# La recherche d'information assistée par ordinateur: quelle représentation des connaissances?

*Article inédit. Mise en ligne : 10 février 2004.*

**Maria Caterina Manes-Gallo et Céline Paganelli**

*Maria Caterina Manes Gallo est maître de conférences en psychologie à l'UFR de Psychologie de l'université de Nantes et habilitée à diriger des recherches (HDR) en Sciences de l'information et de la communication. Elle est membre permanente de l'équipe Gresec-Cristal, à l'Institut de la communication et des médias de l'université Stendhal Grenoble 3.*
*Son activité de recherche concerne la représentation des connaissances pour des systèmes d'interface, à fort ancrage linguistique, appliqués à la recherche d'informations textuelles (e. g. manuels techniques).*
*Elle a travaillé surtout sur la modélisation sémantique des opérations prédicatives portées par les verbes, par rapport à la progression du discours.*

*Céline Paganelli est maître de conférences en Sciences de l'information et de la Communication à l'université Pierre Mendès-France Grenoble 2. Elle est membre permanente de l'équipe Gresec-Cristal, à l'Institut de la communication et des médias de l'université Stendhal Grenoble 3.*
*Son activité de recherche concerne les interfaces de recherche d'informations textuelles. Elle travaille notamment autour de deux axes : l'étude des caractéristiques des documents techniques en vue de la représentation des connaissances qu'ils véhiculent ; une réflexion sur une interface de recherche d'information prenant en compte les caractéristiques des utilisateurs visés.*

**Plan**

Introduction
Aspects de la médiation informatique: le Traitement Automatique de la Langue
Représentation des connaissances: entre intelligence artificielle et psychologie cognitive
    Qu'est ce que l'objet «connaissance»?
    Reproduction *vs* imitation du comportement humain
    Un niveau intermédiaire pour *avoir l'air de comprendre*
La recherche d'information : délimitation et problématique
    Définitions
    Vers la conception d'un système de recherche d'information
Conclusion
*Références bibliographiques*

## INTRODUCTION

L'objet du présent article s'inscrit dans une problématique de communication humain/machine. Notamment la conception d'interfaces en Langue Naturelle Écrite (désormais LNE) appliquées à la recherche d'information documentaire. L'aspect qui sera approfondi concerne la *Représentation des connaissances* (désormais RC). Notre objectif est de montrer quelles sont les contraintes que le mode LNE pose à la médiation informatique dans ce type d'application.

Le terme de « recherche d'information » (*information retrieval*) a été utilisé pour la première fois dans les années cinquante par Kelvin N. Moers. D'après les résultats de ces travaux, la difficulté d'accès à l'information induit fréquemment le renoncement ou la suspension de l'activité de *s'informer*. L'usager d'un centre de documentation préfère se priver de l'information dont il a besoin, en interrompant son activité de recherche lorsque celle-ci devient pénible ou difficile à poursuivre. Actuellement, l'importance croissante et le rôle incontournable joué par le dispositif informatique en recherche d'information fait de ce dernier le partenaire principal de l'*agent cognitif* voulant *s'informer*.

Le but de l'utilisateur est d'accéder à une information qui va lui permettre de mettre à jour ses connaissances ou de réaliser une *tâche*. Les modes d'expression des objectifs de l'utilisateur peuvent être multiples (graphique, parole, gestuel, écrit). L'intérêt du mode langue naturelle est de lui permettre de formuler ses requêtes et de poursuivre l'interaction avec l'interface selon une forme qui lui est plus familière que celle d'un langage de commandes. L'utilisateur ne devra pas préalablement intégrer une connaissance sur les fonctionnalités du logiciel d'interrogation ou sur le type d'information que peut fournir le système. Ce qui a l'avantage de diminuer le coût cognitif sous-jacent à l'activité de *s'informer*.

Mais si pour l'humain le mode langue naturelle écrite constitue un vecteur de communication familier, il en va différemment pour l'interface. Pour l'interface, l'information véhiculée par ce mode correspond à des chaînes de caractères dépourvues de sens. D'où la nécessité de fournir au système des données de connaissance (linguistiques) qui lui permettent d'analyser le sens des messages produits par son partenaire humain. Notamment, pour reconnaître *ce dont parle* l'utilisateur et *comment il en parle* (Rouault, Manes Gallo, 2003). On retrouve ici le problème de la *représentation des connaissances* pour un système de Traitement Automatique de la Langue Naturelle Écrite (désormais Talne).

## ASPECTS DE LA MÉDIATION INFORMATIQUE: LE TRAITEMENT AUTOMATIQUE DE LA LANGUE

Dans le cadre de la communication humain/machine en LNE, le traitement automatique de ce mode d'interaction passe par la conception de dispositifs capables à la fois de « comprendre » ce que lui communique l'utilisateur et de « produire » une séquence qui soit pertinente par rapport au message de ce dernier. Ces deux phases du traitement, traditionnellement définies comme l'analyse et la génération, sont contraintes par des critères de conception différents. Ils posent toutefois un problème commun qui est celui de la définition, de la modélisation et de la formalisation des connaissances linguistiques nécessaires et suffisantes au système pour « calculer du sens » à la fois en analyse et en génération.

En analyse, le système doit à la fois identifier *ce qui est dit* et interpréter *pourquoi il est dit*, à partir des formes linguistiques occurrentes dans le message de l'utilisateur. Par exemple, il doit reconnaître si le but sous-jacent à un message renvoie à une requête d'information pour obtenir un certain état de choses : « *Peut-on être dispensé des valeurs de langue si on a un enseignement équivalent dans l'école ?* » Ou bien si la requête concerne la mise en œuvre d'une procédure : « *Est-il possible de faire une dispense d'assiduité de TD ?* » (Chanet, 1996).

Tandis qu'en génération, le système doit à la fois déterminer *quoi dire* en réponse, c'est à dire le contenu conceptuel de son message et choisir *comment le dire*. C'est à dire trouver une modalité linguistique correcte afin d'exprimer ce contenu. La RC, d'un côté guide le traitement pour l'analyse du sens profond des messages de l'utilisateur et de l'autre détermine la pertinence du contenu conceptuel de la réponse que le système va réaliser linguistiquement.

Historiquement le travail de recherche poursuivi au sein de l'équipe Cristal a été centré sur la représentation des connaissances véhiculées par la langue naturelle, aux différents niveaux de son fonctionnement dans le discours (morphologie, syntaxe, sémantique et pragmatique), mais indépendamment du domaine d'application du discours.

Dans le présent document nous focalisons l'attention surtout sur la fonction de la RC au cours de la phase d'analyse et dans le cadre d'une application spécifique qui est la *recherche d'information* (RI) à partir de gros documents. Pour un approfondissement sur la fonction

de la RC en génération nous renvoyons aux travaux d'Antoniadis (1995), Balicco (1993), Ponton (1996) et Balicco (2000).

