Les nouveaux systèmes sont des systèmes intelligents d'aide à la conception. Ils utilisent souvent le raisonnement par cas, en interaction avec l'expert qui peut modifier les solutions proposées. Le recueil d'expertise est progressif et situé, chaque problème résolu en interaction entre la machine et l'expert améliore les connaissances de la machine pour les problèmes suivants. Dans le domaine de la conception, en Architecture par exemple, l'expert qui conçoit un plan cherche d'abord à ramener la situation qui lui est proposée à des cas qu'il a déjà rencontrés, à se remémorer les solutions qu'il avait alors trouvées, puis à les adapter au nouveau problème (Guéna, 1997). Le raisonnement par cas cherche à automatiser ce processus. La machine va simuler les processus de création d'une solution à partir de celles qu'il a mémorisées pour des problèmes analogues. Pour réaliser une telle machine, il faut déterminer comment les cas seront organisés en mémoire et comment les cas pertinents seront évoqués à partir de la donnée d'un nouveau problème. On utilise alors des mécanismes de catégorisation hiérarchique à partir des attributs les plus importants. Lorsqu'un nouveau cas est traité, la machine propose une solution par analogie avec d'anciennes solutions. Le concepteur humain l'examine et peut la modifier. Puis il valide la solution quand il est satisfait. Le problème et sa solution sont alors examinés par la machine pour déterminer si la solution est suffisamment originale par rapport aux cas déjà connus pour qu'il soit intéressant de se la rappeler. Si elle est suffisamment originale, le cas est stocké dans la base de cas pour être réutilisé dans des situations analogues. Dans ce modèle de raisonnement, il y a donc prise en compte de l'interaction dans la résolution de problèmes et évolution de la mémoire à long terme en fonction des problèmes résolus.

2.4. Mémoire individuelle, mémoire sociale

La mémoire comme activité individuelle est caractéristique du vivant, elle peut être simulée sur machine dans le paradigme de la vie artificielle, en amorçant un système autonome qui apprend par interaction avec son environnement. En Intelligence Artificielle classique, les machines ont été vues d'abord comme un moyen de simuler ou de remplacer la mémoire et les mécanismes de raisonnement de l'homme en tant qu'individu, dans une optique de compétition. Les modes de représentation des données et des connaissances qui ont été développés dans ce but ne sont pas purement descriptifs, ils comportent des mécanismes puissants de recherche, de présentation et de calcul.

Dans les systèmes multi-agents (Ferber, 1995), la question de la mémoire concerne les agents et les collectifs. Elle doit faciliter l'action et l'interaction, permettre l'instauration d'un terrain commun pour les échanges en vue de la coopération et de la négociation. Les systèmes multi-agents ont posé la question de la mémoire partagée, de son inscription dans l'environnement physique et de son inscription dans des organisations sociales avec des conventions, des règles et des lois.

Si la question des représentations mentales comme préalable à l'intelligence a été remise en cause (Brooks, 1991), il existe bien des représentations attestées : des signes, des écrits, des dessins, des plans. Ce sont des représentations externes à l'individu, dont le fonctionnement passe par des codes sociaux et dont chaque individu va faire l'apprentissage. Cette notion « d'objet intermédiaire » a fait l'objet de travaux récents en psychologie sociale (Grégori, 1999). Cette dimension sociale de la mémoire humaine passe par des systèmes sémiotiques, principalement par le langage, et elle ne peut pas se comprendre comme émergeant de calculs d'individus isolés à partir de leur boucle sensori-motrice. Cette mémoire est partagée à travers les langues, les récits et les mythes, les institutions sociales (Castoriadis, 1975) mais aussi à travers des représentations graphiques comme les dessins, les cartes de géographie, les plans d'architecture. Les traces produites par cette instrumentation de la mémoire sont observables alors que les représentations mentales ne le sont pas. Nous pouvons étudier les traces de cette catégorisation et de ses propriétés dans les objets relevant des systèmes sémiotiques, et en faire des théories réfutables car basées sur des observations partageables et pas sur l'auto-observation où la part de l'imaginaire ne peut pas être contrôlée. Après le cinéma, la radio et la télévision, le Web et plus généralement le multimédia amplifient aujourd'hui l'instrumentation de la mémoire et sa dimension

sociale. Et les mémoires d'entreprises sont actuellement l'objet de recherches visant à développer des technologies spécifiques pour accompagner de nouvelles pratiques sociales (Assadi, 1998).

Le modèle que nous proposons, nommé Anadia en référence à (Coursil, 2000) s'appuie sur une théorie de la mémoire compatible avec les théories de la mémoire de l'individu comme activité (Rosenfield, 1989). En tant qu'activité de catégorisation/ reconstruction, elle enregistre ce qui fait différence et seulement cela. Elle s'appuie aussi sur la mémoire sociale dont les traces observables sont des documents variés, en particulier des textes en langue naturelle. Le modèle donne naissance à un dispositif de mémorisation externe et partageable.

3. Principes et fondements d'un analogue machine de la mémoire

On appelle analogue machine d'une compétence humaine, par opposition à une simulation qui serait tenue à la vraisemblance biologique, un modèle fonctionnel qui s'inspire du fonctionnement humain et l'adapte aux compétences propres des machines. Un analogue peut faire mieux certaines choses, moins bien d'autres, en tout cas, il le fait autrement. Le modèle de la mémoire que nous proposons pour donner aux agents logiciels un analogue machine de la mémoire s'appuie sur l'observation de la construction d'un terrain commun dans un dialogue. La mémoire d'un agent comporte un noyau extensible de compétences à dialoguer qui se développe par son activité propre, en interaction avec des interlocuteurs humains, grâce à des mécanismes de référenciation partagée. Pour l'implanter, il faut donc définir une structure mémoire et des opérations sur cette structure qui soient fonctionnellement analogues aux nôtres. Nous en attendons que le dialogue entre des personnes et des agents logiciels munis de cette mémoire devienne plus naturel.

Une analyse de dialogues techniques en présence des objets et des instruments concernés, comme nous l'avons faite dans le projet PIC[2], permet d'observer la mémoire à l'œuvre, en observant comment se construit la co-référenciation aux choses (objets, événements, processus) (Beust et al., 1997). Pour montrer les choses dont on veut parler, pour établir leur *valeur* sémiotique, on les décrit juste assez pour les différencier des choses avec lesquelles elles pourraient être confondues dans la situation. La présentation des choses par la parole dépend des interlocuteurs, de l'activité en cours et de l'histoire de l'interaction. Le sujet entendant peut identifier la chose dont on parle par son activité propre si sa place a été visée de manière adéquate. Ces observations renforcent l'hypothèse que la mémoire est une organisation des différences perçues dans l'interaction avec le monde et avec les autres, et que l'environnement est le support de la reconstruction qui s'opère lors de la référenciation.

Toutes les composantes de la mémoire présentées dans la section 2.2. ne sont pas définies dans ce modèle. Nous pensons actuellement que la mémoire de travail peut être représentée par les objets créés et détruits au cours de l'exécution d'un programme informatique. Nous pensons aussi que le logiciel de base est un analogue acceptable de la mémoire procédurale. Nous proposons un modèle de la mémoire sémantique, qui organise ce qui est invariant dans le temps : les concepts, les structures, les règles, les contraintes, les relations et les fonctions. Nous montrons son articulation avec un modèle de la mémoire épisodique - qui organise ce qui dépend de l'espace et du temps, les états des choses, les événements, l'histoire des processus - défini pour l'instant de façon sommaire par des bases de données. Nous allons d'abord montrer dans les sections suivantes que le noyau d'un analogue de la mémoire sémantique nécessaire à des agents logiciels pour entrer en interaction ne peut être ni un système de représentation de connaissances, ni un système de classification basée sur des mesures, ni une base de données, mais qu'il doit être construit sur une activité langagière.

3.1. Mémoire et classification

Nous allons décrire les principales caractéristiques des deux principaux modes de classification. Nous montrerons ensuite comment ils s'articulent dans un modèle de la mémoire.

1) la classification fondée sur les attributs continus, en termes de ressemblances, par regroupements de notions proches, sans que les frontières soient stables (notion de prototype en psychologie cognitive, méthodes d'agrégation de l'analyse des données).

2) la catégorisation fondée sur les différences, par partition d'une catégorie d'objets en fonction de critères discrets qui s'excluent les uns les autres. La représentation des connaissances telle que nous la mettons en œuvre dans notre modèle, s'appuie sur cette forme de catégorisation.

En analyse des données une classe est un **cluster.** La définition des classes est donnée en terme de

2 Le corpus PIC provient d'une étude expérimentale et une modélisation des processus cognitifs et sociaux de conception distribuée à travers les traces qu'ils laissent dans les dialogues. Dans la situation expérimentale mise en place, trois partenaires ont été invités à se rencontrer pour une séance de travail dont le but était de commencer à concevoir la documentation utilisateur d'un logiciel. Leurs échanges ont été par la suite intégralement retranscrits et analysés. Pour plus d'informations: www.info.unicaen.fr/ ~fgerard/pic/pic.html

distance, donc relativement à des attributs mesurables, ou de ressemblance (Rialle, 1995). Une bonne décomposition minimise les distances entre les objets qui appartiennent à une même classe et maximise les distances entre les objets appartenant à des classes différentes. La mesure est un moyen objectif de décrire les phénomènes. Elle a produit des théories et des instruments de mesure, qui ont fait la force des sciences physiques. Pour qu'une caractéristique C soit mesurable, il faut pouvoir définir, pour tout couple d'objets A et B ayant cette caractéristique :

l'égalité C(A) = C(B)
la somme C(A) + C(B)

On peut mesurer les hauteurs, les distances, les volumes, les poids, les fréquences et les intensités. Pour les choses mesurables, on dispose de bonnes théories mathématiques, avec les théories de la continuité, qui permettent de faire des modèles computationnels. Or tout n'est pas mesurable au sens de la théorie de la mesure. Bien qu'on dise, « mesurer la température », et que ce soit une caractéristique du monde physique, la température n'est pas une mesure parce qu'elle n'est pas additive. On peut s'en rendre compte facilement en comparant les températures sur les échelles Celsius et Fahrenheit : la traduction, l'addition puis la traduction inverse ne donnent pas le même résultat que l'addition directe. On ne peut pas non plus mesurer les événements parce qu'ils sont de nature discrète, on peut seulement les compter. On ne peut mesurer ni les lois, ni les organisations sociales, ni la liberté, ni la justice, ni la douleur, ni la mémoire. Pour les choses qui relèvent du biologique, du mental, du social, on peut mesurer certaines caractéristiques de leur substrat physique ou de leurs productions dans leur dimension physique, mais elles ne se réduisent pas à cette dimension : un livre plus lourd qu'un autre n'en est pas pour autant plus intéressant. Les premières unités de mesure sont le pied, le pouce, le pas, elles partent donc de caractéristiques biologiques, mais leur adoption nécessite des pratiques sociales, comme la négociation d'un accord sur une norme. Pour avoir des théories et des instruments de mesure, il faut une structure sociale organisée, ce qui suppose déjà le dialogue entre les membres de cette société. Les représentations fondées sur la mesure ne peuvent donc pas précéder la co-référenciation dans le dialogue. La mesure suppose que soient déjà partagés le nombre cardinal et le continu.

Les choses qu'on ne peut pas mesurer, on peut cependant les comparer, les classer et les juger. Aristote a proposé dans Organon (Aristote) une catégorisation par genres et différences qui sous-tend la plupart des méthodes de catégorisation postérieures. La représentation sous forme d'arbres de raffinement des genres porte le nom d'arbres de Porphyre à qui on doit cette représentation six siècles plus tard (Figure 5). Remarquons que des attributs continus peuvent être discrétisés par les opérations langagières et servir alors à des catégorisations socialement marquées. Par exemple, dans une école, on pourra distinguer les petits, les moyens et les grands et leur attribuer des caractères distinctifs qui en font des genres : une salle de classe, un maître, un emploi du temps. La couleur, bien que correspondant à une fréquence, est la plupart du temps nommée et par là même discrétisée : elle peut alors servir de base à des distinctions conceptuelles (feux verts, feux rouges...). Il reste que la mesure n'intervient jamais dans la catégorisation, seulement le jugement, et qu'il doit être possible d'énumérer les valeurs distinctes et finies qu'un attribut peut prendre pour qu'il entre dans le processus de catégorisation. Les méthodes de catégorisation ultérieures ont proposé des améliorations et des extensions des arbres de Porphyre sans s'en éloigner pour l'essentiel, mais les interprétations varient entre la représentation des connaissances et les modèles différentiels.

Ces deux modes de classification ne s'opposent pas mais sont complémentaires. La notion de classe fondée sur la mesure permet de partitionner les choses tombant sous la même catégorie au sens d'Aristote. Elle permet d'organiser des entités de la mémoire épisodique qui correspondent en informatique à des entités regroupées dans des bases de données. Cette partition ne produit pas de nouvelles catégories car elle est arbitraire et elle ne révèle pas de nouveaux attributs sur les différentes classes produites. Les méthodes de raisonnement associées seront algébriques et logiques dans le cas de la catégorisation par différenciation, elles seront statistiques ou heuristiques dans le cas de la classification par ressemblances. Ainsi, à l'intérieur des catégories telles que nous les organisons avec le modèle Anadia, les différences accidentelles entre les objets tombant sous une catégorie, celles qui portent sur les valeurs des attributs mesurables (pas sur leur existence, mais sur les valeurs prises dans les objets particuliers) sont le support de classifications.

3.2. Mémoire et représentation

La question de la représentation du monde a souvent été traitée en intelligence artificielle comme une capacité d'un agent à décrire mentalement les objets, les structures et les processus pour raisonner ensuite sur cette représentation (Davis et al., 1993) qui sert de substitut à la réalité. En représentation des connaissances, les choses représentées sont présupposées connues et la représentation qui en est faite est supposée complète et intrinsèque. Ainsi les conclusions des raisonnements qui s'exercent sur cette représentation peuvent être transférables sur le

monde. Une telle représentation ne dépend en principe ni de l'agent qui la construit, ni de son activité physique et langagière. Elle est stable et observable : c'est une représentation scientifique ou ontologique du monde. Ce sont ces qualités qui les rendent partageables par nature.

En représentation des connaissances le mot **classe**, est utilisé au sens de description générique dans une perspective ontologique de description du monde. Les descriptions génériques sont de deux ordres : elles correspondent à des **concepts** ou à des **prototypes** (Nyckees 1998, ch. 12 et 13.)**.** Les concepts décrivent les propriétés caractéristiques des objets qui en relèvent. Ces propriétés sont habituellement décrites en terme de conditions nécessaires et suffisantes. Les prototypes décrivent les propriétés typiques en terme de ressemblance. C'est une autre façon de regrouper les choses, autour d'un objet central. On peut voir la distinction entre concepts et prototypes comme relevant de la distinction entre mémoire sémantique et mémoire épisodique. La distinction que nous faisons entre mémoire sémantique et mémoire épisodique est d'abord fondée sur le rapport au temps. Les prototypes sont très liés à l'expérience des sujets, à leur boucle sensori-motrice. Ils ne se stabilisent pas par l'enseignement, mais ils résultent d'une généralisation. Ils forment une organisation pré-conceptuelle de la mémoire épisodique. Les concepts structurent la mémoire sémantique et résultent d'une abstraction, ils sont hors du temps. Or la construction d'une telle représentation comme objet social partagé présuppose le dialogue entre les humains et elle ne peut donc pas être un préalable à l'activité langagière. Les réseaux sémantiques et les graphes conceptuels, qui cherchent à être des analogues des images mentales, se placent du côté « positif », du côté de la description des choses, du côté de la substance. Dans le courant linguistique initié par Saussure et Hjelmslev où nous nous situons, il s'agit au contraire de se placer du côté des différences et des relations entre les choses, du côté des formes, qui peuvent être saisies par les individus et nommées ensuite. La distinction entre un terme ontologique et un mot peut être comprise à partir d'un exemple : le mot « chat » peut désigner aussi bien un animal vivant ou mort, un jouet en peluche, une photo, un dessin… Il sert à désigner une forme, quelles que soient ses réalisations physiques. Dans une ontologie, on s'intéresse d'abord aux substances organisées et le terme « chat » se référant à nos différents exemples apparaîtra à des endroits complètement différents du dispositif de catégorisation : le chat vivant dans le genre *animal*, le chat en peluche dans le genre *jouets* et la photo ou le dessin dans le genre *signes analogiques*.