Le second élément que nous ne pourrons approfondir concerne l'analyse du discours, telle qu'elle a été inaugurée par Pécheux (1969). Ici le problème est celui d'avoir une RC à partir de laquelle dégager de façon automatisée les unités discursives qui constituent les invariants d'un corpus de discours produits dans des conditions homogènes, assurant une certaine répétition dans le vocabulaire. Le système 3AD95, développé dans le travail de Timimi (1999), est un système d'analyse automatique du discours fondé sur la notion de paraphrase. Ce système met en œuvre une analyse qui sauvegarde la composition séquentielle de la phrase et il fournit un ordonnancement des réponses dans les termes de leur pertinence par rapport à la requête de l'utilisateur. Cette capacité du système 3AD95 constitue une méthode alternative dans la RI, qui permet d'extraire et de classer a posteriori, à partir de leur degré de parenté, les segments pertinents par rapport à la requête de l'utilisateur.

## REPRÉSENTATION DES CONNAISSANCES ENTRE INTELLIGENCE ARTIFICIELLE ET PSYCHOLOGIE COGNITIVE

L'intelligence artificielle et la psychologie cognitive partagent la même problématique de RC, mais en l'approfondissant selon des finalités différentes. En ce sens elles s'inscrivent dans le vaste archipel des sciences cognitives. Dans ce qui suit nous ne prendrons pas en considération les raisons historiques qui ont favorisé ce partage et qui dépendent de l'objet d'étude que ces deux disciplines se sont donné au cours de leur évolution (Dupuy, 1994). On essayera plutôt, d'un côté d'expliquer pourquoi la RC est un secteur de recherche privilégié sur la cognition, et de l'autre d'esquisser quelles sont les similitudes et les différences d'approche à l'étude de la RC, en Intelligence Artificielle (IA) et en psychologie. Les secteurs sur lesquels nous focaliserons notre attention sont le Traitement Automatique de la Langue Naturelle Ecrite (Talne) et la psychologie des processus cognitifs. Né d'une rupture épistémologique avec le behaviorisme, ce secteur de la psychologie générale fonde l'explication du comportement humain sur la modélisation de l'activité mentale qui contribue à son émergence. Dans ce qui suit, nous nous limiterons au paradigme théorique dit de la manipulation symbolique, visant à analyser les facultés mentales dites de haut niveau (e. g. résolution de problème, raisonnement, activités de compréhension et/ou de production langagière). Nous ferons donc intentionnellement abstraction du paradigme alternatif représenté par le courant connexionniste, plus focalisé sur l'étude de l'activité mentale sub-symbolique (Tiberghien, 1992).

### *Qu'est ce que l'objet «connaissance»?*

L'objet d'étude commun aux deux disciplines qui nous préoccupent concerne *la connaissance ou la cognition*, non pas comme un contenu définissable, ou un *savoir* spécifique, mais comme un thème de recherche fédérateur. Thème fédérateur qu'elles partagent avec un ensemble d'autres disciplines (neurologie, linguistique, philosophie...) qui analysent ce même objet à partir de perspectives très différentes. Notamment, en référence à ses origines (le système cérébral et le système mental, mais aussi la culture et la collectivité sociale dans son ensemble), à ses supports (les ordinateurs) et au vecteur principal qui en permet la circulation au sein de la communauté sociale (les langues naturelles) (Ganascia, 1996).

En ce sens, la transversalité de l'objet d'étude *cognition/connaissance* impulse et favorise l'émergence de collaborations pluridisciplinaires entre des secteurs de recherche, parfois très éloignés, mais dont l'objectif commun est de donner un fondement scientifique à la fois formalisé et empirique à l'objet *cognition* ou *connaissance*, à partir de différentes

perspectives et en faisant appel aux sciences de l'information (Le Ny, 1989 ; Sabbah, 1989 ; 1991). En particulier l'*Intelligence Artificielle*.

En tant que métaphore computationnelle, l'hypothèse fondamentale de l'IA est que le processus de compréhension est automatisable et peut de ce fait être imité et/ou reproduit sur ordinateur. L'intelligence humaine est considérée comme le produit d'un ensemble de lois, complexe mais fini, et chaque opération du système nerveux ou mental est identifiable par une séquence d'opérations élémentaires. L'IA essaye de traiter sur ordinateur les problèmes qui sont résolus par l'homme de façon sémantique, sans que celui-ci utilise un ou plusieurs algorithmes définis. D'où le rôle central accordé à la RC. L'objet *connaissance* a une fonction cruciale : primo, pour comprendre le système naturel qui la produit (neurologie, psychologie des processus cognitifs), secundo, pour construire des systèmes artificiels qui puissent contribuer à la diffusion et à l'échange de ce même objet *connaissance* (Talne). Il n'est donc pas surprenant que la psychologie et l'IA aient comme premier élément en commun l'utilisation de l'ordinateur soit comme moyen de validation de modèles théoriques (e. g. reproduction sur ordinateur des processus mentaux : vision, compréhension, raisonnement...) soit comme véhicule de *connaissance*s (e. g. construction d'interfaces en langue naturelle pour la communication homme/machine, ou construction de systèmes experts).

### *Reproduction vs Imitation du comportement humain*

L'IA analyse l'objet *cognition* par rapport à la modélisation et à la formalisation des caractéristiques de cet objet dans différents domaines (communication H/M, diagnostic médical, robotique...). En revanche, la psychologie des processus cognitifs étudie l'objet *cognition* par rapport aux facultés mentales humaines qui produisent et utilisent ce même objet. Dans le premier cas on vise une RC qui permette de modéliser la *cognition* nécessaire à un système automatique pour imiter le comportement humain. Dans le deuxième cas, la RC visée doit permettre de définir des modèles, implémentables sur ordinateur, de l'activité mentale, sous-jacente à la mise en œuvre du comportement humain. Cette centralité accordée à la RC dans l'imitation/reproduction des activités humaines dites de *haut niveau* (résolution de problème, raisonnement, activités de compréhension et/ou de production langagière) prend donc des contours différents selon que le système final visé est artificiel ou humain. L'opposition entre imitation et reproduction de la *cognition* a été définie dans les années 80 dans les termes de deux thèses dites de l'*IA forte* et de l'*IA faible* (Pylyshyn, 1984).

L'*IA forte* vise à élaborer des programmes qui, outre la reproduction d'une certaine conduite humaine, permettent aussi une validation de la modélisation des processus cognitifs, sous-jacents à la mise en œuvre de l'activité simulée (Johnson-Laird, 1983). Les études en RC sont finalisées à la modélisation des données de *connaissances* nécessaires au système humain pour *se représenter*, au niveau conceptuel l'information en provenance de l'environnement, indépendamment des caractéristiques individuelles du sujet. La traduction du modèle en un algorithme doit permettre au système artificiel de fournir en sortie :

— le résultat de la tâche pour laquelle il a été programmé (e. g. résumer un fragment de texte présenté en entrée) ;

— la description des étapes du traitement sous-jacent à l'exécution de cette tâche (e. g. reconnaissance des faits principaux relatés et détermination de leurs relations causales) ;

— la description des erreurs d'interprétation plus typiques que peut faire un humain.

Le but visé n'est pas l'efficacité du système, mais plutôt sa pertinence psychologique. La retombée est épistémologique et à plus court terme. La centralité accordée à la RC fait

partie du paradigme théorique et de la méthodologie au sein desquels étudier les processus mentaux, par définition non-observables. Le programme qui correspond à l'implantation d'un modèle théorique sur les processus cognitifs d'un sujet virtuel constitue donc aussi une méthode de vérification « reconstructive », complémentaire à la traditionnelle méthodologie expérimentale, utilisée en psychologie cognitive (Gallo, 1992).

L'*IA faible* , en revanche, vise à élaborer des programmes qui, pour une tâche donnée, fournissent en sortie la même et/ou une meilleure prestation qu'un sujet humain. Par exemple, reconnaître que dans l'énoncé *le secrétaire transporte des livres* le sujet grammatical se réfère à un être animé et pas à un meuble. Les *connaissance*s représentées dans le système contribuent de façon fondamentale à la mise en œuvre de la tâche pour laquelle le système a été programmé. Le travail théorique de modélisation des *connaissances* est finalisé à long terme à une avancée technologique, i. e. concevoir un système plus efficace. Le problème est de *représenter* les connaissances nécessaires et suffisantes pour que le système effectue une tâche « intelligente » plus rapidement et en évitant les erreurs caractéristiques de « l'intelligence naturelle » de son partenaire humain.

En récapitulant, la psychologie des processus cognitifs vise en partie à modéliser les différentes *connaissances* impliquées et/ou sous-jacentes à la mise en œuvre de l'*activité mentale*, qui permet l'attribution d'une signification à l'information en entrée, et donc à l'émergence du comportement. Tandis que, en Talne où le but est de construire des *systèmes intelligents,* le problème est de définir et de modéliser les *connaissance*s nécessaires et suffisantes au système pour attribuer une signification à l'information en entrée.

### *« Se représenter » la connaissance pour comprendre*

L'aspect de l'objet *cognition* privilégié par les psychologues et les neurologues concerne, d'un côté, l'analyse des rapports esprit/cerveau, et de l'autre l'étude des modalités de fonctionnement de l'esprit i. e. le système cognitif qui produit et utilise l'objet cognition au niveau symbolique. Le siège de l'activité mentale comme processus d'attribution de sens à l'information en entrée au système est la mémoire. La mémoire est un système articulé en plusieurs sous-systèmes : la Mémoire à Long Terme (MLT), la Mémoire à Court Terme (MCT), et la Mémoire de Travail (MT).

De façon très synthétique la MLT sert à stocker les données de connaissances acquises par le sujet au cours de son existence. Par exemple, les connaissances relatives aux savoirs et aux savoir-faire. Tandis que la MCT et la MT permettent au sujet, à partir de l'activation de ses connaissances en MLT, de se construire des représentations circonstancielles des situations avec lesquelles il doit interagir. Par exemple, formuler des anticipations sur le développement de l'intrigue du film qu'il est en train de regarder. Ainsi le terme RC, au sein de la psychologie, sert à indiquer deux phénomènes différents. Il s'agit d'un côté des connaissances, considérées comme des structures stabilisées et stockées en MLT - i. e. savoirs de base acquis et/ou modifiés au cours de l'existence - et de l'autre des représentations considérées comme des états provisoires de connaissance en MCT, résultants de l'activité d'attribution de signification à l'information en entrée. L'acquisition de nouvelles *connaissances* en MLT se fait par stabilisation des représentations plus fréquentes en MCT.

Cette relation entre RC et activité mentale permet de différencier les caractéristiques spécifiques de la *cognition* humaine, par rapport à la fonction de la RC en Talne. Par exemple, pour ce qui concerne le *traitement des significations* qui dépendent de la gestion des *connaissances* et des représentations, en psychologie on distingue deux grandes catégories de processus : les processus automatiques et les processus contrôlés. Les *processus automatiques* impliquent la mise en œuvre irrépressible et non-consciente des don-

nées de *connaissance*, tandis que les *processus contrôlés* impliquent une mise en œuvre consciente des données de *connaissance* soit par échec des automatismes, soit par nouveauté des problèmes à affronter. Par exemple, la compréhension d'un énoncé ironique est considérée comme le résultat d'un processus contrôlé mis en œuvre par le sujet après l'échec du calcul automatique de sa signification littérale (Munch, Brouillet, 1997).

L'existence de deux types de processus (automatiques *vs* contrôlés) de gestion des *connaissances* implique aussi la possibilité de dériver de la nouvelle information en générant des types d'inférences fonctionnellement différentes. Par exemple, dans la compréhension de récits on distingue les *inférences élaboratives* et les *inférences nécessaires*. Les *inférences élaboratives* permettent de dériver le sens implicite véhiculé par la surface linguistique, à partir de l'activation d'une *connaissance* générale sur le monde, e. g. repérage des relations causales ou instrumentales véhiculées par le contenu de deux énoncés. Leur production systématique entraînerait une explosion inférentielle, sans la formulation d'*inférences nécessaires*. Ces dernières permettent au sujet de rétablir la cohérence de sa représentation en MCT, par rapport au contenu explicite du texte. En ce sens elles permettent de contrôler la production des i*nférences élaboratives* plus pertinentes aux différents moments du récit (Coirier, Gaonac'h, Passerault, 1996).

### *« Représenter » la connaissance pour imiter la compréhension*

D'un point de vue historique, on peut faire remonter la naissance de la problématique de la représentation des connaissances RC en IA à un article de Mc Carthy et Hayes (1969) qui distingue deux aspects conceptuellement distincts - heuristique et épistémologique - de la conception de programmes de résolution de problèmes. L'*aspect heuristique* correspond à la définition d'algorithmes efficaces pour une recherche automatique de la solution, tandis que l'*aspect épistémologique* concerne la modalité à travers laquelle on représente les problèmes que le système doit résoudre. Les difficultés soulevées par la formulation de programmes efficaces sont proportionnelles à la complexité de l'aspect épistémologique, complexité qui rejaillit sur l'aspect heuristique. Par exemple, la RC nécessaire pour un programme qui joue aux échecs est beaucoup moins complexe que la représentation des *connaissance*s (linguistiques et extra-linguistiques) nécessaires pour la reconnaissance du sens implicite d'une phrase et/ou d'un texte. Le programme pour un jeu d'échecs en effet, présuppose la définition d'un ensemble d'états correspondants aux possibles configurations de l'échiquier et un ensemble d'opérateurs simples permettant de passer d'un état à un autre.

Dans la communication humain/machine en LNE, indépendamment du type d'application de l'interface (interrogation de base de données, recherche d'informations, dialogue...), un des principaux problèmes qui se pose est celui de définir ce que signifie pour une machine comprendre et/ou produire un texte ou un discours en LNE. Les traitements en analyse concernent le passage de la séquence de surface au sens implicite véhiculé, aux différents niveaux (morphologique, syntaxique, sémantique, pragmatique), par les formes marquées par les pré-traitements (Rouault, 1987).