3.3. Mémoire et sémantique des langues

Anadia est une méthode de catégorisation alternative aux méthodes usuelles en représentation de connaissances, bien que n'échappant pas à la filiation aristotélicienne. Elle répond au fait que « un des problèmes majeurs que pose au sémanticien la description des structurations lexicales réside dans le fait qu'elles tiennent à la fois des systèmes diacritiques (non hiérarchiques) et des systèmes taxinomiques (hiérarchiques) » (Kerbrat-Orecchioni, 1988) car elles résultent des usages qui sont multiformes. À l'inverse des interprétations ensemblistes, qui du reste ne pouvaient pas être celles de l'époque, nous prenons la catégorisation comme un principe de discrimination et non pas comme un principe de représentation. Pour exprimer cette différence d'une autre façon, une catégorisation ensembliste part des choses qui existent pour les organiser, alors qu'une classification discriminante définit des types sous lesquels tombent où ne tombent pas des objets du monde connu. C'est un principe de catégorisation qui laisse sa place à l'inconnu et à l'imaginaire, à l'invention permise par le langage. Ainsi, les concepts décrivent des propriétés nécessaires, mais pas suffisantes. En effet, c'est seulement dans le domaine de l'abstrait que cette notion de condition suffisante a un sens, quand on a la liberté, comme les mathématiciens, de définir ses objets et quand ils sont contenus tout entiers dans leur définition. Un modèle sémiotique renvoie à des choses incomplètement connues, les concepts peuvent déterminer des conditions nécessaires pour que les choses tombent sous ce concept, mais pas des conditions suffisantes afin de gérer des mondes ouverts. Le but n'est pas la représentation du monde, mais son évocation dans des pratiques sémiotiques.

La question du rapport entre la mémoire et la sémantique des langues peut être posée ainsi : il y a une mémoire des choses (objets, processus, événements…) correspondant à l'expérience propre des individus, mais avec le langage, elle prend une dimension sociale liée à la mémoire des énoncés. La mémoire des choses et la mémoire des mots sont-elles construites à l'identique ? Autrement dit, les théories de la catégorisation et de la typicalité sont-elles adéquates à la description du lexique des langues ? Les classes lexicales rassemblent plusieurs mots sémantiquement proches dans le sens où si l'un apparaît dans un énoncé, on peut l'interchanger avec un autre de la même classe sans rendre l'énoncé incohérent. Les catégories classiques n'ont pas obligatoirement de pertinence linguistique en tant que classe lexicale. Par exemple, si l'on convient selon des critères ontologiques de placer l'autruche dans la catégorie des oiseaux, on imagine très mal affecter le mot 'autruche' à la même classe lexicale que 'canari' car on a difficilement d'exemple de co-texte linguistique où l'échange entre les deux mots paraît

naturel. Ceci pose la question de la pertinence des critères ontologiques pour l'analyse et la représentation du matériau lexical. Il apparaît que des critères sociaux et culturels sont plus justes. Par exemple, (Rastier, 1987) rapporte le résultat d'une étude sur le contenu du mot *caviar* suite à une enquête au sein d'une population de collégiens. Il apparaît que le trait luxueux était le plus fréquemment cité tandis que d'autres traits ontologiques tels que granuleux ou encore salé n'étaient jamais évoqués.

Les classes lexicales n'ont donc pas de visée dénotationnelle. Nous allons voir dans cette partie qu'elles sont directement construites par l'usage des langues mais qu'elles ne préexistent pas à cet usage. La mémorisation d'un mot tient donc avant tout à sa mise en co-texte et plus précisément aux interprétations des chaînes dans lesquelles il apparaît. Pour étudier cette mise en co-texte nous allons décrire la construction d'effets de sens au moyen d'opérations interprétatives. Ceci nous amènera à nous placer dans le cadre d'une sémantique componentielle. Les contraintes d'une implémentation pour un agent logiciel conduisent à redéfinir la notion de sème, base de la sémantique componentielle.

3.3.1. Les opérations interprétatives

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, l'interprétation d'une chaîne linguistique ne se réduit pas à une structure syntaxique "décorée" par la sémantique. Les significations qui composent la chaîne ne sont pas premières par rapport au sens de la chaîne. Le sens de l'énoncé et les significations des mots de la chaîne sont identifiés en même temps dans une articulation de deux dimensions, l'axe syntagmatique et l'axe paradigmatique (Figure 3) :

- L'axe syntagmatique est l'axe de la chaîne linguistique. C'est l'axe du temps dans l'apparition des constituants de la production langagière (énoncé, phrase, texte, récit …)
- L'axe paradigmatique représente les significations que l'on peut commuter les unes aux autres dans une chaîne linguistique en garantissant qu'elle aura toujours un sens. C'est l'axe des classes lexicales décrivant les différentes terminologies et les domaines thématiques.

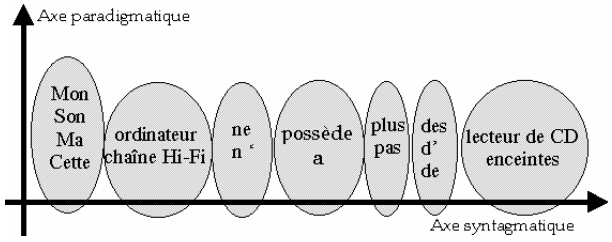

Figure 3 — *Axe paradigmatique, axe syntagmatique*

L'articulation entre l'axe paradigmatique et l'axe syntagmatique est décrite dans la sémantique interprétative de Rastier (Rastier, 1987) en terme d'*opérations interprétatives*. L'étude des opérations interprétatives vise la dynamique des significations des constituants de la chaîne. Il s'agit d'étudier et de modéliser l'effet de l'enchaînement syntagmatique (c'est-à-dire les effets de co-texte) sur la mémorisation et la remémoration des mots. Les mots ne peuvent être appréhendés de façon isolée un à un en dehors de leur co-texte d'apparition. Ce sont de purs artéfacts que les dictionnaires inventorient. C'est aussi ce que montrent les *fonctions lexicales* de Mel'cuk (Mel'cuk , 1986) construisant les significations par des rapports co-textuels entre significations[3].

La dynamique des significations tend en général à renforcer les propriétés communes des composants de la chaîne et par-là, à effacer celles qui ne donnent pas lieu à une répétition dans le co-texte. D'un point de vue théorique, l'interprétation consiste en une mise en évidence d'*isotopies*. Depuis les travaux de Greimas (Greimas, 1966), l'isotopie est définie comme la récurrence d'un même trait sémantique au sein d'une entité textuelle (énoncé, phrase, paragraphe…). C'est le principe de base d'un modèle sémantique de l'axe syntagmatique que nous avons expérimenté pour l'interprétation d'énoncés en langue naturelle par un agent logiciel (Beust et al., 1998).

L'établissement des isotopies est le résultat d'opérations interprétatives sur le sens des mots appelées *actualisation* et *virtualisation*. L'actualisation consiste à identifier un trait sémantique dans un contexte. Par exemple, dans l'énoncé

(1) Le facteur m'a donné une lettre.

Le trait /courrier/ est actualisé dans le contenu du mot 'lettre' parce qu'il se répète dans le contenu de

[3] Il existe dans le dictionnaire explicatif et combinatoire environ un soixantaine de fonctions lexicales. Par exemple la signification de voiture est liée à celle d'automobile par la fonction lexicale de synonymie selon la relation SYN(voiture)=automobile. De même la signification de respect est liée à celle d'irrespect par la fonction lexicale d'antonymie selon la relation ANTI(respect)=irrespect. D'autres fonctions lexicales expriment également des rapports de collocations comme dans BON(conseil)=précieux qui exprime qu'on qualifie en général en français un bon conseil par l'adjectif précieux.

'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*. C'est ainsi que l'actualisation permet une résolution de la polysémie lexicale dans l'interprétation. Ainsi, dans cet exemple, le trait /courrier/ est renforcé par le co-texte alors que ce n'est notamment pas le cas du trait /en papier/ appartenant également au contenu de 'lettre' ; à l'inverse, ce trait 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 trait en contexte. Par exemple, dans le mot composé *pilote automatique*, on dira que le trait /humain/ appartenant à *pilote* est virtualisé car non seulement il n'est pas répété mais, de plus, il est invalidé par le contenu de *automatique*.

L'actualisation permet d'attribuer à des mots des traits que l'on n'aurait pas prévus à priori. C'est ce qu'on appelle l'affférence. Il y a deux sortes d'affférence : les affférences co-textuelles et les affférences socialement normées. Une affférence co-textuelle (i.e. le trait afférent est inhérent dans d'autres mots du co-texte) est une assimilation. Ce serait le cas du trait /légume/ qui serait assimilé dans le contenu de 'scoubidous' du fait de l'énumération dans l'énoncé *Voici des choux, des concombres et des scoubidous*. L'assimilation est aussi l'affférence qui décrit le rattachement des anaphoriques à leurs antécédents. Par exemple, le trait /objet/ n'est pas défini dans le contenu du pronom 'il' mais il y est afférent dans l'énoncé *J'ai retrouvé mon stylo, il était caché.* C'est donc en contextualisant des mots que l'on fixe leurs contenus. L'affférence, en décrivant un effet du syntagmatique sur le paradigmatique, explique l'acquisition des langues par leur pratique. L'opération interprétative inverse de l'assimilation est la dissimilation. Alors que l'assimilation diminue, par affférence, les contrastes forts, la dissimilation, quant à elle, augmente les contrastes faibles. C'est le cas lorsqu'un mot générique est utilisé pour exprimer une idée spécifique. Par exemple, dans *routes et autoroutes*, le contenu de *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 du même mot dans une coordination :

(2) *Il y a musique et musique.*

Ici, l'effet de sens que produit la dissimilation consiste à distinguer les deux significations de *musique*. Ainsi, on peut afférer à la première le trait /agréable/ et à la seconde /désagréable/. Une telle dissimilation due à la répétition d'un même mot conduit à ne pas interpréter l'énoncé comme une tautologie, alors que ce serait le cas dans un langage formel.

Quand elle n'est pas co-textuelle, l'affférence est socialement normée, c'est-à-dire qu'elle est le fait d'une norme sociale partagée au sein d'une communauté linguistique. C'est, par exemple, le cas du trait /tristesse/ afférent au mot 'noir' dans *il broie du noir* ou encore le cas du trait /bonheur/ dans 'rose' dans *la vie en rose*. L'affférence socialement normée, montre en quoi les langues et les mots, qui sont en perpétuel changement, sont des productions de sociétés humaines.

Ce que montrent les opérations interprétatives, c'est que le contenu d'une occurrence d'un mot dépend beaucoup plus de sa contextualisation que d'une représentation à priori. Ce n'est pas le cas pour les objets du monde pour lesquels on peut construire des catégories dont les critères définitoires indiquent les usages de l'objet et délimitent une notion d'objet prototype de la classe. Les opérations interprétatives indiquent qu'une classe lexicale n'est pas une catégorie classique et qu'il n'y a pas de sens prototypique d'un mot.

3.3.2. La sémantique componentielle

Les opérations interprétatives décrites ci-dessus sont l'objet d'étude de la sémantique interprétative (Rastier, 1987) qui se place dans le cadre théorique de la sémantique componentielle. La sémantique componentielle s'est inspirée de travaux de chercheurs comme Katz et Fodor (Katz et Fodor, 1963) qui proposaient de définir à priori un ensemble de traits et d'indiquer dans les dictionnaires, pour chaque sens d'un mot, le sous-ensemble des traits présents (ex : *BLEU 1. qualificatif, couleur, objet physique, inanimé... 2. qualificatif, être humain, novice...*) (Sabah, 1996). Elle a ensuite trouvé son assise grâce à des linguistes comme Greimas (Greimas, 1966) et Pottier (Pottier, 1964). Le but premier de cette sémantique est de décrire scientifiquement la signification des mots.

«De la même façon qu'un phonème est défini comme un faisceau de traits phonétiques simultanés, la signification peut être définie comme un faisceau de traits sémantiques simultanés » (Osgood, 1963)

À l'instar de la terminologie de la phonologie (*phèmes*, *phonèmes*) dont cette sémantique s'est inspirée (cf. citation de Osgood), le trait sémantique est appelé *sème*. Il se définit comme une unité minimale de signification qui, de même que le *phème*, n'est pas susceptible de réalisation indépendante. Les *sémèmes,* équivalents en sémantique des *phonèmes* de la phonologie, sont définis comme des combinaisons de *sèmes* décrivant des significations.

La sémantique componentielle permet d'envisager la construction de la signification selon deux méthodes :

l'une dite référentielle consiste à définir le sème comme une qualité du référent, la seconde dite différentielle définit les sèmes comme des critères de différenciation entre les significations d'une même classe lexicale. Rejetant une vision dénotationnelle des classes lexicales, nous nous plaçons dans le cadre épistémologique d'une description componentielle différentielle. Il s'agit en d'autres termes de "*rechercher les clefs des significations dans les relations qu'elles entretiennent entre elles plutôt que dans les relations que chacune d'entre elles, considérée isolément, entretient avec le monde"* (Nyckees, 1998).

3.3.3. Redéfinition de la notion de sème

La vision classique du sème correspond à une approche linguistique des systèmes de significations, *i.e.* à une approche descriptive du sens qui présuppose une interprétation humaine. Ainsi, le sème est défini comme *« toute paraphrase métalinguistique ne contenant rien d'autre que l'interprétation que l'on peut en faire »*, par exemple /animé/ ou /sert à découper le fromage/.

Dans une approche computationnelle, nous ne devons plus considérer le sème comme une entité autosuffisante issue de l'interprétation mais plutôt comme une entité relationnelle. Ce qui nous paraît ainsi pertinent pour fonder une combinatoire de représentation pour un agent logiciel, ce n'est pas uniquement le trait sémantique en lui-même mais également ce à quoi on peut l'opposer (Beust p.85, 1998). Nous proposons donc de redéfinir le sème comme un jeu d'opposition de signes structurant au niveau même des éléments de base de la signification. Comme nous avons pu le vérifier à travers l'étude d'un corpus de conversation (le corpus PIC[4]), 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, son *domaine d'interprétation*.

Considérons par exemple les mots *livre* et *billet*. Dans une description linguistique du contenu de ces mots et dans le cadre d'une sémantique componentielle classique, on pourrait tous deux leur affecter le sème /en papier/. Cependant, il ne nous semble pas dans un tel exemple qu'il soit question du même sème car les oppositions de valeurs ne sont pas les mêmes. En ce qui concerne le contenu de *livre* le trait papier est pertinent dans les oppositions papier *versus* logiciel ou encore papier *versus* oral. Ces oppositions concernant le domaine d'interprétation des supports de l'information, nous utiliserions donc ici le sème suivant :

Domaine d'interprétation : supports de l'information
Oppositions : papier *vs.* logiciel
papier *vs.* oral

Concernant le mot *billet* l'opposition pertinente concernant les types de monnaie est papier *versus* métallique *versus* électronique. Nous utiliserions donc pour billet le sème :

Domaine d'interprétation : type de monnaie
Opposition : papier *vs.* métallique *vs.* électronique

De manière similaire, une même opposition peut s'inscrire dans des domaines d'interprétations différents, entraînant ainsi des phénomènes d'emplois métaphoriques :

(3) - « Tu ne peux pas tuer tous les gens qui ne te plaisent pas ! »
- « Pourquoi ? »
- « …c'est mal. » [5],

Dans le dernier énoncé de ce dialogue nous sommes en présence du sème :

Domaine d'interprétation : morale
Opposition : mal *vs.* bien

(4) - « Mère, je respire à peine ; j'ai mal à la tête » : [6],

Dans cet énoncé nous sommes en présence du sème :

Domaine d'interprétation : santé et bien-être
Opposition : mal *vs.* bien

Notre propos n'est pas ici de nous attarder sur le phénomène de la métaphore. L'exemple précédent montre l'importance du *domaine d'interprétation* quant aux sèmes évoqués dans un énoncé. Dans cette optique, le sème est donc redéfini comme étant le couple [domaine d'interprétation + un jeu d'oppositions]. Il n'est plus l'extrémité d'une relation fonctionnelle binaire entre significations (Bachimont p.133, 1996) mais **cette relation** même entre significations.

Ainsi, les sèmes et les représentations componentielles que l'on forme avec les sèmes n'ont pas pour but de fixer de la meilleure façon possible une connaissance objective partagée par le plus grand nombre. Au contraire, ces représentations sont situées dans une interaction entre un humain et une machine et elles ne sont justifiées que par le point de vue de l'usager relativement à son projet. Il s'agit de donner à

[4] c.f note 2.