Le traitement en analyse est articulé en plusieurs sous-traitements, définis à partir des niveaux de fonctionnement du système de LN, tels qu'ils ont été formulés par les linguistes. L'efficacité d'un système dépend de sa capacité de résolution des ambiguïtés à chaque niveau et de la reconstruction du sens implicite. D'après l'option méthodologique adoptée au sein de l'équipe Cristal, cette efficacité est fondée sur la pertinence des modèles linguistiques formalisés plutôt que sur des recettes informatiques. Avec toutefois un bémol. En effet, les logiques contraignantes de la programmation et du fonctionnement interne de

la machine interdisent une adoption a-critique des modélisations trop descriptives formulées par les linguistes. Pour chaque phénomène linguistique considéré, il est toujours nécessaire de déterminer le niveau de granularité auquel l'analyser – notamment dans le cas des opérations de prédication, leur signification littérale impliquant la prise en compte du sens véhiculé par les verbes selon le nombre et le rang des actants qui l'encadrent. Par exemple, en sémantique on vise une modélisation des connaissances qui prenne en compte la signification littérale des indicateurs linguistiques en étude, en faisant volontairement abstraction de leur possible signification figurée (Manes Gallo, Rouault, 1999).

Comme son nom l'indique la problématique de la RC implique toujours deux aspects. À savoir, une modélisation des connaissances linguistiques nécessaires pour le calcul du sens et une formalisation des modèles définis en vue de leur implantation. Ce deuxième aspect correspond à la nécessité de mettre en place une stratégie de résolution qui permette au système d'exploiter de façon optimale, en analyse et en génération, les connaissances définies. Par exemple, en analyse le problème essentiel est celui de réduire le plus possible la production d'ambiguïtés ou de solutions parasites aux différents niveaux d'analyse (morphologique et/ou syntaxique). En d'autres termes, le problème de la représentation au sens strict implique le choix de l'architecture du système. L'option choisie au sein de l'équipe Cristal est celle développée par les travaux de Stéfanini (1992) et Warren (1996). Il s'agit de l'architecture multi-agents Talisman qui permet une meilleure interaction entre les différentes connaissances linguistiques que doit gérer le système, par rapport à la traditionnelle architecture séquentielle.

### *Un niveau intermédiaire pour avoir l'air de comprendre*

La principale différence entre la perspective informatique et la perspective psychologique concerne d'un côté l'ontologie - le système humain sait tandis que le système automatique ne saura jamais sans un humain - et de l'autre la finalité de l'étude sur la RC (cf. supra, « Reproduction *vs* Imitation du comportement humain »). Un système traite de l'information en manipulant des séquences de caractères qui pour lui n'ont aucun sens. Pour les systèmes automatiques, l'*attribution d'une signification* à l'information en entrée dépend des données de *connaissance* qui ont été représentées dans le programme par le concepteur du système. Tandis que dans le cas du sujet humain, il y a acquisition de nouvelles *connaissances* en MLT par stabilisation des représentations qu'il a plus fréquemment construites en MCT.

La RC est donc à la base de la capacité du dispositif informatique d'inférer et/ou de dériver des données informatives nouvelles qui, dans le cas d'une interface en LNE, coïncident avec la reconnaissance et la reconstruction du sens implicite des messages de l'utilisateur. Sens implicite qui renvoie d'un côté à l'interprétation de la signification véhiculée par les marques linguistiques de surface, ou repérage de ce qui est dit, et de l'autre au but communicationnel qu'elles permettent de réaliser, ou repérage des intentions de l'utilisateur (cf. infra, « Vers la conception d'un système de recherche d'information »).

Cette focalisation sur l'inscription sémiotique des intentions de l'utilisateur rend l'opposition entre *IA forte* et *IA faible* beaucoup moins contrastée ou profonde dans le cas des systèmes de Talne appliqués à la recherche d'information. En d'autres termes, la construction d'interface pour la communication humain/machine permet d'infléchir l'opposition entre imitation et reproduction du comportement humain. D'un côté la *convivialité* de ces systèmes (*IA faible*) repose en partie sur des investigations empiriques relatives aux dimensions cognitives et/ou aux stratégies langagières mises en œuvre par les sujets pour véhiculer et construire à travers le discours la/les *connaissance(s)* qu'ils ont

l'intention de communiquer. Mais d'un autre côté, cette caractéristique de *convivialité* du système final, ne contraint pas à viser une modélisation des processus cognitifs du sujet utilisateur, au niveau de profondeur requis par la reproduction de l'activité mentale, sous-jacente à la mise en œuvre de sa compétence langagière (*IA forte*). Le système final doit *avoir l'air* d'interagir avec l'utilisateur, ce qui n'implique pas de reproduire ses processus mentaux. Si tel était le cas, le système final serait un clone incomplet d'un sujet humain virtuel. En ce sens, les problèmes posés par l'interaction humain/machine en LNE permettent de repenser la RC en référence à un niveau intermédiaire entre reproduction et imitation de la compétence humaine dans l'activité de compréhension et de production langagière.

Pour ce qui concerne la modélisation des connaissances linguistiques nécessaires au système, l'adoption d'une approche cognitiviste se rapporte à la volonté d'étudier la langue comme support et véhicule de *connaissances*. Par exemple, à partir de l'analyse des stratégies énonciatives et des opérations prédicatives mises en œuvre par le sujet, repérer les objets de discours et leurs relations, construites par la progression du discours (Rouault, Manes Gallo, 2003). Ce type de linguistique se différencie d'autres approches plus traditionnelles qui sont focalisées sur la description comparative de différents systèmes de langue (étude synchronique) ou sur leur évolution dans le temps (étude diachronique).

Le Talne s'inscrit dans cette perspective cognitiviste, à cause de l'objectif qu'il se donne, i. e. faire effectuer par l'ordinateur sur des textes ce que l'humain lui-même est capable d'effectuer (comprendre, raisonner, résumer...). D'où une plus grande importance accordée à l'analyse des stratégies langagières mises en œuvre dans des discours, qu'aux langues elles-mêmes. Cette focalisation sur la « langue en fonctionnement », telle qu'elle peut être repérée à partir de l'étude de corpus, permet de donner un fondement empirique à la modélisation théorique (Rouault, 1987). En effet en Talne le problème est surtout celui d'identifier les problèmes réels posés par la mise en œuvre du système de signes à des fins communicationnelles. Par exemple, au niveau sémantique, on distinguera entre les *connaissances* linguistiques communes à tous les corpus et les *connaissances* spécifiques déterminées par le domaine du discours. Dans ce cadre, la prise en compte du type d'application visée est essentielle. En recherche documentaire, le syntagme nominal joue un rôle plus central que dans les systèmes de dialogue H/M (consultation interactive d'une base de données) dans lesquels le verbe joue un rôle plus prégnant (Paganelli, 1997).

## LA RECHERCHE D'INFORMATION: DÉLIMITATION ET PROBLÉMATIQUE

Dans le cadre de la recherche d'information, la problématique que pose la représentation des connaissances consiste à doter le système de RI de connaissances lui permettant d'être intermédiaire entre l'utilisateur et le texte, sachant que ce rôle d'intermédiaire se joue entre un utilisateur qui a du mal à exprimer ses besoins et un texte dont le système doit pouvoir interpréter le contenu. En d'autres termes, la RC pour un système de RI doit d'un côté lui permettre d'inférer les intentions de l'utilisateur et de l'autre lui expliciter le contenu du texte.

### *Définitions*

La recherche d'information est une activité dont la finalité est de mettre en regard une information et un utilisateur. C'est une activité par laquelle un utilisateur accède à une information (un document, une partie de document, des données) consignée sur un support, information qui va lui permettre de mettre à jour ses connaissances ou de réaliser une tâche.

Ici, nous nous intéressons aux utilisateurs en situation professionnelle qui recherchent de l'information en vue de réaliser une tâche. Cela signifie que l'activité de recherche d'information est motivée par un besoin initial qui s'inscrit dans un contexte social déterminé, où l'information est recherchée pour une utilisation précise (Chevallet, 2002).

### *L'utilisateur*

L'utilisateur s'engage dans une recherche d'information parce qu'il a un manque à combler. En effet, nous supposons que la demande d'information d'un utilisateur représente un manque d'information. Cependant, le fait que l'utilisateur s'engage dans une activité de RI signifie déjà qu'il est conscient de ce manque et qu'il cherche à y remédier. Quand Dervin (1983 ; 1992) propose, pour étudier les besoins d'information, une méthode qui emploie la métaphore « *situation-gap-use* » selon laquelle a) tous les besoins d'information viennent d'une lacune dans les connaissances d'un individu ; b) cette lacune entraîne une situation spécifique à laquelle l'individu peut remédier par différentes tactiques, l'auteur sous-entend déjà que l'utilisateur a conscience de cette lacune.

De la même manière, Belkin (1984) juge que le besoin d'information émerge d'une anomalie dans l'état des connaissances d'un utilisateur. Il souligne la difficulté pour l'utilisateur d'exprimer ses besoins, qui représentent ce qui lui manque ou ce qu'il ne connaît pas. Taylor (1986), lui aussi, considère que le point de départ du développement d'un besoin d'information est un malaise inexprimable par l'utilisateur. Le processus qu'il décrit, de la lutte de l'utilisateur pour exprimer son besoin à la recherche d'information pour le satisfaire, est cognitif. Les travaux de Belkin et de Taylor insistent sur la difficulté des utilisateurs à verbaliser leur manque et leur besoin d'information.

### *L'information*

En regard de l'utilisateur, se trouve l'information. Elle peut être véhiculée par du texte, du son, des images... L'activité de RI, qui consiste en une mise en correspondance entre une demande et une information, nécessite que, au préalable, il y ait une organisation de la documentation et des outils de représentation de l'information.

Classiquement, les traitements effectués pour représenter les documents recouvrent deux types de description : la description physique et la description du contenu du document. La recherche d'information à laquelle s'intéressent la plupart des travaux de l'équipe Cristal est une RI textuelle. Elle porte sur des documents intégraux (et non des références de documents) qui sont majoritairement constitués de texte. Dernièrement, un nouveau chantier de recherche a été démarré (Badjo-Monnet, 2000), qui s'intéresse à l'indexation de documents multimédias et vise à prendre en compte le média image.

### *Les logiques de conception*

L'idée d'un malaise de l'utilisateur au départ du processus de RI fait que nous sommes enclins à fonder notre démarche de conception d'un système sur une méthode privilégiant l'utilisateur aux fonctions ou, autrement dit, privilégiant la logique de l'ergonome à celle du concepteur. Cette distinction est issue notamment de Moran (1981) qui oppose la logique du concepteur à la logique de l'ergonome.

Dans le premier cas, le concepteur se base sur ses intuitions pour prédire ce que veut l'utilisateur. Cette logique privilégie le système, et l'interface est optimisée par rapport aux tâches à exécuter. Il en résulte une multiplication d'interfaces en fonction des tâches : interface de traitement de texte, de messagerie, de recherche documentaire (Dalbin, 1992). Selon Bisseret (1983), les concepteurs sont ici orientés par les problèmes et non par les utilisateurs visés. Les limites de cette logique sont évidentes puisque les intuitions du

concepteur ne peuvent suffire à recouvrir tous les besoins des utilisateurs ni à les cerner de manière forcément adéquate.

La logique de l'ergonome privilégie l'utilisateur. Elle considère qu'une réflexion sur l'interface doit se situer en amont de la conception d'un système et doit recouvrir un travail sérieux et profond d'étude des utilisateurs. Selon Bisseret (1983), l'objectif de l'ergonomie est d'adapter les comportements de l'ordinateur, c'est à dire toutes ses manifestations externes, à l'usager dans le sens d'une assistance la plus évidente et la moins contraignante possible de son activité. Ainsi, on peut envisager qu'une même application ait des interfaces différentes selon les utilisateurs auxquels elle s'adresse car il est invraisemblable qu'un mode d'interaction soit satisfaisant pour tous (Shneidermann, 1978). Dans ce cas, la conception d'un système ne peut se faire qu'après un recueil d'informations sur ses utilisateurs potentiels.

Dans le cadre ici défini de la recherche d'information textuelle par des utilisateurs dans une situation professionnelle, la conception d'un système nécessite que soient pris en compte à la fois les caractéristiques des utilisateurs visés et les spécificités du contexte de la tâche de recherche d'information (Paganelli, 2003).

## *Vers la conception d'un système de recherche d'information*

### *L'utilisateur dans le processus de recherche d'information*

Le schéma présenté ci-dessous représente le processus de RI et insiste sur la place de l'utilisateur au sein de ce processus, de son besoin d'information à son interprétation de la réponse du système comme l'information attendue.