[5] in *Terminator 2 : Judgement Day* – scénario de James Cameron, W.Wisher – 1991.

[6] in *Les Chants de Maladoror* – Chant I, strophe 11 – Lautréamont – 1869

la machine des représentations lexicales liées à une dynamique d'apprentissage et d'interaction, et non des représentations fixant des significations de façon universelle ce qui est souvent l'objectif dans d'autres applications de la sémantique componentielle. Ici, comme dans la langue, tout peut être discuté à tout moment par chacun. Les représentations paradigmatiques n'ont pas d'autre objectif que d'être contextualisées dans l'enchaînement syntagmatique. L'analyse interprétative de cet enchaînement en terme d'isotopies, d'actualisation et de virtualisation conditionne en retour les représentations paradigmatiques et c'est ainsi que la sémantique des langues est une activité mémoire sans fin. Cette articulation de représentations paradigmatiques différentielles et d'opérations interprétatives est un premier analogue de cette activité pour un agent logiciel.

3.3.4. Fondements du modèle

Nous avons vu que l'activité de catégorisation est une activité primitive des êtres vivants qui permet leur adaptation au milieu (Dubois, 1991). Parmi tout ce qui peut être observé du monde qui nous entoure et de notre activité, seulement ce qui fait différence pour les résultats de l'activité du sujet sera construit dans la mémoire. Avec l'activité langagière, la catégorisation prend une dimension sociale car si les choses dont on parle peuvent être connues de l'interlocuteur par ses sens et ses actions, elles peuvent aussi être absentes et même être inconnues de lui : elles seront alors présentées dans le dialogue par différentiation avec les choses connues. Le jugement, qui ne suppose que la comparaison et la possibilité de faire des différences, formera donc la base de la co-référenciation entre agents humains et logiciels. C'est la catégorisation au sens langagier qui peut servir de base à la constitution d'un terrain commun pour la communication entre des agents logiciels. Nous proposons donc d'organiser la mémoire sémantique d'un agent logiciel comme une représentation praxéologique, en vue d'une pratique, comme un modèle du terrain commun nécessaire aux interactions, qui n'est ni une représentation ontologique des choses, ni une représentation fondée sur des mesures. On peut ainsi construire une connaissance partagée du monde physique et biologique, mais on peut aussi construire un monde imaginaire partagé à travers les mythes, les contes, les romans, les projets, et un monde social (Nicolle et Saint-Dizier, 1998).

4. Le modèle Anadia

Ce chapitre présente les principes généraux mis en jeu dans le modèle de la mémoire sémantique comme champ de jugements, comme fondé sur la valeur, notion introduite par Saussure pour rendre compte de la sémantique des langues. Un exemple permet d'expliquer les avantages du mode de représentation choisi. Ensuite, le processus interactif de construction des réseaux de discrimination, figurant l'analogue de la mémoire sémantique humaine, est exposé d'abord en soi, sans préjuger d'une automatisation.

Concevoir la mémoire sémantique comme un champ de valeurs, qui suppose la comparaison mais pas la mesure, reporte tout ce qui concerne les attributs mesurables dans la compétence de la mémoire épisodique. Mais il faut bien entendu concevoir ces deux formes de mémoire de manière articulée, comme deux dimensions orthogonales d'une même compétence des sujets. La mémoire sémantique est l'objet du modèle Anadia, la mémoire épisodique est représentée de manière duale à la mémoire sémantique, par des listes ou des bases de données relatives à chaque catégorie, avec un pointeur sur le genre qui correspond. Dans ce modèle chaque catégorie effective fait une place :

1) à la définition d'une classe-type des objets tombant dans cette catégorie, au sens des classes en conception et programmation par objets,

2) à la définition d'un prototype relatif à la catégorie,

3) à des listes ou des bases de données correspondant à la mémoire épisodique, contenant les objets connus relatifs à cette catégorie,

4) aux classifications correspondantes.

4.1. Les principes

Le principe de discrimination des valeurs en tables a été utilisé manuellement par Jacques Coursil à de nombreuses reprises pour l'analyse du discours (Coursil, 1993), (Coursil, 2000). L'articulation de ce principe avec le modèle du séme de Pierre Beust, présenté dans la section 3.3. est l'objet de cette section. On définit un réseau de discrimination, qu'on appelle dispositif, comme analogue de la mémoire sémantique humaine. Les dispositifs sont constitués de tables à l'intérieur desquels on partionne un genre en catégories. Les tables d'un même dispositif sont reliées entre elles par des relations de sous-catégorisation (Figure 11). Comme dans tous les modèles de catégorisation le processus est récursif. À chaque niveau de catégorisation, on appellera « genre » la catégorie de départ et « catégories » les catégories obtenues par décomposition[7]. Pour

[7] En référence aux définitions courantes pour ces deux termes : 'genre' désignant un ensemble d'éléments présentant des caractères communs et 'catégorie' désignant une classe dans laquelle on range des objets, des personnes présentant des caractères communs.

partitionner un genre, on utilise un ou plusieurs critères discrets ayant un petit nombre de valeurs. Les critères retenus ne sont pas hiérarchisés, ils sont croisés et la combinatoire des valeurs forme autant de sous-catégories potentielles. Parmi ces catégories, certaines peuvent être nommées car des choses leur correspondent et d'autres non. Si toutes les catégories, ou presque toutes, sont nommées, le choix initial était correct. Si trop de catégories sont vides, les critères choisis pour la sous-catégorisations doivent être ré-étudiés.

La constitution des dispositifs de discrimination part d'une situation concrète où on va chercher à différentier **n** objets d'un même genre. Pour différencier deux objets, il suffit d'une caractéristique binaire. Pour en différentier quatre, il faut deux caractéristiques binaires ou une caractéristique à quatre valeurs, une caractéristique binaire et une caractéristique ternaire pour en différentier six. C'est en principe le maximum qu'on envisagera simultanément car quand le nombre d'objets à différentier augmente, il faut procéder par paliers pour que chaque table de catégorisation puisse être maintenue en mémoire de travail. Au-delà de six classes, on commence en principe par construire et nommer des catégories intermédiaires. Chaque objet est alors placé dans sa catégorie, puis dans une sous catégorie en choisissant indépendamment les attributs de chaque sous catégorie. Cette catégorisation progressive est nécessaire pour notre fonctionnement cognitif, elle n'est pas importante pour les machines, qui pourraient gérer des tables et des topiques beaucoup plus grandes, le modèle informatique permettant d'utiliser un nombre indéterminé d'attributs ayant un nombre indéterminé de valeurs. Une raison interne à la méthode amène aussi à procéder par paliers : tous les attributs ne sont pas pertinents pour tous les objets. Faire une grande table avec tous les attributs en colonne et tous les objets en ligne comme dans les bases de données relationnelles est donc une mauvaise solution pour la catégorisation car elle crée des places inutiles en très grand nombre. Cette structure de base de données est pertinente pour la mémoire épisodique, pour représenter les valeurs des attributs communs des objets d'un même genre, pas pour la mémoire sémantique, dont le rôle est de distinguer des genres. Il faut donc créer des tables différentes pour chaque catégorie, avec des attributs pertinents seulement. Certains attributs deviennent pertinents dans une sous-catégorie du fait que d'autres attributs ont une certaine valeur (par exemple, seuls les objets matériels peuvent avoir un poids et un volume), c'est ce qu'Aristote appelait les attributs propres de la catégorie.

Le modèle du sème tel que nous l'avons défini dans la section 3.3 trouve un mode de représentation informatique dans le modèle Anadia. Les attributs des tables sont les domaines d'interprétation. Les valeurs des attributs forment des oppositions cohérentes pour le genre à partionner. Reprenant le mode de présentation des sèmes de la section 3.3, le registre de la table (Figure 4) correspond aux sèmes suivants :

Domaine d'interprétation : support de la documentation
Opposition : papier *versus* logiciel

Domaine d'interprétation : présentation de la documentation
Opposition : globale *versus* précise

Documentation	***Support***	***Présentation***
Manuel	papier	globale
Référentiel	papier	précise
Aide en ligne	logiciel	précise
Tutoriel	logiciel	globale

Figure 4 — *Une catégorisation des documentations*

Ainsi, notre modèle de la mémoire sémantique pourra, comme nous le verrons à travers les exemples d'utilisation, être la base d'opérations interprétatives automatiques car il offre un modèle de représentation implémentable en machine intégrant lui-même un modèle calculatoire de représentation sémantique : le sème tel que nous l'avons redéfini.

Les dispositifs Anadia présentent deux différences avec les catégorisations usuelles. La première différence est qu'un principe de discrimination se substitue aux descriptions en positif proposées par les modes de représentation habituels, soit à partir de conditions nécessaires et suffisantes, soit à partir d'un air de famille. Cette première différence a été expliquée dans la section 3.1., elle consiste à décrire non les choses elles-mêmes, mais leurs différences, à produire un dispositif de différentiation entre les choses, qui se placent dans les trous du dispositif. La deuxième différence est que les méthodes en catégorisation arborescente supposent qu'il existe un arbre unique où toutes les choses sont représentées, des plus générales aux plus spécifiques, alors que nous proposons des tables décentralisées, qui dépendent de points de vue variés, et dont le recoupement n'est ni obligatoire, ni automatique. Comme la mémoire sémantique se constitue dans l'expérience de l'agent, elle n'a pas de raison d'être un dispositif de discrimination globale, centralisée par les différences les plus primitives. Bien sûr, il peut arriver que les agents aient pour but d'établir une ontologie et donc de se poser ces questions, mais ce n'est pas un fonctionnement premier. Les dispositifs sont propres aux situations et ne fusionnent que s'il y a remémoration d'une situation à propos d'une autre situation, ce qui peut alors provoquer des remises en ordre.

4.2. Un exemple

Pour montrer l'avantage d'une représentation en table (ou en dispositif), sur la représentation en arbre, reprenons un exemple donné par (Eco, 1988 p. 99-100) où le genre de base à partitionner est « animal ». Les arbres de Porphyre, qui ont été cités dans la section 3.1., partitionnent un genre en choisissant un attribut, puis partitionnent chaque sous-catégorie à partir d'un autre attribut. Si deux attributs sont au même niveau, une fois qu'on a choisi de placer un attribut avant un autre, il faudra redécomposer les catégories obtenues avec l'autre attribut, et on peut créer des arbres alternatifs sans pouvoir les départager. (Figure 5.)

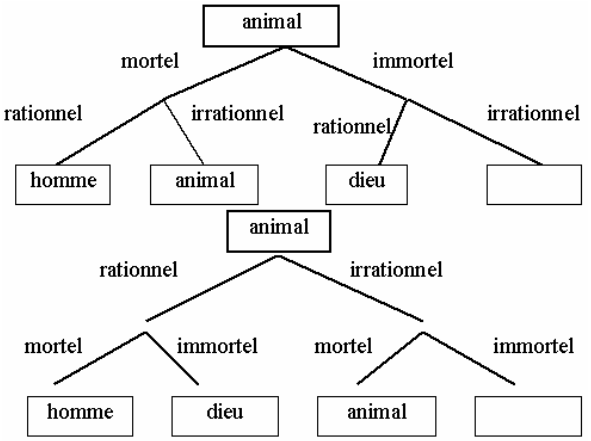

Figure 5 — *Les sous-catégorisations alternatives du genre animal avec l'arbre de Porphyre*

Remarque : un mot comme animal peut être utilisé à plusieurs niveaux de la catégorisation, puisqu'il prend son sens relativement aux termes auxquels il s'oppose. Ceci correspond dans la langue aux relations d'hyponymie.

Les tables prennent en compte simultanément tous les attributs pertinents pour un genre donné sans être pertinents pour le genre supérieur (Figure 6). On évite ainsi les problèmes de choix de l'attribut à chaque étape de la discrimination, car ce choix est souvent arbitraire. Le cas où une catégorie est partitionnée par les valeurs d'un seul attribut est un cas particulier dans la méthode Anadia qui nous ramène ponctuellement aux arbres de discrimination. Dans le cas général, on prend en compte en même temps tous les attributs partitionnant une catégorie, il n'y a donc pas d'arbres alternatifs. Et même plus, si le processus de sous-catégorisation amène à utiliser le même attribut pour re-partitionner deux sous-catégories de même niveau, il est prescrit de reconsidérer la sous-catégorisation pour remonter cet attribut au niveau de la catégorie, ce qui peut amener à retarder la prise en compte d'un autre. La partition de toutes les sous-catégories de même niveau n'utilise jamais le même attribut.

mortel	rationnel	
oui	oui	homme
oui	non	animal
non	oui	dieu
non	non	

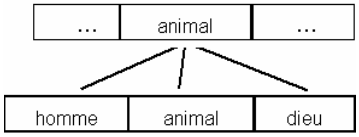

Figure 6 — *La sous-catégorisation du genre animal avec Anadia*

Même si l'exemple d'Eco est de caractère ontologique, les atouts d'une représentation en table (ou en dispositif) se manifestent de la même façon lorsque les descriptions mises en jeu relèvent du langage, sans souci d'universalité.

4.3. Définitions

Chaque table Anadia partitionne un genre en catégories en fonction de la combinatoire des différences observées entre les choses qui relèvent de ce genre. Les catégories obtenues peuvent à nouveau être partitionnées et ainsi de suite. Définissons les termes utilisés dans ce modèle :

attribut : propriété qu'on peut ou non attribuer à une chose. Par exemple, les choses matérielles ont une couleur et les événements n'en ont pas, les processus ont une durée, et pas les figures géométriques.

domaine : les valeurs possibles d'un attribut.

registre : liste des attributs discrets à domaine fini les plus pertinents pour un genre à partitionner dans un certain contexte[8]. Comme Aristote, nous distinguons deux grands types d'attributs, les attributs à valeurs continues et les attributs à valeurs discrètes. Seuls les attributs à valeurs discrètes et ayant un petit nombre de valeurs, peuvent être utilisés pour différentier des catégories et donc intervenir dans les registres.

table : combinatoire des places obtenues par croisement des valeurs des attributs d'un registre. Une table est un moyen de distinguer des catégories dans un genre. Chaque place correspond à un concept, une catégorie potentielle et on espère que les objets à distinguer vont tomber dans des catégories différentes.

dispositif : jeu de tables reliées entre-elles par des relations de sous-catégorisation. Un dispositif est le

[8] Rappelons qu'il ne s'agit pas de faire des catégories universelles mais des catégories utiles pour une tâche interactive, ce qui amène à tenir compte du contexte dans le choix du registre.

reflet de la catégorisation d'un domaine par un sujet pour un sujet.

sélection : ensemble des places de la table reconnues comme valides parce qu'on peut les nommer ou en donner des exemples. Chaque place sélectionnée correspond donc à une catégorie effective. Les autres places sont vacantes, soit pour placer des objets inconnus, soit comme support de la création imaginaire. Une catégorie devient un genre quand on cherche à la sous-catégoriser. Il n'y a pas de différence de nature entre genre et catégorie, il y a une différence de rôle.

topique : graphe des relations entre les places sélectionnées (Figures 7 et 8). La relation entre deux places est un nombre de différences de valeurs d'attributs. La topique la plus intéressante est la topique des différences *à un trait près* , c'est-à-dire sur une valeur d'un seul attribut. Les topiques révèlent la structure du domaine initial relativement aux attributs considérés. Si toutes les places ont été sélectionnées, la topique des différences *à un trait près* est un graphe connexe ayant des propriétés spécifiques (tous les sommets ont même degré). On sait alors que tous les attributs choisis sont indépendants et pertinents. La structure des topiques complètes ne dépend pas des attributs ni de leurs valeurs. Les figures ci-dessous donnent les structures des topiques complètes les plus simples. Les valeurs des attributs viendront remplir les rectangles arrondis (les places). Si la topique n'est pas complète, il manque des places et les liens correspondants.