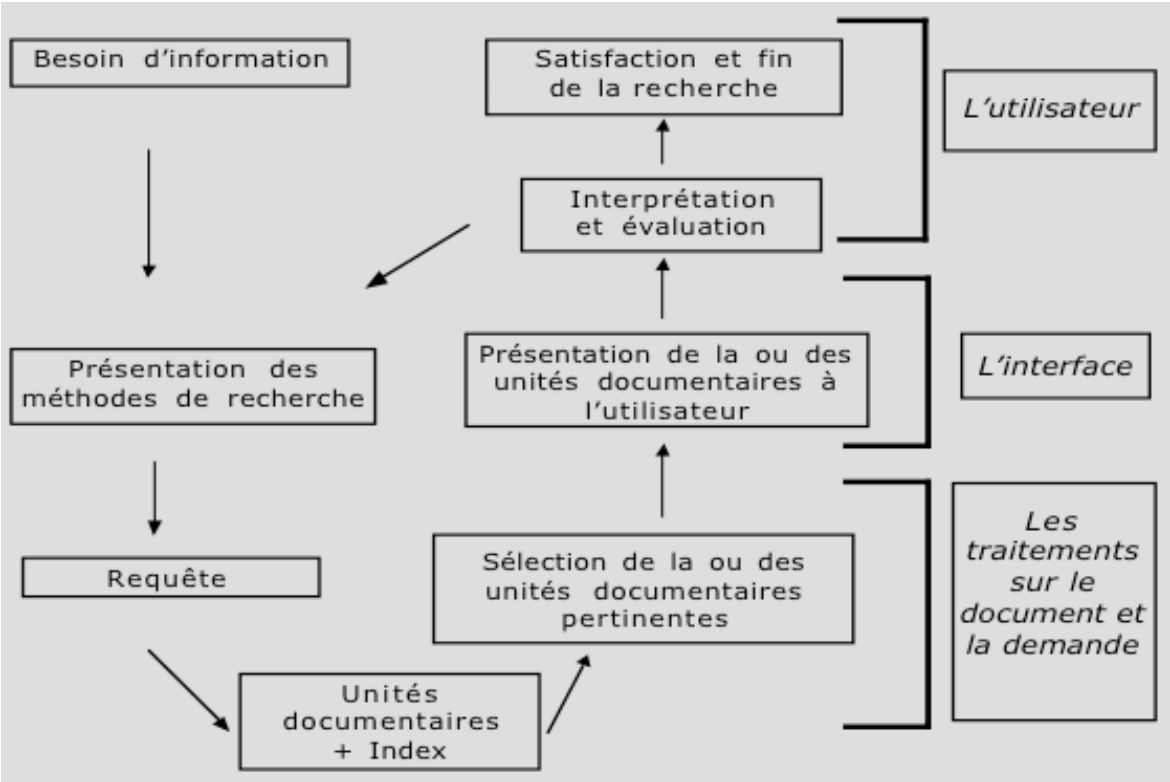


**Schéma du processus de RI**

Le schéma présenté ci-dessus permet de représenter grossièrement le processus de recherche d'information. Ce schéma comporte trois parties principales : l'utilisateur et le module de traitements (sur le document et sur la demande d'information) qui sont reliés entre eux par l'interface. En effet, pour être « en phase », ces deux parties nécessitent la

présence d'un intermédiaire qui est l'interface. Celle-ci a pour but d'assurer la communication entre l'utilisateur et le système et de guider l'utilisateur dans son travail.

Le module des traitements segmente le document en unités qui sont utilisées pour servir de base à l'indexation. Puis il effectue une représentation de ces unités documentaires par l'opération d'indexation. L'utilisateur a un besoin d'information dans le cadre de son activité. L'interface permet à l'utilisateur de traduire son besoin d'information dans le langage du système et de le transformer ainsi en requête. Celle-ci constitue une représentation du besoin d'information qui se fait au travers de l'utilisation du vocabulaire d'indexation. L'interface présente notamment à l'utilisateur les différents moyens qu'il peut utiliser pour exprimer son besoin d'information : listes issues de l'indexation, opérateurs de combinaison, de proximité... et les différentes stratégies de recherche qui peuvent être mises en œuvre.

Par rapport à cette requête, le module de traitements sélectionne, dans le document, les unités documentaires dont la représentation est conforme à celle de la requête. C'est l'interface qui prend alors en charge la manière dont ces segments de texte sont présentés et mis à disposition de l'utilisateur. C'est aussi d'elle que vont dépendre les éventuelles possibilités d'exploiter ces segments réponses, avec la possibilité de relancer une recherche à partir d'une réponse ou avec l'utilisation éventuelle de liens de type hypertextuels.

L'utilisateur interprète alors les unités documentaires qu'il reçoit en réponse. C'est l'utilisateur qui, en consultant ces morceaux de document, leur donne du sens en fonction de ses besoins et de l'information attendue. Suite à une recherche, soit l'utilisateur est satisfait et il retourne à son activité principale, soit il tente une nouvelle recherche.

### *La représentation du texte*

On considère ici que l'indexation et la RI textuelle dans un document posent des problématiques nouvelles, notamment par rapport à l'indexation et la recherche documentaire classique. Si la seconde porte sur des bases de références de documents, la première s'effectue dans des documents en texte intégral.

Pour la recherche d'information textuelle dans un document :

– la première question est de savoir pour quel type de besoin un utilisateur effectue une recherche d'information dans un gros document. Sommes-nous en présence d'un besoin thématique comme dans la recherche documentaire où les besoins sont du genre : « Quels sont les documents qui parlent de tel sujet ? » ? Ou bien l'utilisateur adresse-t-il un autre type de demande à ce document ?

– La seconde question concerne l'unité documentaire à choisir pour servir de base à l'indexation dans un gros document, et pour être utilisé comme réponse à une requête. Pour la recherche documentaire, les unités documentaires à indexer sont données par avance puisque c'est le document dans son entier qui est indexé, alors que pour la recherche d'information textuelle, on ne s'intéresse plus à une collection de documents mais à un seul document en texte plein. Ainsi, en réponse à l'utilisateur, on ne fournira plus les références d'un document mais une partie de texte appartenant à ce document. De ce fait, il faut segmenter le document en parties de texte qui soient indexées et discriminantes les unes par rapport aux autres, et qui seront ensuite fournies à l'utilisateur.

– Enfin, la dernière question concerne l'indexation des unités documentaires. Dans le cadre de la recherche documentaire, le vocabulaire d'indexation est constitué de formes nominales qui représentent le ou les thèmes d'un document. Pour la RI en texte intégral, on se place dans le cadre d'une indexation automatique. Mais on peut se demander s'il y a des contraintes particulières qui pèsent sur l'index. En effet, est ce que la représentation

d'un document nécessite l'indexation de toutes ses chaînes de caractères ? Est-ce que l'on doit conserver uniquement les formes nominales, ou bien utiliser aussi d'autres types de formes (notamment les formes verbales) ? En fait, il faut s'interroger pour savoir quelles sont les méthodes d'indexation les plus appropriées à ce type de recherche d'information et quels traitements doivent être mis en œuvre.

Ainsi, nous considérons que les points à déterminer pour la conception d'un système de RI dans un document intégral sont les suivants :

*1. Les méthodes de recherche* : quels besoins émanent des utilisateurs ? quelles méthodes leur proposer pour effectuer la RI et exprimer leur besoin ? Dans le cadre de la RI dans des documents techniques par des utilisateurs experts du domaine (Paganelli, 1997), des entretiens ont été effectués auprès d'utilisateurs potentiels du système à concevoir, pour savoir quels types de demandes les amènent à effectuer une recherche d'information dans un document technique. Il est apparu que les experts recherchent dans des documents techniques pour remplir un besoin de type opératif : ils veulent savoir pour faire. Ce besoin se traduit par deux types de demandes d'information adressées au document : les premières concernent des demandes de définition ou description d'objets, les secondes sont des demandes de description de procédures. Suite à ce recueil d'information, nous supposons que les méthodes classiques de RI (formes nominales et opérateurs booléens) ne sont pas adaptées pour exprimer ce type de demandes.