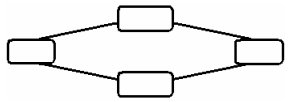

Figure 7 — *La topique pour 2 attributs à 2 valeurs chacun.*

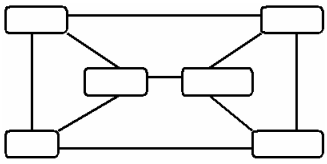

Figure 8 — *La topique pour un attribut à deux valeurs et un attribut à 3 valeurs.*

Dans le cas général, la topique n'est pas complète puisque les places vides n'y figurent pas. Mais si trop de places sont vacantes, on va tenter de modifier le registre. Si la topique n'est pas connexe c'est que trop d'attributs ont été envisagés au même niveau, alors qu'ils n'étaient pas indépendants. L'observation de la topique permet de valider ou de remettre en cause le registre dont elle provient. Voici une table différenciant les logiciels en ligne pour la recherche de documents sur Internet et la topique résultante (Figure 9) :

Logiciels en ligne pour la recherche de documents	Accès à l'information	Architecture
moteurs de recherche	requête	index
	browsing	index
annuaire de recherche	browsing	hiérarchie
	requête	hiérarchie

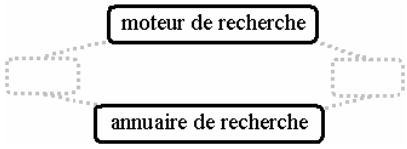

Figure 9 — *Table Anadia et topique résultante des logiciels en ligne pour la recherche de documents sur Internet.*

Dans cet exemple trivial, la topique résultante de la table n'est pas connexe (les parties grisées ne font pas partie de la topique). Le registre correspondant n'est pas judicieux car les deux attributs présents ne sont pas indépendants.

4.4. Propriétés

L'opération de sélection dans la table et l'observation des topiques révèlent les propriétés des attributs considérés pour construire la table :

1)- Une propriété relative à la relation entre les catégories et les attributs : la pertinence.

Un attribut est pertinent pour une catégorie si tous les exemplaires de la catégorie ont cette propriété, et s'ils n'ont pas tous la même valeur pour cette propriété. Si tous les exemplaires d'une catégorie ont la même valeur pour un attribut, cette valeur est une caractéristique de la catégorie et cet attribut est pertinent à un niveau supérieur de la décomposition. Si certains exemplaires n'ont pas cette propriété, elle ne doit pas être utilisée pour cette catégorie, mais elle sera peut être utilisée ensuite pour une sous catégorie.

2)- Deux propriétés de relation entre les attributs : l'indépendance et la subordination.

Deux attributs sont indépendants si toutes les places de la combinatoire de la table formée à partir de leurs registres sont sélectionnées, ou si toutes les places moins une le sont. S'il y a plus de places vides, le doute est possible et il faut examiner si un des attributs n'a pas une valeur fixe quand un autre attribut a une certaine valeur.

Un attribut B est subordonné à un attribut A si B est pertinent seulement quand A prend certaines valeurs.

Un attribut subordonné à un autre apparaîtra à un niveau inférieur de la catégorisation, car mettre un attribut subordonné à un autre dans la même table produit des places vides.

3)- Deux propriétés d'un domaine d'un attribut : la complétude et la minimalité.

Le domaine d'un attribut est complet si toutes les choses évoquées peuvent être classées selon les valeurs définies. Cette propriété est triviale si le domaine d'un attribut est oui/non, mais on risque alors d'obtenir de nombreuses places vides dans la table, car les attributs ne seront pas indépendants. Par exemple, on peut réinterpréter l'analyse que les structuralistes ont faite du symbolique (Deleuze, 1972) en disant que le registre de la réalité avec deux valeurs réel/imaginaire est incomplet et doit être transformé en registre réel/imaginaire/symbolique pour pouvoir y placer les objets des systèmes de signes. Avoir 2 attributs : réel oui/non, imaginaire oui/non, fait bien une place pour le symbolique (réel oui, imaginaire oui) mais crée une place vide (réel non, imaginaire non).

Comme les éléments d'une catégorie ne sont pas définis par des propriétés nécessaires et suffisantes, on peut montrer qu'un registre est incomplet en exhibant des exemplaires qui ont des valeurs imprévues, mais on ne peut jamais démontrer qu'il est complet, sauf de manière triviale, et c'est encore une raison pour ne pas interpréter les résultats de la catégorisation proposée en termes ensemblistes.

Le domaine de l'attribut A est minimal pour les attributs subordonnés si chaque attribut subordonné est pertinent pour une seule valeur de A. On peut aussi dire, comme Aristote, que la catégorie correspondante possède cet attribut en propre. Lorsqu'on s'aperçoit qu'un attribut n'est pas minimal, ceci amène une refonte de la grille pour le faire remonter à sa place.

4)- Une propriété d'un registre : la minimalité.

Pour des raisons d'économie d'espace et surtout pour la rapidité des processus de recherche, qui sont la plupart du temps combinatoires, une table ne doit pas introduire de critères superflus. Un registre est minimal si on ne peut enlever ni un attribut, ni une valeur d'un attribut, sans que deux sous-catégories tombent sous le même concept. Il peut arriver qu'il y ait des places vides dans la grille et que le registre soit minimal. Dans la figure 9, le registre choisi n'est pas minimal. L'un des deux attributs est superflu au vu de la sélection de table.

4.5. La méthode

La décomposition s'appuie alternativement sur l'examen des différences entre les catégories et sur l'examen de leurs attributs propres. Elle découpe récursivement une catégorie en sous-catégories en croisant les valeurs des attributs discrets les plus pertinents à ce niveau pour une activité donnée. Il existe un grand nombre d'attributs qui pourraient être utilisés mais beaucoup ne sont pas indépendants, leur prise en compte simultanée introduirait donc des places vides dans la table. Pour minimiser la mémorisation et l'activité psychique, il faut construire la combinatoire des différences les plus saillantes et placer les objets relativement à cette combinatoire de telle sorte que chaque objet puisse être distingué des autres avec le minimum d'effort.

Décider de construire une table, c'est fixer un genre à partitionner. Pour construire une table, il faut d'abord choisir un registre d'attributs et déterminer pour chaque attribut choisi le domaine de ses valeurs. Un registre est composé d'attributs discrets, pertinents et indépendants. Ces propriétés sont recherchées mais si elles n'ont pas été prouvées au départ, elles vont être éprouvées par le processus tout entier. La table est construite par une opération algorithmique qui énumère la combinatoire des croisements entre les valeurs de tous les attributs.

Lorsque la table est construite, trois questions se posent, qui doivent être examinées par un ou plusieurs agents car leur réponse provient de l'observation du monde :

1. Peut-on trouver des choses qui tombent sous chacun des concepts ? Si oui, la catégorie est effective. Elle peut être nommée et elle pourra être prise comme genre à partitionner plus tard.
2. Y a-t-il des choses qu'on ne sait pas classer dans une de ces catégories ? Si oui, il manque des valeurs dans le domaine d'un attribut.
3. Toutes les distinctions voulues sont-elles obtenues ? Si non, il manque un attribut.

Lorsque cet examen révèle des insuffisances du registre, il faut construire un nouveau registre et recommencer.

Lorsque le résultat est satisfaisant, la topique des relations à un trait près peut être construite. D'autres topiques, sur les relations à deux ou trois traits près peuvent aussi être construites par le même algorithme, mais leur utilisation est plus rare (Section 5.1.). Si la topique des différences à un trait près n'est pas connexe, il y a certainement trop d'attributs qui ont été pris en compte en même temps et donc il y a trop de places vides. En regardant ainsi les sélections et les topiques, les agents peuvent à nouveau être amenés à supprimer ou à transformer des attributs, ou à transformer le domaine d'un attribut, « pour parler du monde en même temps qu'ils le constituent » (Kerbrat-Orecchioni, 1988). Le processus recommence jusqu'à

ce qu'une solution satisfaisante soit trouvée, c'est-à-dire qu'un accord entre les agents sur la table soit obtenu.

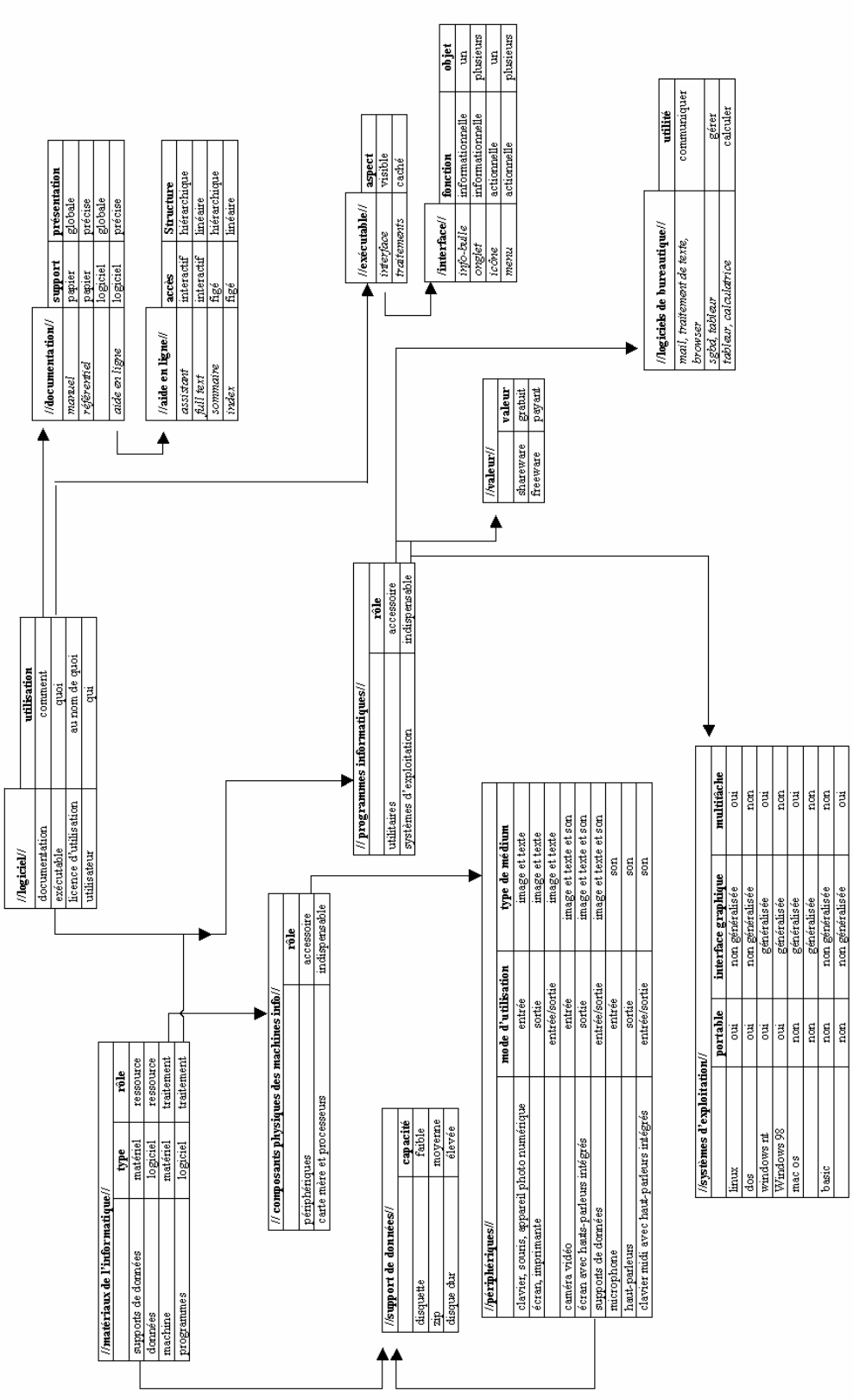

Figure 11 — *Un exemple de dispositif Anadia en rapport avec l'informatique*

Le processus est récursif : les places sélectionnées définissent chacune une catégorie dans laquelle de nouveaux attributs peuvent être pertinents, ce qui amène à raffiner la catégorisation en construisant une nouvelle table. Le processus est récursif d'une autre manière, une catégorie partitionnée peut servir d'attribut pour partager une autre catégorie, les sous-catégories servant de valeurs discrètes pour cet attribut.

Ce processus de catégorisation est opportuniste, récursif et redondant : la même chose peut être classée plusieurs fois sous des points de vue différents. La méthode produit un dispositif de tables de catégorisation où chaque niveau regroupe les sous-catégorisations des places des tables de niveau supérieur. Ce dispositif n'est pas toujours un arbre comme on peut le voir sur l'exemple de la Figure 11. Il dispositif présente de plus chaque sous-catégorisation sous forme d'un graphe des différences à un trait près, la topique. L'examen des propriétés de connexité de la topique permet de valider les choix des attributs faits pour cette catégorie ou plus exactement met en évidence les mauvais choix, ce qui permet de les corriger.

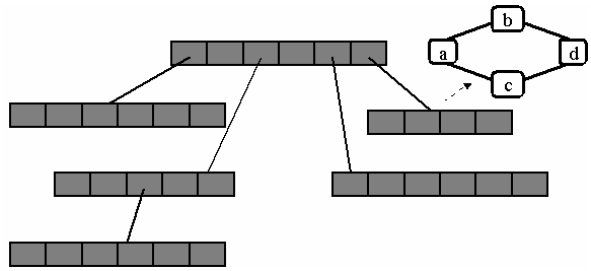

Figure 10 — *Un arbre de catégorisation*

Dans la figure 10, qui est un schéma abstrait, nous voyons en haut une première catégorisation d'un genre en six catégories, quatre de ses catégories ont à nouveau été découpé en catégories. Un exemple concret est donné figure 11. Il n'y a pas de relation entre les sous-catégorisations des places d'un même niveau du dispositif, elles utilisent en général des attributs différents. Pour la sous catégorie la plus à droite, nous avons fait figurer sa topique. La catégorisation peut se développer par partition d'un sous-genre si c'est nécessaire. Dans la figure 10, nous avons fait figurer des sous catégorisations pour les catégories 1, 3, 5 et 6 et en bas une sous-catégorisation supplémentaire pour la troisième sous-catégorie.

Dans les cas concrets, une tâche relative à une catégorie de chose amène à construire un réseau de tables : un dispositif. La figure 11 en présente un exemple dans le domaine des technologies de l'information. Cette représentation en dispositif de tables n'a pas à être complète pour être optimale : elle doit être suffisante pour différencier toutes les entités qui sont visées comme différentes avec le moins de charge cognitive possible. On peut extraire de ces tables une certaine représentation des choses qui tombent dans les catégories en énumérant les attributs des tables et leur valeur. Par exemple :

index : *mode figé, structure linéaire, support logiciel, présentation précise,…*

Une représentation des exemplaires de catégories peut donc être produite comme image d'un état du dispositif, au sens où Peirce dit que le signe nous fait savoir quelque chose de l'objet (Peirce, 1978). Elle permet de rendre compte de manière computationnelle de l'instauration d'un terrain commun entre les participants d'un dialogue. Elle peut servir au raisonnement logique ou à l'argumentation. La description obtenue pour un exemplaire d'une catégorie en déroulant les catégories auxquelles il appartient et en précisant les valeurs de ses attributs propres en donne une description partielle et provisoire qui se restructure par l'interaction. Elle correspond à ce que le signe nous donne à voir de l'objet en tant qu'il est placé dans un système, et donc au sens où il contient tous les autres par défaut. Cette représentation n'est pas permanente, elle n'est pas conservée en tant que telle parce que l'évolution des tables la rendrait obsolète. Cette représentation sous jacente est toujours provisoire, elle n'a donc pas à être mise à jour.

L'introduction d'une nouvelle entité dans un réseau de tables peut se faire sans remettre en cause les catégories existantes si sa place est vacante. Elle peut amener à considérer une nouvelle valeur pour un attribut, ou à considérer un nouvel attribut pour tous les objets d'une catégorie, et donc à refondre l'organisation de la mémoire. Cette réorganisation peut être locale si l'attribut ajouté ou modifié est dans un registre local, mais elle peut avoir lieu même au plus haut niveau. La grille peut aussi s'améliorer par observation des relations entre les attributs, révélées par les opérations de manipulation de ces attributs pour construire les tables.