L'observation du processus de RI par des utilisateurs experts nous a amenés à différents résultats (Paganelli, 1997 ; Mounier, 1999). D'une part, il apparaît que les utilisateurs experts veulent aller vite dans leur recherche d'information et ne veulent pas consulter des listes de résultats trop importantes ni des réponses trop longues. D'autre part, il ressort de cette observation que les experts ont du mal à exprimer leurs demandes avec des mots-clés et de ce fait reformulent plusieurs fois leurs requêtes avant satisfaction.

Ces différents résultats expérimentaux nous amènent à réfléchir à des méthodes de RI plus adaptées aux utilisateurs visés par le système, et à une présentation des réponses du système qui pondère ces réponses notamment en fonction de leur taille.

*2. La segmentation du document* : comment découper le document en unités documentaires qui soient discriminantes les unes par rapport aux autres et qui soient autonomes i. e. qui constituent des unités réponses pour l'utilisateur ?

Différents choix peuvent être faits pour segmenter le document en unités d'information indexables et susceptibles de constituer des réponses pertinentes. Les travaux de Starfill et Waltz (1992) cités dans Hearst *et alii*, (1993) proposent un découpage du texte en segments de taille égale, à savoir trente « mots ». Chaque segment est indexé de manière autonome et une requête d'un utilisateur est alors comparée à chacun des segments. Cette proposition d'un découpage arbitraire du texte prend seulement en compte des critères de taille et laisse de côté les aspects linguistiques ou structurels du texte.

L'étude de Salton *et alii* (1993) considère que le texte est un assemblage de morceaux de texte de taille variable qui vont des phrases aux unités hiérarchiques à savoir les unités issues de la structure logique du document. Ce travail propose de garder la phrase comme unité pertinente et de découper le texte en phrases. Ainsi, chaque phrase se voit attribuer un poids statistique par rapport au poids des mots qu'elle contient. La segmentation en phrases a des inconvénients car dans les documents volumineux, la phrase est une unité qui semble trop petite pour être significative. De plus, elle n'a pas une véritable autonomie lui permettant d'être comprise hors de son contexte. Ainsi, si les phrases sont syntaxiquement autonomes, elles comportent des éléments implicites comme les anaphores (Paice *et alii*, 1993).

La plupart des découpages tiennent compte de l'organisation logique du document. Comme on l'a dit plus haut, le document a une structure logique et est organisé en entités. Ces unités logiques ajoutent un niveau de signification supplémentaire et une information n'aurait pas la même valeur selon l'unité à laquelle elle appartient (Guthrie, 1988 ; Caro, 1993).

Une étude (Mounier, 1998) a consisté à travailler sur l'unité documentaire à indexer en déterminant au préalable l'unité documentaire qui est attendue comme réponse par les utilisateurs. Suite à une expérimentation, il apparaît que les unités qui sont choisies comme des réponses satisfaisantes par les utilisateurs sont des unités plus fines que les unités référencées dans la structure du document (dans la table des matières) et sont comparables au paragraphe. Ce type de résultat implique donc que la segmentation du document à indexer se fasse au niveau des paragraphes. D'autres études expérimentales (Ouerfelli, 2000) ont été menées sur ce sujet.

*3. Les méthodes d'indexation :* comment extraire et représenter les connaissances véhiculées par le document ?

Ce dernier point soulève le problème de l'indexation automatique de documents textuels et du choix des termes à conserver pour représenter le contenu des documents. Il s'agit de repérer les descripteurs qui représenteront chaque unité issue de la segmentation. Le terme de « descripteur » est pris ici non pas dans le sens strict documentaire, c'est à dire comme élément appartenant à un langage documentaire, mais comme un syntagme nominal extrait automatiquement à partir du texte. Le descripteur est envisagé ici comme un élément du discours (Amar, 2000). Il peut s'agir soit d'une forme nominale simple soit d'une forme nominale complexe.

Dans sa thèse, Bentes-Pinto (1999) s'est intéressée à l'indexation de documents techniques, en particulier les manuels de maintenance. Son travail montre que ces documents ont des contenus spécifiques qui ne peuvent être représentés par une indexation classique (liste de formes nominales). En alternative, elle propose de représenter le contenu de ces documents dans des schémas de tâche dont l'ancrage linguistique sont les constructions prédicatives. Ce schéma vise à rendre compte de l'information procédurale contenue dans ces manuels à partir du type de procès (actions *vs* états) référent des constructions prédicatives occurrentes dans le document. Il permet d'identifier notamment les descriptions d'opérations à réaliser pour sortir d'une panne et les pré-requis nécessaire à leur réalisation.

## CONCLUSION

Dans le présent article, nous avons essayé d'expliquer quelle est la fonction de la RC dans la construction de systèmes de Talne et pourquoi elle s'inscrit à un niveau intermédiaire entre reproduction et imitation du comportement langagier humain.

Nous avons focalisé l'attention sur la phase d'analyse (i. e. interprétation de *ce qui est dit* et de *pourquoi il est dit*) dans le cadre de la RI. Par rapport à ce type d'activité la *convivialité* de l'interface dépend de son rôle d'intermédiaire entre le texte et l'utilisateur. D'après l'approche adoptée nous considérons que le niveau de convivialité du système dépend d'une modélisation des connaissances fondée à la fois sur des études d'ergonomie cognitive et de linguistique. D'après la première perspective, on part de données empiriques sur le comportement langagier de l'utilisateur afin de pouvoir adapter le système à ses attentes. Tandis que d'après la seconde perspective, on essaie de modéliser les connaissances linguistiques nécessaires et suffisantes au système pour interpréter la

langue en fonctionnement dans un texte, le plus indépendamment possible du domaine de référence de ce dernier.

## RÉFÉRENCES BIBLIOGRAPHIQUES

Amar Muriel, *Les fondements théoriques de l'indexation : une approche linguistique*, ADBS éditions, 2000

Antoniadis Georges, *Génération automatique de textes : essai de caractérisation du domaine*, Séminaire franco-suisse d'Archamps Linguistique et Informatique, 6 avril 1995

Badjo-Monnet Bernadette, Bertier Marc, « Indexation pour la recherche d'informations dans des documents techniques structurés multimédias », *Les enjeux de la et de l'information et de la communication*, 2000

Balicco Laurence, *Génération des répliques en français dans une interface homme-machine en langue naturelle*, thèse de doctorat en Informatique, université Pierre Mendès-France, Grenoble 2, février 1993

Balicco Laurence, Ben-Ali Salahedine, Ponton Claude, Pouchot Stéphanie, « Two applications for a non context French generator », *Natural Language Processing 2000*, June 2000

Belkin Nicholas J., « Cognitive models and information transfer« , *Social science information studies*, n° 4, 1984