Construire une bonne grille de catégorisation pour un domaine est un processus expérimental, où il s'agit d'essayer des combinaisons d'attributs, de les déplacer jusqu'à ce que le résultat ait de bonnes propriétés algébriques (il y a peu de places vides et tous les attributs sont minimaux) et jusqu'à ce qu'il ait de bonnes propriétés conceptuelles (on trouve des attributs propres dans chaque sous-catégorie). La grille est toujours complète au sens où tout ce qui est connu entre dedans, et toujours provisoire car elle peut constamment prendre en compte de nouvelles

différences lorsqu'il devient nécessaire de distinguer des choses nouvelles qui sans cela tomberaient dans la même case. Le résultat n'est jamais définitif, car l'observation de nouveaux objets peut amener à prendre en compte de nouvelles différences qui vont amener à restructurer l'ensemble. Ceci tout à fait normal : le modèle a été conçu comme intégrant l'apprentissage dans son principe de fonctionnement (Nicolle et Vivier, 1997), alors que la plupart des modèles d'apprentissage distinguent une phase d'apprentissage et une phase d'exploitation. La grille de catégorisation produite par la sélection dans une table peut toujours s'améliorer par apprentissage de différences pertinentes, relatives à l'expérience de l'agent, aux tâches qui utilisent la catégorisation, ou consécutives à ses échanges langagiers qui amènent à confronter une partie des catégorisations des interlocuteurs

La méthode a d'abord été présentée en soi, sans préjuger de son automatisation (dont les premières étapes seront décrites dans la section 5). La construction d'une table fait alterner des étapes de choix et de définition (choix des attributs du registre, définition de leur domaine, sélection des places), qui nécessitent la mise en relation avec une situation, et donc l'intervention d'un agent, et des opérations algorithmiques (construction des tables et des topiques). Dans cette méthode, certaines étapes sont faciles à automatiser de manière générique. Certaines étapes sont impossibles à automatiser sans un modèle d'agent logiciel interactif complet. Construire de tels agents est notre objectif depuis plusieurs années, et ce travail en est une étape. Mais comme l'objectif global n'est pas atteint, il est trop tôt pour savoir si oui ou non la méthode dans son ensemble peut être mise en œuvre par des agents logiciels. C'est aujourd'hui une hypothèse de travail. La première étape de l'automatisation consiste à faire gérer les arbres de grilles de catégorisation par la machine et à implanter toutes les étapes algorithmiques. Nous allons maintenant décrire un logiciel interactif qui aide un utilisateur dans ce processus en réalisant pour lui les opérations algorithmiques et en vérifiant certaines propriétés des objets construits.

5. Réalisation logicielle

Nous terminons cet article par la description d'une réalisation logicielle permettant la construction semi-automatique des dispositifs Anadia et d'un exemple d'utilisation de tels matériaux. Le problème crucial l'aide à leur construction par un utilisateur ou un groupe d'utilisateurs sera abordé dans une dernière section.

5.1. Le logiciel Anadia

Le logiciel permettant la création dispositifs de tables de catégorisation est une première réalisation expérimentale mettant en œuvre les principes de construction des tables et de leurs relations dans des dispositifs et permettant leur utilisation dans divers champs d'application.

Dans sa dernière version en Java[9], le logiciel offre des outils de production, d'analyse, de modification et de visualisation des dispositifs. Il assure toutes les étapes algorithmiques et offre un environnement d'interaction pour les étapes de choix qui sont de la responsabilité de l'usager. La construction des tables à l'aide du logiciel facilite leurs modifications éventuelles tout en vérifiant la cohérence des relations de catégorisation au sein des dispositifs. Les interfaces proposées permettent également de présenter des points de vue dont l'utilisateur n'avait pas conscience, en montrant les tables construites automatiquement mais aussi les topiques résultantes.

La première des fonctionnalités proposées est la création d'attributs en donnant pour chacun le domaine de ses valeurs et son nom (Figure 12).

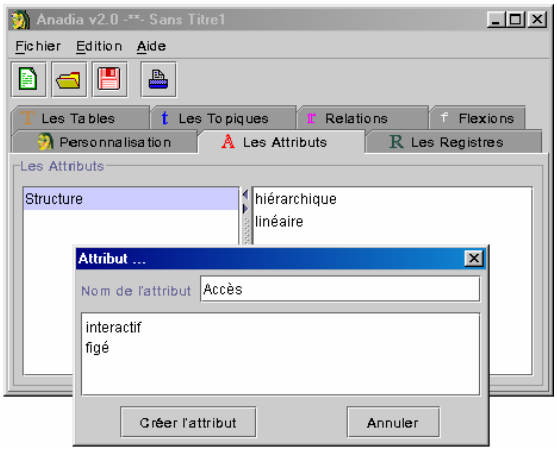

Figure 12 -- *Logiciel Anadia - Création d'un attribut.*

Pour créer un registre, l'utilisateur choisit les attributs qui lui semblent pertinents à ce niveau de son projet (Figure 13). Une fois un registre créé, le logiciel calcule la table qui résulte de la combinatoire des domaines des attributs sélectionnés (Figure 14). La table affichée par le logiciel devient alors le principal lieu d'interaction entre l'utilisateur et le système. L'utilisateur examine chaque ligne afin de trouver celles qui décrivent des jeux de valeurs représentant des catégories effectives (*i.e.* dont il peut connaît un exemplaire). Quand c'est le cas, il a la possibilité de donner un nom à cette ligne et de construire ainsi sa sélection à partir de la combinatoire.

[9] Le logiciel ainsi que les sources, sont disponibles à l'adresse suivant : www.info.unicaen.fr/~perlerin

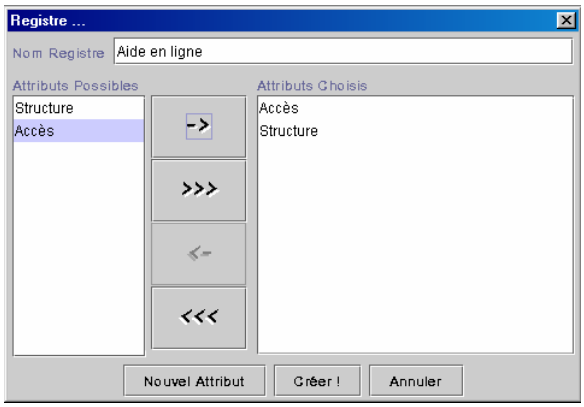

Figure 13 -- *Logiciel Anadia - Création d'un registre.*

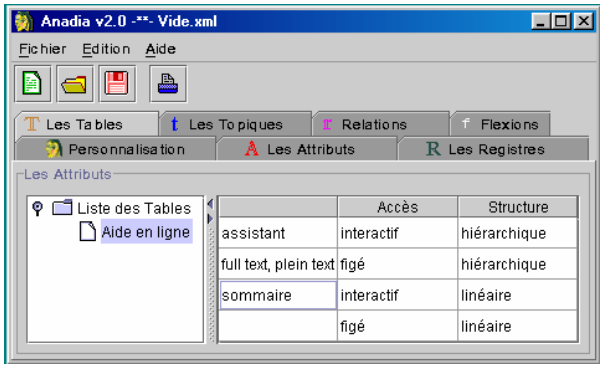

Figure 14 -- *Logiciel Anadia – Présentation de la table calculée par le logiciel.*

Remarque : Dans le cas de la figure 14, l'utilisateur n'a pas encore trouvé de catégorie correspondant à la dernière ligne de la table. Nous pouvons noter que la possibilité est donnée de nommer plusieurs catégories pour une seule ligne (cas de relation de synonymie) .

À partir de la fenêtre où est présentée la table, l'utilisateur peut :

- créer un attribut dont le domaine a pour valeurs les noms des sous-catégories de la sélection. (Dans notre exemple, l'utilisateur peut demander que soit formé un attribut ayant pour domaine *Assistant et Sommaire* dans le but de faire de nouvelles différences dans d'autres catégorisations.)
- demander à re-catégoriser une sous-catégorie de la sélection.

demander le calcul de la topique correspondant à l'état actuel de la sélection (Figure 15.).

Les tables crées peuvent aussi être mises en relation de sous-catégorisation par l'intermédiaire du panel « *Relation* » qui propose une vue d'ensemble du dispositif ainsi formé (Figure 16). Dans ce même panel, l'utilisateur a la possibilité d'attribuer une couleur à chaque table construite pour une visualisation plus aisée des résultats de l'utilisation du dispositif le cadre d'une recherche documentaire (comme dans la section 5.2.).

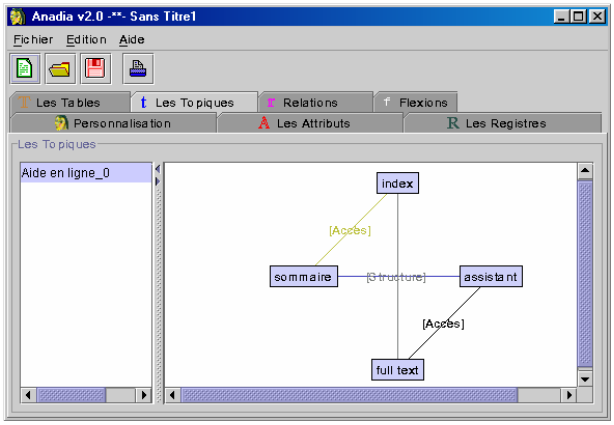

Figure 15 -- *Logiciel Anadia – Présentation de la topique calculée automatiquement. Les arcs de la topique sont étiquetés par les différences mises en jeu.*

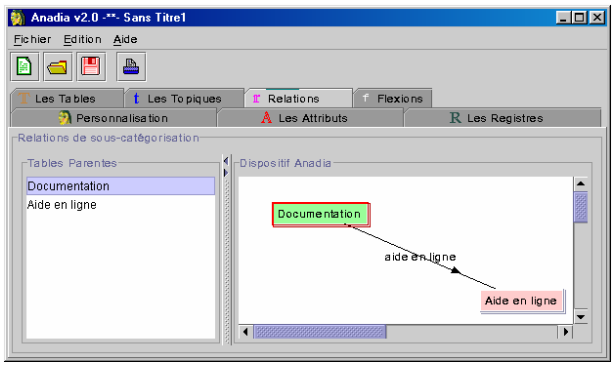

Figure 16 -- *Logiciel Anadia – Présentation du dispositif formé par l'ensemble des tables de la session de travail. Les relations de sous-catégorisation sont étiquetées par la première catégorie de la ligne correspondante appartenant à la table parente.*

Tous les mots en présence dans les tables peuvent faire l'objet d'un calcul automatique ou semi-automatique de flexions (panel « *Flexions* ») grâce à la base de données BDLex du laboratoire IRIT de l'Université Paul Sabatier de Toulouse[10].

Ces fonctionnalités permettent à l'utilisateur de réaliser des sessions de travail regroupant une ou plusieurs tables reliées en dispositif et de les sauvegarder dans des fichiers au format XML[11] facilement utilisables par d'autres applications. Ces documents XML sont tous soumis à la même DTD[12] ce qui permet une interopérabilité pratique. À chaque étape de création,

---

[10] www.irit.fr/ACTIVITES/EQ_IHMPT/ress_ling/

[11] www.w3.org/XML

[12] La DTD (Document Type Definition) définit la structure interne du document XML. Il s'agit en fait de la description du langage, des données utilisées par un programme pour interpréter et présenter un document codé en accord avec l'usage d'une communauté.
www.w3.org/ People/danield/xmlglfr/all.htm

l'utilisateur peut voir l'état de la feuille XML de sa session dans un des menus prévus à cet effet[13] (Menu *« Édition »,* sous-menu *« Voir le source »*).

Anadia offre plusieurs choix de configuration. Outre quelques modalités de confort, on peut choisir le mode d'affichage des topiques (en cercle ou en ligne, avec ou sans l'affichage des différences) et paramétrer la relation servant à les construire. La plus utilisée est la relation de différence à *un* trait prés (sélectionnée par défaut). Elle est la relation exprimant la plus forte proximité entre deux sous-catégories.

5.2. Utilisation des tables Anadia pour la recherche documentaire

Pierre Beust (Beust, 1998) a mis en place avec Anadia un traitement sémantique de surface des énoncés pour l'analyse d'énoncés de dialogue (les contraintes de modalité, de prédication et d'argumentation ne sont pas l'objet de son analyse interprétative). Dans un contexte de recherche documentaire, ce traitement de surface est-il suffisant pour amener une réponse au problème consistant à savoir si un document donné traite de tel ou tel sujet, aborde tel ou tel thème ? Des travaux expérimentaux sont actuellement en cours pour le savoir et cette section présente nos propositions dans ce domaine.

5.2.1. L'état du problème

L'utilisateur souhaitant rechercher un document sur l'Internet a principalement deux possibilités : il peut soit interroger des moteurs de recherche, soit utiliser des annuaires de recherche organisés en répertoires – à l'heure actuelle les principaux sites de recherche proposent simultanément les deux options. La première solution consiste à interroger une base de données de documents indexés par l'intermédiaire d'une requête, la seconde à naviguer dans une arborescence de sujets jusqu'à aboutir à la liste des URL connues par l'annuaire et censées correspondre au chemin discriminant parcouru.

Un nombre important de requêtes soumises aux moteurs de recherche de l'Internet ne satisfont pas pleinement les attentes des utilisateurs (Bellot et Elbèze, 2000). La liste de documents proposée en retour est souvent trop longue : son exploration représente un travail exagérément laborieux pour l'auteur de la requête. Pourquoi ?

Les moteurs de recherche indexent les documents explorés grâce aux mots qu'ils contiennent. La mise en corrélation entre un document indexé et une requête fonctionne selon le principe du *pattern matching* (ou reconnaissance de motif). Un document est considéré comme conforme à une requête si les mots contenus dans cette dernière y apparaissent. Cette technique se heurte souvent au problème de la polysémie ou au cas des documents abordant plusieurs sujets (Perlerin pp.6-12, 2000). Par exemple, la Figure 17 présente les cinq premiers résultats d'une requête soumise au moteur de recherche Google™ [14]. La requête est : 'microsoft virus'. Cette requête est interprétée par défaut comme correspondant à l'expression logique 'microsoft Et virus' : les URL présentées dans la liste des résultats doivent contenir à la fois le mot 'microsoft' et le mot 'virus'[15].

Voici quelques informations sur les 5 premiers documents retournés :

Document 1 :
www.acbm.com/num_01/pages/page27.html
Nombre de mots : 1045
Nombre d'occurrences du mot 'virus' : 0
Nombre d'occurrences du mot 'microsoft' : 9
Description du document : Intitulé «Microsoft se moque-t-il des utilisateurs ?», le document rapporte que Windows 4.0 NT Server et NT WorkStation possèderaient le même noyau. Le mot virus n'apparaît pas dans le document mais une fois dans les balises <Title> du code HTML de la page.

Le document 1 présente une particularité remarquable : alors que la requête soumise le prévoyait, le mot 'virus' n'est pas présent au sein du document à proprement dit. Il n'apparaît qu'une fois au sein des balises <Title> du code HTML et n'est donc visible à l'écran que dans la partie supérieure des butineurs. Comme nous allons le voir par la suite, cette présence ne reflète en rien les thèmes abordés dans le document.

Document 2 :
www.microsoft.com/france/technet/default.asp
Nombre de mots : 855
Nombre d'occurrences du mot 'virus' : 3
Nombre d'occurrences du mot 'microsoft' : 22
Description du document : Le document rassemble une série d'articles décrits en quelques lignes auxquels les utilisateurs peuvent accéder par l'intermédiaire de hyperliens. Les articles ont tous comme sujet les produits Microsoft. Sur la droite, un petit encart intitulé «Truc de la semaine» recèle les trois occurrences de 'virus' et traite des problèmes liés aux fichiers attachés infestés dans les logiciels de messagerie.

[13] Des dispositifs au format XML sont téléchargeables sur http://users.info.unicaen.fr/~perlerin/recherche/anadia/dispositifs.

[14] www.google.com

[15] www.abondance.com/outils/goo_syntaxe.html pour plus d'informations sur le fonctionnement du moteur de recherche Google™.

Document 3 :
www.virus-fr.com
Nombre de mots : 711
Nombre d'occurrences du mot 'virus' : 12
Nombre d'occurrences du mot 'microsoft' : 2
Description du document : 5 virus sont décrits en quelques lignes à la suite d'un court éditorial. Le mot 'microsoft' apparaît une fois dans un paragraphe de 72 mots au sujet d'un macro virus infectant le logiciel *Word 2000*. Ce mot est également présent dans un hyperliens intitulé « *Microsoft victime d'une attaque* » dans la partie droite de l'écran.

Document 4 :
www.ornitho.org/numero24/articles/micro_hack.html
Nombre de mots : 4074
Nombre d'occurrences du mot 'virus' : 18
Nombre d'occurrences du mot 'microsoft' : 20
Description du document : Le document est un article intitulé « Virus I Love You, Microsoft : je t'aime moi non plus ». Il s'agit d'un entretien avec un spécialiste en sécurité Internet au sujet son métier, de la vulnérabilité de Microsoft aux virus et de problèmes généraux liés au piratage informatique.

Document 5 :
www.aful.org/presse/pr-virus.html
Nombre de mots : 1180
Nombre d'occurrences du mot 'virus' : 16
Nombre d'occurrences du mot 'microsoft' : 9
Description du document : L'article explique pourquoi les utilisateurs de Linux ou de systèmes d'exploitations compatibles POSIX ont été épargnés par le virus « I love you ». L'AFUL (Association Francophone des Utilisateurs de Linux et des Logiciels Libres) demande au gouvernement d'inscrire à l'ordre du jour la proposition de loi Le Déaut-Paul-Cohen pour la promotion des logiciels libres insensibles aux attaques des virus et explique pourquoi les produits Microsoft y sont vulnérables.

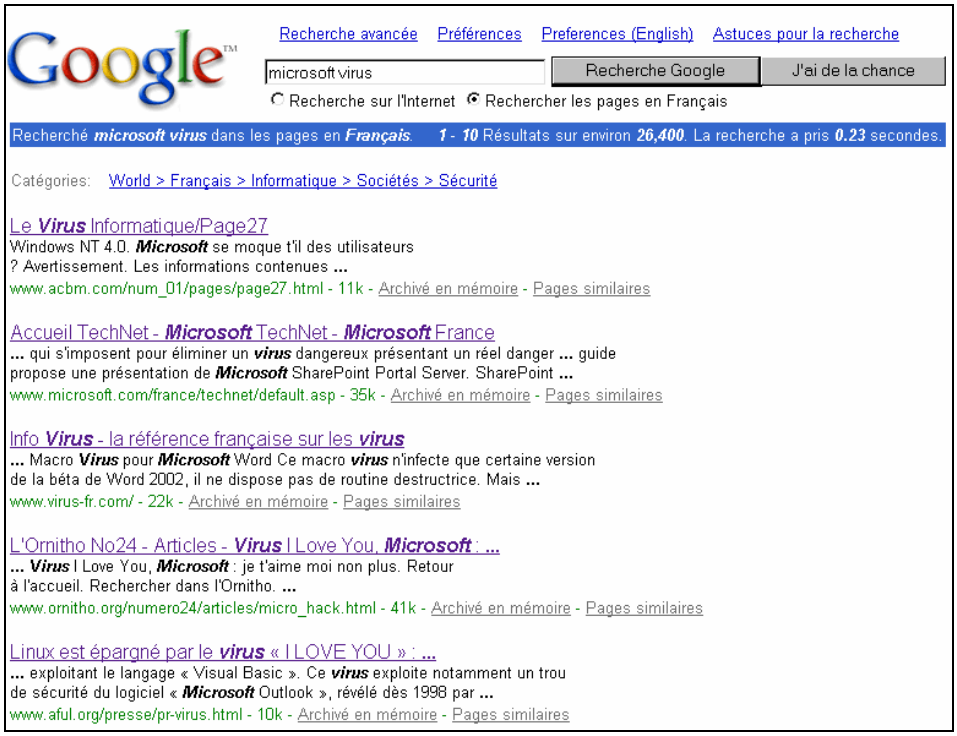

Figure 17 -- *Les 5 premières URL obtenues suite à l'interrogation du moteur de recherche Google™ avec la requête 'microsoft' et 'virus*

'.

Pour les annuaires de recherche les manques proviennent essentiellement des techniques de catégorisation employées : comment décider de la catégorie dans laquelle doit apparaître un document ? Soit ce choix est effectué à la main[16] et est donc très coûteux en temps et en main d'œuvre qualifiée, soit il est assuré par des agents logiciels et l'on se heurte aux mêmes problèmes que pour les moteurs de recherche (utilisation des méta-données subjectives ou trompeuses pour l'indexation, analyse partielle des documents par les robots, polysémie…) (Perlerin pp. 6-12, 2000). Le principe de catégorisation proposé par les annuaires de recherche est purement ontologique et les problèmes abordés dans la section 4.2. s'y retrouvent donc. Par exemple, les documents placés dans une catégorie « Arts » dans un site donné pourront être fort différents de ceux placés dans une catégorie éponyme sur un site différent (Chaffee et Gauch, 2000). Il ne s'agit pas uniquement de

[16] Voir [Salton, 1966] [Salton, 1986] pour une étude sur les limite de l'indexation manuelle.

considérations subjectives quant à la catégorisation d'un document, mais de l'inadéquation d'une pratique de catégorisation générique pour une recherche finalisée.

Certaines propositions tendent à dépasser la simple utilisation du *pattern matching* dans les moteurs de recherche grâce à des ontologies (Pretschner et Gauch, 1999), des thésaurus (Yang et al., 1998) ou des techniques d'indexation en langage naturel (logiciel Tropes d'Acetic[17]). Ils retrouvent alors les mêmes problèmes que les annuaires pour classer les documents. Or la question de la généralité du sens peut être posée autrement. Dans (Rastier et al., 1994), les auteurs décrivent les degrés de systématicité de la sémantique unifiée comme étant : le système fonctionnel de la langue (le dialecte qui pose les normes de la langue), les normes sociales (le sociolecte qui découle de considérations essentiellement culturelles) et les normes idiolectales (l'idiolecte lié au locuteur-individu et qui représente l'ensemble de ses régularités personnelles ou « normes individuelles » dans ses actes linguistiques). Rastier avoue lui-même que *« la description de leur interaction pose des problèmes délicats qui ne peuvent être éludés »*. Notre approche anthropocentrée[18] trouve sa justification dans cette remarque. Nous pensons que la recherche documentaire a de nombreux points communs avec une activité linguistique et que les critères énoncés par Rastier doivent prendre part au modèle que nous proposons. Lors de la construction des tables Anadia, l'utilisateur imprime ses normes dialectales : il choisit la langue qu'il désire voir utilisée dans les documents qu'il recherche, il restreint le sens de certains mots polysémiques… Ce qui relève des normes idiolectales est implicite au modèle, car les relations sémantiques sont construites par un individu ou un groupe d'individus. Les normes sociolectales font appel à une activité particulière de la part de l'utilisateur. Dans le cadre de la recherche documentaire, celui-ci se trouve en effet dans une situation dialogique dont il est le seul acteur : il doit prendre en compte pour son objectif, un auteur qu'il ne connaît pas à priori (à travers des documents dont il ne connaît pas à priori l'existence) mais avec qui il partage un centre d'intérêt : le sujet de sa recherche. L'individu construisant ces relations sémantiques doit tenir compte d'une certaine utilisation possible de la langue par un auteur. Cette remarque relève de la définition de la langue en tant que phénomène social (et donc partagé). Lorsqu'on utilise des ontologies ou des thésaurus pré-définis, les normes idiolectales et sociolectales sont délaissées : l'utilisateur doit s'adapter aux définitions préexistantes quelle que soit sa recherche et quel que soit son profil. Certaines méthodes plus récentes de construction des ontologies tendent à replacer l'utilisateur au cœur de la recherche à travers l'analyse de son comportement ou la construction de plusieurs ontologies mises en correspondance lors d'une recherche (une ontologie du site construite par un agent, une ontologie personnelle de l'utilisateur) (Obiwan, 2000) (Chaffee, 2000).

5.2.2. Nos propositions

Les tables Anadia engagent une implication importante de l'utilisateur qui, a travers une co-construction du sens de mots avec la machine, définit ses propres relations sémantiques qui serviront à l'analyse des documents. Comme Anadia est un modèle de la mémoire sémantique basé sur la notion de valeur, c'est cette valeur qui ; à travers les tables construites, va nous permettre de décider de la pertinence d'un document par rapport à une recherche et de la place à donner à ce document dans la liste des résultats proposée. Le nombre de tables devant être pris en compte pour une recherche est défini par l'utilisateur qui fixe ainsi lui-même les considérations sémantiques à prendre en compte.

Concrètement, nous proposons d'apporter une valeur ajoutée aux systèmes de recherche documentaire existants en y ajoutant un filtrage et un ré-ordonnancement des documents trouvés par un moteur de recherche en n'utilisant que des données fournies par l'utilisateur (Figure 18). L'objectif de notre étude est de confronter le modèle Anadia à une tâche nécessitant une compétence interprétative de la part des machines. Nous souhaitons dépasser la *sémantique lexicale* couramment utilisée dans ce champ d'application pour aboutir à l'utilisation d'une *sémantique des textes* et accroître par ce biais, à la fois la qualité des résultats et la qualité de leur présentation aux usagers.

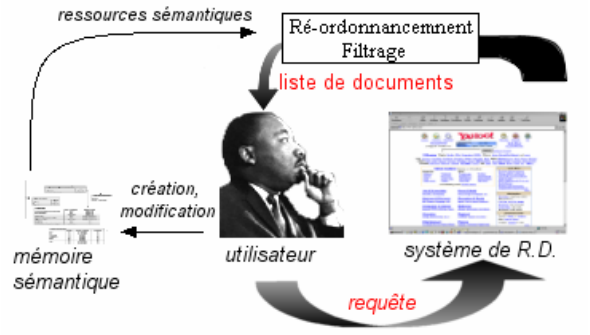

Figure 18 -- *Utilisation d'une mémoire sémantique (Dispositifs Anadia) pour la RD*

Le filtrage et le ré-ordonnancement proposés se basent sur le repérage des isotopies dans les

[17] www.acetic.fr/

[18] Théodore Thlivitis [Thlivitis, 1998] indique que dans les systèmes anthropocentrés, *« ce n'est pas l'homme qui essaie d'* entrer *dans un monde informatique quasi-autosuffisant, mais la machine qui se construit autour des besoins précis de l'homme pour mieux l'assister. »*

documents en fonction des dispositifs construits et des tables sélectionnées pour une recherche donnée. Le repérage des isotopies dans une portion de texte ou à l'intérieur d'un document entier permet deux opérations majeures pour la recherche documentaire : la désambiguïsation sémantique et la recherche d'impressions référentielles.

La désambiguïsation sémantique :

Le mot 'virus' est polysémique. Même si une approche diachronique permet de rapprocher les deux emplois, son utilisation en tant que terme médical est différente de celle en tant que terme informatique. Dans le cadre d'une recherche documentaire, une requête incluant les mots 'virus' et 'microsoft' peut aboutir à deux textes présentant ces deux termes en nombres d'occurrences identiques mais recelant en fait deux utilisations distinctes du mot 'virus' :

(5) Un peu plus tard, l'équipe de Robert Gallo au National Cancer Institute (NCI) et celle de Jay Levy à l'université de Californie à San Fransisco isolèrent un rétrovirus sur des patients atteints du sida et des personnes en contact avec des malades. Les trois équipes isolèrent ce qu'on appelle maintenant le VIH, l'agent étiologique du sida. Un second ***virus*** *du sida humain (VIH2) sera découvert trois ans plus tard par l'équipe de l'institut Pasteur. (Source : Sida, Encyclopédie* ***Microsoft*** *Encarta® 1997).* (www.ressy.org/public/spsfe/sida.htm)

(6) Au mois de mars dernier, le ***virus*** *I love you aurait touché près de 50 millions d'ordinateurs équipés de la messagerie Outlook. Exploits de pirates surdoués ou failles grossières de sécurité dans les produits* ***Microsoft****? Hubert Tournier, spécialiste en sécurité internet, répond aux questions de l'Ornitho et fait le point sur les pirates qui écument l'Internet.* (www.ornitho.org/numero24/articles/micro_hack.html)

Le repérage des isotopies de ces deux fragments de texte peut permettre la levée de l'ambiguïté du terme ou au moins, de ne retenir que l'un des deux fragments dans le cadre d'une recherche précise. Dans le cas d'une recherche sur les 'virus informatiques' et leurs relations avec 'microsoft', les deux termes doivent apparaître au sein d'un même dispositif (Figure 20). Dans ce cas de figure, on repère dans le premier fragment de texte une isotopie[19] n'ayant pour support que deux mots : 'virus' et 'microsoft', alors que le second fragment en présente une en ayant au moins 6 mots supports. Alors que l'utilisation d'ontologies ou de thésaurus aurait fortement augmenté le nombre de mots à prendre en compte pour le *pattern matching* nécessaire à l'analyse des documents, l'utilisation d'un dispositif Anadia regroupant par essence un nombre limité de termes liés sémantiquement permet une désambiguïsation sémantique au moins de qualité équivalente.

Envisagée indépendamment d'une tâche précise (comme pour son évaluation dans le projet SENSEVAL[20]), la désambiguïsation sémantique nécessite de larges bases de données comme des dictionnaires ou des thésaurus. Pour la recherche documentaire, l'objectif de la recherche circonscrit le domaine. Dans notre modèle, la désambiguïsation sémantique consiste alors, dans le cas où un même mot serait décrit de façon à exprimer plusieurs sens différents, à ne retenir qu'un seul de ces sens en fonction des isotopies découvertes. Il ne s'agit pas d'aboutir à une levée exhaustive de tous les cas d'ambiguïté sémantique, mais de ne retenir que les sens décrits par l'utilisateur et retenus par lui dans le cadre d'une recherche donnée pour décider de la pertinence d'un document. Les ressources utiles à cette tâche sont donc réduites à celles fournies par l'utilisateur.

La recherche d'impressions référentielles :

Les isotopies relevées dans l'exemple précédent permettent également la découverte des effets de sens au sein des fragments de texte analysés. Au vu des sèmes mis en jeu dans les isotopies repérées, un module d'analyse peut conclure que l'on parle dans le premier fragment de 'virus' en terme de maladie et dans le second en terme de programme informatique (pour peu que des dispositifs Anadia correspondant à ces deux sujets aient été utilisés pour l'analyse). Pour les textes, l'analyse à l'échelle de la phrase, du paragraphe, du document ou d'un ensemble de documents permet la découverte de faisceaux d'isotopies. C'est en fonction de la présence et de la couverture de ces faisceaux que la pertinence d'un document par rapport aux attentes de l'utilisateur est décidée. L'impression référentielle calculée à partir des isotopies découvertes au sein des documents est donc au cœur du filtrage et du ré-ordonnancement proposé. La machine met ainsi en place une interprétation sélective et supervisée car l'utilisateur a toujours la possibilité de modifier son dispositif pour affiner l'analyse sémantique et de paramétrer au mieux l'outil logiciel par rapport à ses exigences (Section 5.3).

L'analyse des documents est effectuée en deux étapes. La première consiste à en repérer les différentes parties (paragraphes, tableaux…) en fonction du codage utilisé (HTML, XML…), la seconde à localiser les isotopies dans chacune de ces parties. En premier lieu, cette analyse peut aboutir à l'élaboration de rapports d'exploration regroupant à la

[19] Il s'agit ici d'isotopies faisant intervenir plusieurs tables de niveaux différents dans même dispositif.

[20] www.itri.bton.ac.uk/events/senseval/

fois une représentation graphique du document et une liste des thèmes abordés. Les isotopies utiles à cette tâche peuvent se limiter à celles mettant en jeu des mots appartenant à une même table Anadia. Dans les rapports, les parties des documents sont numérotées de bas en haut et de gauche à droite. Dans les graphiques, les traits horizontaux préfigurent les isotopies repérées. Ces traits ont la même couleur que la table correspondante dans le dispositif :

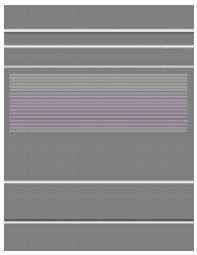

Figure 19 – *Représentation graphique du rapport d'exploration du document 1.*

Document 1 Rapport d'exploration :

Nombre de mots : 1045
Nombre d'occurrences du mot 'virus' : *0*
Nombre d'occurrences du mot 'microsoft' : *9*
Thèmes abordés
Partie n°3 :
//Matériaux de l'informatique// (6 mots supports pour l'isotopie) - //Systèmes d'exploitation// (11 mots supports) - //Acteurs de l'informatique// (1 mot support)

Le thème des virus informatiques n'est jamais abordé dans le document 1. Seul le thème des systèmes d'exploitation (correspondant à la table //Systèmes d'exploitation// du dispositif dans laquelle se trouve le mot 'microsoft') est présent.