Bentes Pinto Virginia, *La représentation des connaissances dans le contexte de la documentation technique : proposition d'un modèle d'indexation*, thèse de doctorat en Sciences de l'information et de la communication, université Stendhal, Grenoble, 1999

Bisseret André, « Pour une psychologie ergonomique des systèmes documentaires », *Documentaliste*, volume 20, n° 1, janvier-février 1983

Chanet Catherine, *La demande dans le dialogue finalisé : de la surface linguistique aux représentations de l'action*, thèse de doctorat en Sciences de l'information et de la communication, université Stendhal, Grenoble, septembre 1996

Chevallet Jean-Pierre, Nigay Laurence, « Les interfaces pour la recherche d'information » in *Interaction homme-machine et recherche d'information* (Paganelli Céline, dir.), Hermès, Lavoisier, 2002

Coirier Alain, Gaonac'h Daniel, Passerault Jean-Michel, *Psycholinguistique textuelle : une approche cognitive de la compréhension et de la production des textes*, Armand Colin, Paris, 1996

Dalbin Sylvie, « Interfaces dans les systèmes d'aujourd'hui », *Les interfaces intelligentes dans l'IST*, Rapport Inria, 1992

Dervin Brenda, « Information as a user construction : the relevance of perceived information needs to synthesis and interpretation », *Knowledge structure and use : implications for synthesis and interpretation*, Philadelphia, Temple university press, 1983

Dervin Brenda, « From the mind's eye of the user : the sense making qualitative-quantitative methodology », *Qualitative research in information management*, Englewood, Libraries Unlimited, 1992

Dupuy Jean-Pierre, *Aux origines des sciences cognitives*, La Découverte, Paris, 1994

Gallo Maria Caterina, « Levels of integration of unexpected elements in story comprehension », *Ricerche di Psicologia*, n° 1, 1992

Ganascia Jean-Gabriel, *Les sciences cognitives*, Flammarion, Paris, 1996

Guthrie John T., « Locating information in documents : examination of a cognitive model », *Reading Research Quartely*, 1988

Hearst Marti A., Plaunt Christian, « Subtopic structuring for full-length document access », *Proceedings of the 16th annual international ACM SIGIR, conference on research and development in information retrieval*, Pittsburg USA, 1993

Johnson-Laird Philip N., *Mental models*, Harward University Press, Cambridge MA, 1983

Le Ny Jean-François, *Science cognitive et compréhension du langage*, PUF, Paris, 1989

Manes Gallo Maria Caterina, Rouault Jacques, « Le couple Sémantique/Pragmatique et le calcul du sens », *Vextal*, 200-207, 1999

Manes Gallo Maria Caterina, « Communication humain/machine en langue naturelle : un nouvel enjeu pour la psycholinguistique », revue *Interaction Homme/Machine*, vol. 2, n° 4, 2003

Mc Carthy John, Hayes Pat J., « Some philosophical problems from the stand point of Artificial Intelligence », in Meltzer Bernard et Michie David (Eds) : *Machine Intelligence*, n° 4, Edimburgh University Press, 1969

Moran Thomas P., « An applied psychology of the user », *Computing surveys*, vol. 13, n° 1, 1981

Mounier Evelyne, Paganelli Céline, « Texts' structures and information retrieval in large textual documents », *Fifth International ISKO Conference*, Lille, France, 1998

Mounier Evelyne, Paganelli Céline, « L'accès à l'information pertinente dans les documents techniques volumineux », *Secondes Journées du Chapitre Français de l'ISKO*, Lyon, octobre 1999

Munch Christelle, Brouillet Denis, « Les processus de compréhension des expressions ironiques », *Fifth International Congress of the International Society of Applied Psycholinguistics*, Porto, Portugal, June 1997

Ouerfelli Tarek, *Recherche d'information dans un document technique : restructuration du texte dans une perspective de consultation sur indices textuels*, thèse de doctorat en sciences de l'information et de la communication, université Stendhal, Grenoble, 2000

Paganelli Céline, *La recherche d'information dans des bases de documents techniques en texte intégral. Étude de l'activité des utilisateurs*, thèse de doctorat en Sciences de l'information et de la communication, université Stendhal, Grenoble, 1997

Paganelli Céline, Mounier Evelyne, « Information retrieval in Technical documents : from the User's Query to the Information-Unit Tagging », *Proceedings of ACM Sigdoc*, San Francisco, octobre 2003

Paice Chris D., Jones Paul A., « The identification of important concepts in highly structured technical papers », *Proceedings of the 16th annual international ACM SIGIR, conference on research and development in information retrieval*, Pittsburg USA, Juin 1993

Pêcheux Michel, *Analyse automatique du discours*, Dunod, Paris, 1969

Ponton Claude, *Génération automatique de textes en langue naturelle : Essai de définition d'un système noyau*, thèse de doctorat en Informatique, université Stendhal, Grenoble, novembre 1996

Pylyshyn Zenon, *Computation and Cognition*, Cambridge MA, Harvard University Press, 1984

Richard Jean-François, Tiberghien Guy, « Épistémologie et Psychologie », *Psychologie française*, t. 44, n° 3, 1999

Rouault Jacques, *Linguistique automatique : applications documentaires*, P. Lang, Berne, 1987

Rouault Jacques, Manes Gallo Maria Caterina, « Intelligence Linguistique : Le calcul du sens des énoncés élémentaires », *Hermès*, 2003

Sabah Georges, « L'intelligence artificielle et le langage », « Représentation des connaissances », vol. 1, *Hermès*, 1989

Sabah Georges, « Le traitement automatique des langues », in Vergnaud G., (Ed.), *Les sciences cognitives en débat*, Éditions du CNRS, 1991

Salton Gerard, Allan James, Buckley Chris, « Approaches to passage retrieval in full text information systems », *Proceedings of the 16th annual international ACM SIGIR, conference on research and development in information retrieval*, Pittsburg USA, Juin 1993

Schneidermann Ben, « Improving the human factors aspect of database interaction », *ACM Transactions*, vol. 13 n° 4, décembre 1978

Stéfanini Marie-Hélène, *Talisman : une architecture multi-agents pour l'analyse du français écrit*, thèse de doctorat en Informatique, université Pierre Mendès-France Grenoble 2, janvier 1993

Taylor Robert S., *Value-added processes in information systems*, Norwwod, Ablex, 1986

Tiberghien Guy, « Les Sciences Cognitives : Un nouveau programme scientifique ? », in Sfez Lucien (Ed.), *Dictionnaire critique de la communication*, PUF, Paris, 1992

Timimi Ismail, *De la paraphrase linguistique à la recherche d'information, le système 3AD : théorie et implantation (aide à l'analyse automatique du discours)*, thèse de doctorat en Sciences de l'information et de la communication, université Stendhal Grenoble 3, septembre 1999

Vignaux Georges, *Les Sciences Cognitives : une introduction*, La Découverte, Paris, 1991

Warren Karine, *Gestion de conflits dans une architecture multi-agents d'analyse automatique*, thèse de doctorat en Informatique, université Stendhal, Grenoble, 1998