// Traitement de l'information//

	outils
informatique	électroniques

// Acteurs de l'informatique//

	type d'action
programmeur, analyste	créer
utilisateur, client	utiliser

//matériaux de l'informatique//

	type	rôle
supports de données	matériel	ressource
données	logiciel	ressource
machine, ordinateur	matériel	traitement
programmes, logiciel	logiciel	traitement

// programmes informatiques//

	rôle
utilitaires, logiciel	accessoire
systèmes d'exploitation	indispensable

// composants physiques des machines info//

	rôle
périphériques	accessoire
carte mère et processeurs	indispensable

// Pirates informatiques //

	Prédilection
hacker, cracker, pirate	données
phreaker, pirate	téléphone
coder, pirate	virus
carder, pirate	banques

//support de données//

	capacité
disquette	faible
zip	moyenne
disque dur	élevée
réseau, Internet, Ethernet	illimitée

// Virus informatiques //

	// Support //
macro-virus, virus	macro
vers, virus, worms	programme
cheval de Troie, virus	macro. ou prog

//systèmes d'exploitation//

	portable	interface graphique	multitâche
linux, Unix	oui	non généralisée	oui
dos	oui	non généralisée	non
windows nt, mucrosoft	oui	généralisée	oui
windows , mucrosoft	oui	généralisée	non
MacOs	non	généralisée	oui
	non	généralisée	non
	non	non généralisée	non
	non	non généralisée	oui

Figure 20 -- *Dispositif Anadia en rapport avec l'informatique et le piratage.*

Document 2 Rapport d'exploration :

Nombre de mots : 855
Nombre d'occurrences du mot 'virus' : *3*
Nombre d'occurrences du mot 'microsoft' : *22*
Thèmes abordés :
Partie n°7 : //Systèmes d'exploitation// (3 mots supports)
Partie n°8 : //Systèmes d'exploitation// (1 mot support)
Partie n°9 : //Systèmes d'exploitation// (4 mots supports)
Partie n°10 : //Systèmes d'exploitation// (4 mots supports)
Partie n°11 : //Systèmes d'exploitation// (1 mot support)
Partie n°12 : //Virus informatiques// (2 mots supports)

Les deux thèmes //Virus informatiques// et //Systèmes d'exploitations// sont abordés mais dans des parties distinctes à l'intérieur du document 2. Seule une isotopie en rapport avec les virus informatiques (avec 2 mots supports) est présente dans tout le document.

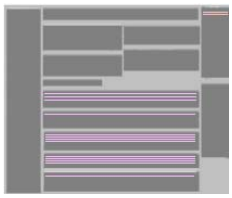

Figure 21 – *Représentation graphique du rapport d'exploration du document 2.*

Document 3 Rapport d'exploration :

Nombre de mots : 711
Nombre d'occurrences du mot 'virus' : *12*
Nombre d'occurrences du mot 'microsoft' : *2*
Thèmes abordés :
Partie n°6 : //Virus informatiques// (1 mot support)
Partie n°7 : //Virus informatiques// (1 mot support)
Partie n°8 : //Virus informatiques// (1 mot support)
Partie n°9 : //Virus informatiques// (1 mot support)
Partie n°10 : //Virus informatiques// (1 mot support)
Partie n°12 : //Virus informatiques// (1 mot support)

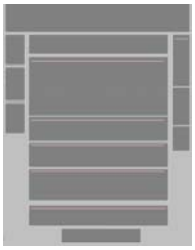

Figure 22 – *Représentation graphique du rapport d'exploration du document 3.*

Aucune isotopie en rapport avec les systèmes d'exploitation n'a été détectée. La présence de deux occurrences du mot 'microsoft' n'a pas amené au repérage d'une isotopie car le terme apparaît dans des parties distinctes.

Document 4 Rapport d'exploration :

Nombre de mots : 4074
Nombre d'occurrences du mot 'virus' : *18*
Nombre d'occurrences du mot 'microsoft' : *20*
Thèmes abordés :
Partie n°2 : //Pirates informatiques// (1 mot supports) - //Support de données// (1 mot supports)
Partie n°3 : //Support de données// (2 mots supports)
Partie n°4 : //Pirates informatiques// (10 mots supports)
Partie n°5 : // Systèmes d'exploitation // (7 mots supports) - //Virus informatiques// (1 mot supports)
Partie n°6 : // Systèmes d'exploitation // (20 mots supports) - //Virus informatiques// (11 mots supports) - //Acteurs de l'informatique// (3 mots supports) - //Supports de données// (1 mot support) - //Technologies de l'information// (1 mot support)
Partie n°7: //Pirates informatique// (2 mots supports) – // Systèmes d'exploitation // (2 mots supports) - //Programmes informatiques// (1 mot support)
Partie n°8 : //Pirates informatique// (1 mot support) - //Matériaux de l'informatique// (1 mot support)
Partie n°9 : //Traitement de l'information// (1 mot support)
Partie n°10 : //Supports de données// (2 mots supports) - // Systèmes d'exploitation // (1 mot support)
Partie n°11 : // Matériaux de l'informatique// (6 mots supports) - //Supports de données// (5 mots supports) - //Traitement de l'information// (2 mots supports) - //Pirates Informatiques// (1 mot support)
Partie n°12 : //Supports de données// (2 mots supports)

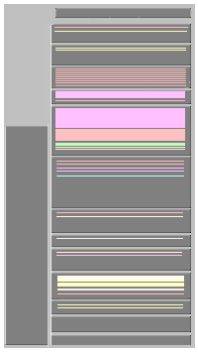

Figure 23 – *Représentation graphique du rapport d'exploration du document 4.*

Les thèmes //Virus informatiques// et //Systèmes d'exploitation// co-existent au sein de deux parties (5 et 6) et se retrouvent isolément dans d'autres parties du texte.

Document 5 Rapport d'exploration :

Nombre de mots : 1180
Nombre d'occurrences du mot 'virus' : *16*
Nombre d'occurrences du mot 'microsoft' : 9
Thèmes abordés :
Partie n°1 : //Systèmes d'exploitation// (1 mot support) - //Virus Informatiques// (1 mot support)
Partie n°2 : //Systèmes d'exploitation// (23 mots supports) - //Virus Informatiques// (15 mots supports) - //Matériaux de l'informatique// (13 mots supports) - //Acteurs de l'informatique// (4 mots supports)

Figure 24 – *Représentation graphique du rapport d'exploration du document 5.*

Les thèmes //Virus informatiques// et //Systèmes d'exploitation// co-existent au sein de deux parties du document (1 et 2).

Les informations recueillies à la suite de l'analyse des documents sont le matériau utilisé pour le filtrage et le ré-ordonnancement. Dans notre exemple, la priorité pourra être donnée aux documents recelant des parties où les thèmes //Virus informatiques// et //Systèmes d'exploitation// sont présents. Il s'agit ici

des documents 4 et 5 qui correspondent véritablement aux attentes d'un utilisateur voulant avoir des informations sur Microsoft et les virus informatiques. L'ordonnancement des résultats pourra déprendre du nombre de mots supports de ces isotopies dans le document (sans pondération en fonction du nombre de mots des documents, l'ordre serait ici document 5 puis document 4).

Ce travail est en cours de réalisation à partir de techniques ayant déjà fait l'objet d'études approfondies (recherche de mots, manipulations de documents…) et offrant des procédés d'implémentation éprouvés et efficaces. Nous sommes de ce fait conscients que nombre de problèmes devront être résolus avant de pouvoir fournir un outil satisfaisant à des usagers. Il s'agira probablement d'usagers professionnels répétant souvent les mêmes requêtes, comme en veille technologique, pour que le coût de construction des tables soit compensé par la qualité et la précision des résultats obtenus.

5.3. Aide à la création des tables Anadia.

Comme les utilisateurs ne sont pas de bons constructeurs de ressources sémantiques, parce que leurs connaissances sont largement implicites, nous devons explorer des voies de recherche pour l'aide à la construction des tables Anadia. Les tables Anadia créées par un utilisateur pour une tâche donnée présentent une double articulation remarquable. Elles reflètent à la fois une certaine appropriation de la langue par l'auteur des tables (normes idiolectales) et l'essence même de la langue : son caractère social (normes sociolectales et dialectales). Le processus d'aide que nous proposons prend en compte ces deux aspects. Il ne s'agit pas ici de tendre vers un colossal dispositif exhaustif et universel permettant de décrire le monde, l'univers et le reste mais de construire des relations de catégorisations décrivant des significations utiles pour une tâche et valables pour un utilisateur ou une communauté restreinte d'utilisateurs. Ainsi, les aides proposées ne sauraient être efficientes que dans un processus laissant le dernier choix à l'utilisateur et dont la substance serait le matériau linguistique observé dans des conditions précises d'utilisation. Comme nous avons pu le voir dans l'exemple précédent, l'utilisation d'un dispositif peut amener à le modifier. La cohérence globale du réseau étant maintenue par le logiciel Anadia, un utilisateur peut à loisir modifier ses tables et les relations qui les joignent à la vue des résultats de leur utilisation.

Cependant, pour amorcer le processus ou pour parvenir à trouver des catégories dont l'utilisateur n'avait pas conscience lors de la constitution d'un dispositif, nous avons esquissé les prémisses d'un système d'aide semi-automatique basé sur l'étude d'un corpus de textes. Pour cela, nous avons étudié statistiquement un ensemble de 1783 dépêches journalistiques au format HTML en relevant toutes les formes lexicales présentes dans différents échantillons et en calculant les co-occurrences de certaines de ces lexies. Cette étude avait un triple objectif (Perlerin pp.35-60, 2000):

1. étudier un ensemble de textes traitant tous d'un même sujet technique (il s'agissait ici de dépêches sur la finance, l'assurance et les opérations boursières) pour pouvoir mieux connaître le matériau linguistique,
2. créer des tables Anadia cohérentes vis-à-vis d'un domaine technique qui nous était étranger pour mener à bien nos expériences futures,
3. observer la possibilité d'une aide à la création des tables Anadia.

   Nous nous focaliserons ici sur ce dernier point.

La première étape de cette étude s'est basée sur la loi de Zipf : George Kingsley Zipf (1902-1950) (Zipf, 1949) a repris et développé les travaux initiés par certains cryptographes et le sténographe Baptiste Estoup relatifs aux fréquences relatives des mots dans un texte. Si l'on dresse une table de l'ensemble des mots différents d'un texte quelconque, classés par ordre de fréquences décroissantes, on constate que la fréquence d'un mot est inversement proportionnelle à son rang dans la liste, ou autrement dit, que le produit de la fréquence de n'importe quel mot par son rang est constant : ce que traduit la formule f * r = C, où f est la fréquence et r le rang. Cette égalité, qui n'est vraie qu'en approximation, est indépendante des locuteurs, des types de textes et des langues. Il semble ainsi qu'il s'agisse véritablement d'un trait général des énoncés linguistiques (il s'avère par ailleurs que cette loi s'applique dans de nombreux autres domaines comme le phénomène de l'immigration[21] ou encore la popularité des pages du W3[22]). Cette théorie a permis entre autre de mettre au jour que dans les listes créées selon ce principe, on rencontrait (approximativement) tout d'abord les mots grammaticaux et par la suite les mots représentatifs du domaine (s'il s'agit d'un texte ou d'un corpus issu d'un domaine technique ou tout du moins restreint) (Giguet, 1998) (Figure 25).

Cette loi nous a donc permis d'extraire la liste des mots du domaine de notre corpus (et au passage d'en apprécier son homogénéité en fonction du nombre d'occurrences de ces mots à forte redondance). En fonction de ces résultats, nous avons construit un ensemble de tables cohérentes avec le domaine,

---

21 http://www.rri.wvu.edu/WebBook/Goetz/Migx2.htm

22 http://www.useit.com/alertbox/zipf.html - Nous proposons au lecteur intéressé par « les lois de distributions

exploitables pour les expériences relatives à l'aide à la création des tables de catégorisation.

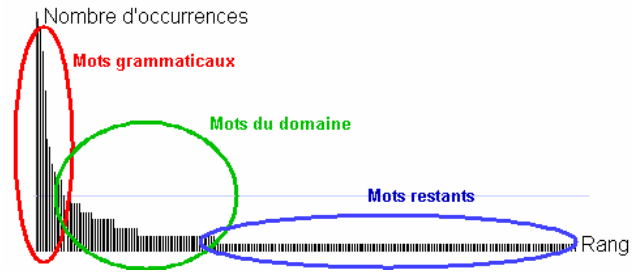

Figure 25 -- *Loi de Zipf.*
Utilisation pour l'étude d'un corpus homogène

L'ensemble des lexies des tables ainsi construites a été rassemblé dans un tableau présentant leurs co-occurrences en pourcentage du nombre de fichiers (Figure 26).

De nombreuses expériences nous ont permis d'envisager une méthode se basant sur le tableau ci-dessus pour proposer une aide à la création des tables. Dans le cas où un utilisateur se trouve dans l'impossibilité de trouver une catégorie satisfaisante correspondant à une ligne d'une table qu'il a déjà partiellement remplie Figure 27), Nous pouvons en nous basant sur le tableau des co-occurrences (Figure 26), lui proposer la liste des lexies présentant les plus fortes co-occurrences avec celles déjà présentes dans la table.

	société	**banque**	**groupe**	**filiale**	**établisse-ment**	**entreprise**	**compagnie**	**agence**	**pôle**	**bancassu-rance**	**holding**	**réseau**	**assurance**
société		77%	66%	18%	40%	43%	18%	11%	9%	5%	23%	28%	27%
banque	80%		49%	18%	44%	40%	21%	12%	9%	6%	18%	27%	30%
groupe	75%	54%		40%	21%	33%	24%	8%	14%	6%	22%	26%	39%
filiale	46%	44%	90%		11%	21%	17%	6%	14%	6%	25%	30%	45%
établissement	95%	100%	44%	10%		64%	16%	14%	4%	1%	12%	29%	19%
entreprises	88%	78%	58%	17%	55%		17%	17%	12%	4%	9%	35%	15%
compagnie	69%	75%	79%	25%	25%	31%		10%	21%	13%	31%	35%	60%
agence	91%	91%	57%	17%	48%	65%	22%		13%	13%	22%	52%	30%
pôle	63%	59%	85%	37%	11%	41%	37%	11%		22%	33%	22%	48%
bancassurance	82%	100%	82%	36%	9%	36%	55%	27%	55%		64%	36%	91%
holding	93%	69%	78%	40%	20%	18%	33%	11%	20%	16%		18%	67%
réseau	90%	81%	71%	36%	38%	53%	29%	21%	10%	7%	14%		31%
assurance	60%	63%	76%	39%	18%	16%	35%	9%	16%	12%	37%	22%	

Figure 26 -- *Table des co-occurrences en pourcentage de fichier de toutes les tables construites. Le tableau se lit : « Dans 77% des fichiers où société apparaît, banque est présent également. »*

	Cadre juridique	***Domaine de prédilection***
société	défini	aucun
entreprise	indéfini	aucun
établissement	indéfini	spécialisé
	défini	spécialisé

Figure 27 -- *Exemple d'une table incomplète : le mot manquant initialement présent dans la table est « compagnie ».*

Dans le cas de la figure 24, le logiciel retourne la liste suivante (le nombre indiqué entre parenthèses représente la somme des pourcentages de co-occurrence) :

agence (204)
réseau (181)
compagnie (169)
banque (164)
groupe (129)
holding (131)
…

Figure 28 -- *Propositions de lexies pour la table incomplète (Figure 27)*

On notera ici que le mot manquant de la table, '*compagnie'* est proposé en 3ème position sur les 13 trouvées (toutes n'ont pas été inscrites ici). L'expérience a été réitérée sur l'ensemble des tables construites pour nos expériences. Les mots manquants étaient alors systématiquement présents au moins dans les 10 premiers de la liste proposée par le logiciel.

L'étude d'un corpus et les tables construites par nos soins à l'issus de cette étude donnent des résultats encourageants. Ils ne sauraient être pleinement validés qu'à la suite d'une expérimentation à plus grande échelle et par rapport à des tables construites par de « véritables » utilisateurs, expérimentation qui doit être mise en place cette année avec des psychologues.

6. Conclusion

Rappelons l'hypothèse dont nous sommes partis : la mémoire est fondée sur une activité de catégorisation de type algébrique. Les exemplaires d'une catégorie n'y sont pas rassemblés selon des critères de ressemblance comme dans les modèles ensemblistes. Ils sont regroupés en fonction des différenciations qui définissent les catégories. Cette distinction est importante car selon un point de vue ensembliste, ce sont les objets qui sont premiers et les regroupements d'objets qui sont des artéfacts tandis que selon un point de vue algébrique ce sont les catégories qui sont premières et qui déterminent les possibles places pour présenter et différencier les objets qui doivent l'être. Ainsi, il n'est pas question de tout représenter mais seulement ce qui doit être distingué à un moment donné de l'activité selon un principe opportuniste. De plus, une catégorisation algébrique produit des représentations systémiques dans le sens où les connaissances ne sont pas représentées indépendamment les unes des autres mais où chacune tient une place dans un réseau qu'elle forme avec les autres. D'une certaine façon, chaque place inclut les autres par défaut, idée que l'on retrouve également dans la notion de valeur chez Saussure.

À partir de cette hypothèse, nous avons construit un modèle de mémoire **unifiée** dans le sens où les principes et la méthode sont valides tant pour la représentation de connaissances extra-linguistiques (la « mémoire des choses ») que pour la représentation de terminologies (la « mémoire des mots »). Plutôt que de chercher à appliquer à la représentation du matériau lexical les principes ontologiques de la représentation de connaissances, nous avons fait le choix inverse qui consiste à appliquer à la représentation de catégories extra-linguistiques les propriétés systémiques de la sémantique des langues. Bien sûr la classification scientifique des êtres vivants par exemple, n'a rien à voir la classe lexicale des animaux. Il n'est pas ici question de plaider en faveur d'une homogénéité des formes de représentation (extra-linguistique ou intra-linguistique) car les activités qui mettent en œuvre ces représentations ne sont pas les mêmes, mais la classification scientifique a sa source dans le langage.

Si les dispositifs Anadia sont un analogue de la mémoire sémantique, la question de leur pertinence pour la psychologie et la psycholinguistique devra être posée. Cette question n'est pas aujourd'hui notre objet d'étude, nous cherchons à concevoir des agents logiciels qui puissent entrer dans des interactions langagières au même titre que leurs partenaires humains. L'analogie que nous visons est donc d'abord une similitude fonctionnelle entre la machine et l'humain. Le monde des machines n'est pas le même que le nôtre, elles n'ont une expérience propre que des flux d'entrée et de sortie, de l'état de leurs registres mémoire, et de leurs processus de calcul. Mais nous faisons l'hypothèse que c'est à partir de cette expérience propre que le même processus de catégorisation différentielle et de reconstruction peut fonctionner. Par exemple, via l'écran, elles sont dans l'espace et peuvent distinguer le haut et le bas, la droite et la gauche, les couleurs. Via les événements et les flux d'entrée, elles sont dans le temps et peuvent distinguer l'avant et l'après.

La mémoire d'un agent logiciel ne peut se limiter à une structure où toutes les choses puissent être représentées. Si l'on veut mettre des machines dans des situations d'interaction et de dialogue, alors la mémoire des agents logiciels doit être le modèle d'une activité. Cette activité est une **activité sémiotique** dans le sens où elle consiste en un aller-retour sans fin entre des expressions et des contenus. Les dispositifs Anadia, à travers les attributs, les registres, les tables et les topiques forment un cadre de représentation pour cette activité. Les objets, processus, événements et faits inscrits dans ces dispositifs acquièrent le statut d'objets sociaux. Il sont liés les uns aux autres. Ils sont situés dans un temps, une interaction et des partenaires et sont les supports des possibles co-références entre un humain et une machine.

La construction de ce modèle va comporter plusieurs phases d'amorçage dont l'expérimentation permet l'évolution. Seule la première phase a été réalisée et est présentée dans cet article. En attendant une

réalisation plus complète d'agents logiciels dialoguant en langue naturelle, Anadia est déjà un bon cadre de travail pour la conception des logiciels interactifs. Actuellement, tout le travail nécessaire d'adaptation entre le mode de fonctionnement logique de la machine et le mode de fonctionnement sémiotique des usagers est à faire par le programmeur ou l'ingénieur cogniticien, quand on ne demande pas à l'usager de le faire lui-même. Anadia peut fournir au concepteur la possibilité de tester l'adéquation de ses représentations à son projet en vérifiant que les critères de représentation retenus sont tous nécessaires, et qu'ils sont suffisants pour différencier tous les objets qui doivent l'être, et en ayant un support pour les faire valider par les usagers. Ainsi, ce processus de catégorisation interactif peut servir à analyser et à structurer les concepts d'un domaine avant de concevoir un logiciel à base de connaissances.

Remerciements

Ce travail a été soutenu par le GIS Sciences de la Cognition.

Les auteurs tiennent également à remercier Georges Perec, qui les a inspirés dans ce travail, en particulier à travers ces lignes :

« Il y a dans toute énumération deux tentations contradictoires ; la première est de TOUT recenser, la seconde d'oublier tout de même quelque chose ; la première voudrait clôturer définitivement la question, la seconde la laisser ouverte ; entre l'exhaustif et l'inachevé, l'énumération me semble ainsi être, avant toute pensée (et avant tout classement), la marque même de ce besoin de nommer et de réunir sans lequel le monde (« la vie ») resterait pour nous sans repères : il y a des choses différentes qui sont pourtant un peu pareilles ; on peut les rassembler dans des séries à l'intérieur desquelles il sera possible de les distinguer.

Il y a dans l'idée que rien au monde n'est assez unique pour ne pas pouvoir entrer dans une liste, quelque chose d'exaltant et de terrifiant à la fois. »

Georges Perec, Penser/classer Hachette Littératures 1985 p.167.

Références Bibliographiques

[Adam-Nicolle, 1990] Adam-Nicolle A. (1990). Le métalangage Airelle, *Revue Française d'Intelligence Artificielle.* Vol. 4, n°1.

[Anderson, 1983] Anderson J. R. (1983).The architecture of cognition. MA: Harvard University Press : Cambridge.

[Aristote] Aristote. Organon I. Les catégories. Trad. J. Tricot (1969). Coll. Bibliothèque des textes philosophiques. Ed. VRIN : Paris.

[Assadi, 1998] Assadi H. (1998). Construction d'ontologies à partir de textes techniques - Application aux systèmes documentaires . Thèse de doctorat de l'Université de Paris 6.

[Bachimont, 1996] 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 de Paris.

[Bellot et Elbèze, 2000] Bellot P., Elbèze M. (2000). Classification locale non supervisée pour la RD. Revue T.A.L. Vol. 41. 335-365.

[Beust et al., 1997] Beust P., Delepine L., Jacquet D. (1997). Rédacteur, Utilisateur, Concepteur : Quelle référence commune ? actes 01DESIGN'97 Théoule-sur-mer. France. 157-162.

[Beust et al., 1998] Beust P., Nicolle A. (1998). Une sémantique interactionniste pour le dialogue homme-machine. Actes du colloque de l'Association pour la Recherche Cognitive ARC'98. Saint-Denis. France. 205-211.

[Beust, 1998] Beust P. (1998). Contribution à un modèle interactionniste du sens. Thèse de doctorat d'informatique de l'Université de Caen. http://users.info.unicaen.fr/~beust/These/these.html

[Bourdon, 1992] Bourdon F. (1992) Un modèle de dérive de connaissances - Application en bureautique. Thèse de doctorat de l'Université du Maine au Mans.

[Brixhe, et al., 1994] Brixhe D., Saint-Dizier V., Trognon A. (1994). Résolution interlocutoire d'un diagnostic, dans Modélisations d'explications sur un corpus de dialogue. Revue Télécom-Paris. Vol. 1, 27-47.

[Brooks, 1991] Brooks R.A. (1991). Intelligence without Representation. *Artificial Intelligence Journal.* Vol. 47, 139-159.

[Castoriadis, 1975] Castoriadis C. (1975). L'institution imaginaire de la société. Seuil : Paris.

[Chaffee et Gauch, 2000] Chaffee J., Gauch S. (2000). Personal Ontologies for Web navigation. actes CIKM'00. McLean, VA, USA. 227-234.

[Chaffee, 2000] Chaffee J. (2000). Personal Ontologies for Web Navigation. Thèse de l'Université du Kansas.

[Coursil et al, 1995] Coursil J., Montlouis C.D., Delépine L., Beust P. (1995) Anadia 1, manuel de l'utilisateur. Université des Antilles-Guadeloupe.

[Coursil, 1993] Coursil J. (1993). Essai d'Intelligence Artificielle et de Linguistique Générale. Rapport pour

l'obtention du diplôme d'habilitation à diriger des recherches. Université de Caen.

[Coursil, 2000] Coursil J. (2000). La fonction muette du langage. Presses Universitaires Créoles. Ibis Rouge Éditions : Guadeloupe.

[Davis et al., 1993] Davis R., Shrobe H., Szolovits P. (1993). What Is a Knowledge Representation. *AI Magazine*. Vol. 14, 17-33.

[Deleuze, 1972] Deleuze G. (1972). À quoi reconnaît-on le structuralisme ? In Châtelet F. (ed.) *La philosophie au XX^{e} siècle.* , réédition Marabout : Paris, Tome IV. 293-329.

[Dubois, 1991] Dubois D. (1991). Les catégories sémantiques naturelles : prototype et typicalité. in D. Dubois (ed.). *Sémantique et cognition : Catégories, concepts et typicalité*, Éditions du CNRS : Paris. 16-27.

[Eco, 1988] Eco U. (1988). Sémiotique et philosophie du langage. PUF : Paris.

[Fahlman, 1979] Fahlman S. E. (1979). NETL, a system for representing and using Real-Word knowledge. MIT Press : Cambridge.

[Ferber, 1995] Ferber J. (1995). Les systèmes multi-agents - Vers une intelligence collective. InterEditions : Paris.

[Giguet, 1998] Giguet E. (1998). Méthode pour l'analyse automatique de structures formelles sur documents multilingues. Thèse de doctorat d'informatique de l'Université de Caen.

[Golberg et Robson, 1984] Golberg A., Robson D. (1984). Smalltalk-80, the langage and its implementation. Addison-Wesley : Reading, Massachusetts.

[Grégori, 1999] Grégori N. (1999). Étude clinique d'une situtation de conception de produit – Vers une pragmatique de la conception. Thèse de doctorant de l'Université de Nancy 2.

[Greimas, 1966] Greimas A.J. (1966). Sémantique Structurale. PUF : Paris.

[Guéna, 1997] Guéna F. (1997). Le raisonnement par Classification appliqué à la CAO. Rapport d'Habilitation à Diriger les Recherches de l'Université de Caen.

[Hjelmslev, 1943] Hejlmslev L. (1943). Prolégomènes à une théorie du langage, Ed. de Minuit (1968) : Paris.

[Katz et Fodor, 1963] Katz J.J., Fodor J.A. (1963). The Structure of a Semantic Theory. In Rosenberg and Travis (1971). *Language.* 34(2). 170-210.

[Kerbrat-Orecchioni, 1988] Kerbrat-Orecchioni C. (1988). Sémantique. In Encyclopedia Universalis, 693-699.

[Laird et al., 1987] Laird J.E., Newell A., Rosenbloom P.S. (1987). SOAR : An Architecture for General Intelligence,. In *Artificial Intelligence.* American Elsevier Pub. Company : New York. Vol. 33(1), 289-325.

[Lehuen, 1997] Lehuen J. (1997). Un modèle de dialogue dynamique et générique intégrant l'acquisition de la compétence linguistique. Thèse de doctorat d'Informatique de l'Université de Caen.

[Lenat, 1982] Lenat D.B. (1982). AM, An Artificial Intelligence Approach to Discovery in Mathematics as Heuristic Search. *Knowledge-Based Systems in Artificial Intelligence* Mc Graw-Hill : New York. 1982.

[Lenat, 1983] D Lenat. (1983). Eurisko: A program which learns new heuristics and domain concepts. In *Artificial Intelligence*. American Elsevier Pub. Company : New York. Vol. 21.

[Lévy, 1990] Lévy P. (1990). Les Technologies de l'Intelligence. Editions La Découverte : Paris.

[Maes et Nardi, 1987] Maes P., Nardi N. (1987). Meta-level architectures and reflection. North-Holland : Amsterdam.

[Mel'cuk, 1986] Mel'cuk I. (1986). Dictionnaire explicatif et combinatoire du français contemporain. Presses de l'Université de Montréal.

[Minsky, 1975] Minsky M. (1975). A Framework for representing knowledge, In P. Winston (ed.) *The Psychology of Computer Vision*. Mc Graw Hill : New York. 211-277.

[Nicolle et Saint Dizier, 1998] Nicolle A., Saint-Dizier de Almeida V. (1998). Vers un modèle des interactions langagières. In Delisle S., Chaïb-Draa B., Moulin B. (ed.) *Analyse et simulation de conversations : de la théorie des actes du discours aux systèmes multiagents*. InterÉditions : Lyon.

[Nicolle et Vivier, 1997] Nicolle A., Vivier J. (1997). Dialogue et apprentissage : humain/humain, humain/machine, machine/machine. In K. Zreik (ed.) *Apprentissage par l'interaction*. Europia productions : Paris. 61-82.

[Nyckees, 1998] Nyckees V. (1998). La Sémantique. Belin : Paris.

[Obiwan, 2000] Ontology Based Informing Web Agent Navigation (2000), projet du National Science Foundation CARREER *"Cooperative Agents for Conceptual Search and Browsing of World Wide Web resources",* www.ittc.ukans.edu/obiwan/index.html.

[Osgood, 1963] Osgood C.E. (1963). On understanding and creating sentences. *American psychologist*. Vol. 18, 735-751.

[Peirce, 1978] Peirce C.S. (1978). Ecrits sur le signe rassemblés, traduits et commentés par Gérard Deledalle. Seuil : Paris.

[Perlerin, 2000] Perlerin V. (2000). Catégorisation lexicale pour la recherche documentaire. Rapport de Dea d'Informatique de l'Université de Caen.

[Pitrat, 1990] Pitrat J. (1990). Métaconnaissance. Hermés : Paris.

[Pottier, 1964] Pottier B. (1964). Vers une sémantique moderne. *Travaux de linguistique et de littérature*. Vol. 2/1, 107-137.

[Pretschner et Gauch, 1999] Pretschner A., Gauch S. (1999). Ontology Based Personalized Search. Actes IEEE Intl. Conf. on Tools with Artificial Intelligence 1999. Chicago. 391-398.

[Rastier et al. , 1994] Rastier F., Cavazza M., Abeillé A. (1994). Sémantique pour l'Analyse. Masson : Paris.

[Rastier, 1987] Rastier F. (1987). Sémantique interprétative. PUF : Paris.

[Rialle, 1995] Rialle V. (1995). Vers la maîtrise informatique de la connaissance. In Dr H. Joly (ed.) *Biomédecine*. Lavoisier : Paris. 52-75.

[Rosenfield, 1989] Rosenfield I. (1989). L'invention de la mémoire. Editions Eshel : Paris.

[Salton, 1966] Salton G. (1966). Information Dissemination and Automatic Information Systems. actes IEEEE 54, 12.

[Salton, 1986] Salton G. (1986). Another Look at Automatic Text-Retrieval Systems. Communications of the ACM. Vol. 29, n° 1, 648-656.

[Sabah, 1996] Sabah G. (1996). Le Sens dans les T.A.L. - Le Point sur le Sens après 40 ans de recherches. http://m17/ limsi.fr/ Individu/ gs/ textes/ ATALA-4.14.96/LePointSurLeSens2_ToC.html

[Saussure, 1915] (de) Saussure F. (1915). Cours de linguistique générale. (1986) Mauro-Payot : Paris.

[Sowa, 1984] Sowa J.F. (1984). Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley : Reading, Massachusetts.

[Thlivitis, 1998] Thlivitis T. (1998). .Sémantique interprétative Intertextuelle : assistance informatique anthropocentrée à la compréhension des textes. Thèse d'Informatique de l'Université de Rennes 1.

[Tiberghien, 1997] Tiberghien G. (1997). La mémoire oubliée. Mardaga : Paris.

[Yang et al., 1998] Yang M.C., Wood W.H., Cutkosky M.R. (1998). Data Mining for Thesaurus Generation in Informal Design Informal Design Information Retrieval. Standford University : Standford.

[Zipf, 1949] Zipf G.K. (1949). Human Behavior and the Principle of Least Effort. Addison-Wesley : London.

Les Auteurs

Anne Nicolle est professeur d'informatique à l'Université de Caen et chercheur au GREYC, UMR CNRS 6072. Après une thèse en intelligence artificielle et des travaux sur la réflexivité et les langages à objets, elle s'est orientée vers les sciences cognitives. Dans le pôle Modescos de la MRSH de Caen, elle mène des travaux interdisciplinaires sur l'interaction langagière entre les humains et les machines et sur la modélisation du langage dans une démarche expérimentale.

Pierre Beust est maître de conférences en informatique à l'université de Caen. Ses recherches en traitement automatique des langues se déroulent au sein du GREYC et du pôle Modescos de la MRSH de Caen. En 1998, il a soutenu à Caen une thèse interdisciplinaire entre l'informatique et la linguistique sur l'interprétation et la sémantique des langues dans les modèles informatiques et plus spécifiquement dans le dialogue homme-machine.

Vincent Perlerin est doctorant moniteur en informatique à l'université de Caen . Au GREYC, il consacre la majeure partie de ses recherches à l'expérimentation du modèle ANADIA dans le champ d'application de la recherche documentaire. Son travail de thèse est co-dirigé par Anne Nicolle et Laurent Gosselin.