Quelques remarques à propos des systèmes de recherche d'information et de traitement des connaissances

Article inédit, mis en ligne le 24 mars 2003

Adrian Staii

Adrian Staii est actuellement attaché temporaire d'enseignement et de recherche à l'IUT 2 de l'université Pierre Mendès France (Grenoble 2), et prépare une thèse de doctorat au sein du Gresec sur la recherche d'information et le traitement des connaissances. Journaliste de formation, il s'intéresse également à la socio-économie des médias et est l'auteur de plusieurs traductions.

Plan

Introduction
Problèmes posés par les systèmes classiques de recherche d'information et de traitement des connaissances
L'utilisation du TAL pour la recherche d'information et le traitement des connaissances
Propositions pour un système de recherche d'information et de traitement des connaissances
Perspectives
Références bibliographiques

INTRODUCTION

On parle souvent d'une société de l'information pour désigner improprement une réalité où l'informatisation et l'automatisation gagnent continuellement en ampleur et en importance (Bangemann, 1994 ; Castells, 1998). Si l'appellation est certes contestable et contestée (Tremblay, 1995 ; Miège, 1997 ; Lacroix et al., 1998), elle renvoie néanmoins à un certain nombre de faits dont l'évidence semble faire l'unanimité : durant les cinquante dernières années, la quantité et la diversité de l'information produite n'ont cessé de s'accroître, de même que les moyens nécessaires pour la traiter, la stocker ou l'échanger.

Pendant longtemps, le moteur de la croissance de ce double processus a été surtout une accumulation quantitative, la masse d'information qu'on pouvait traiter et même les solutions envisageables en matière de traitement étant conditionnées par les capacités physiques des machines (mémoire, vitesse de calcul, débit des réseaux, etc.). Ces dernières années, avec l'avancée de la microélectronique et sous l'impact des divers projets d'informatisation sociale, la donne semble différente. Nous en sommes arrivés à un moment où les progrès en matière de traitements informationnels reposent moins sur l'augmentation des capacités physiques des machines que sur la qualité des solutions proposées.

Dans ce texte, il est question d'un domaine où de telles solutions qualitatives sont aujourd'hui nécessaires : la recherche d'information. Nous voulons montrer qu'une voie possible pour le développement de ce domaine peut être le recours à une architecture complexe, où la recherche d'information repose sur et se combine avec le traitement des connaissances. Cette symbiose semble possible par la mise à contribution des outils appropriés pour le traitement automatique des langues (TAL) et pour la représentation des connaissances.

La structure de ce texte suit la logique de ce questionnement. La première section sera destinée à une présentation des principaux problèmes soulevés par les systèmes classiques de recherche d'information et par les systèmes à base de connaissances. À travers

l'énumération de ces problèmes, nous verrons que des solutions communes pourraient venir de la même direction : le traitement automatique de la langue. Aussi, la deuxième section portera-t-elle sur les difficultés majeures de l'utilisation du TAL pour la recherche d'information et le traitement des connaissances. La troisième section apportera quelques suggestions concernant la conception d'un système qui combine la recherche d'information avec le traitement des connaissances. Enfin, nous terminerons en nous posant quelques questions qui esquisseront autant de perspectives de recherche.

Comme ce texte ne s'adresse pas forcément à un public expert, nous avons été parfois obligé de sacrifier la complétude des propos au bénéfice de la lisibilité, cherchant un juste compromis entre l'exactitude et l'accessibilité.

PROBLÈMES POSÉS PAR LES SYSTÈMES CLASSIQUES DE RECHERCHE D'INFORMATION ET DE TRAITEMENT DES CONNAISSANCES

Un système de recherche d'information est un dispositif censé permettre à l'utilisateur de trouver de manière pertinente et efficace les données dont il a besoin (pour une présentation générale des problématiques liées à la recherche d'information, voir Bertrand, 1991 ; Paganelli, 1997). On peut ranger dans cette catégorie les systèmes de recherche documentaire utilisés dans les bibliothèques, les systèmes d'interrogation directe question-réponse (pour la consultation des horaires des trains, par exemple), les systèmes de recherche d'information des entreprises, les annuaires électroniques, etc. L'architecture de ces systèmes s'organise autour de deux composantes centrales : la base de données et le moteur de recherche. L'appariement entre la requête de l'utilisateur et les informations enregistrées dans la base se fait d'habitude à l'aide d'un index plus ou moins complet de mots-clés. Par exemple, dans nombre de systèmes documentaires, la sélection est opérée à partir d'une description signalétique des documents et d'une représentation condensée de leur contenu (indexation sous forme de mots-clés ou descripteurs et, éventuellement, résumé du document). Ainsi, l'interrogation de la base de données permet-elle d'obtenir uniquement des documents secondaires (descriptions), les utilisateurs devant par la suite se procurer les documents primaires qui leur sont signalés.

En opposition avec ces systèmes, et pour répondre aux besoins de plus en plus divers des utilisateurs, émergent des applications où la recherche s'effectue dans des bases de données en texte intégral, c'est-à-dire qui ne se limitent plus à de simples descriptions secondaires, mais contiennent les textes tout entiers. Si ces applications sont encore difficiles à implémenter dans certains domaines, comme les bibliothèques par exemple (à cause de la quantité considérable de textes qui devraient être enregistrés dans la base, mais également à cause des problèmes juridiques liés aux droits d'auteur), elles sont de plus en plus fréquentes notamment dans les entreprises qui doivent faire face à une masse croissante et diverse de documents techniques (Bertrand, 1991).

La tâche du système est alors double : il doit d'abord trouver les documents pertinents, et, par la suite, faire une recherche dans ces documents pour trouver l'information précise qui lui a été demandée. Cette tâche pose plusieurs problèmes fondamentaux, dont notamment :

— le traitement de la requête et l'indexation automatique. Traditionnellement, on prenait en considération des unités lexicales isolées, mais ce principe a vite montré ses limites. En effet, le traitement d'un mot isolé peut (au meilleur des cas) rendre compte de sa signification hors contexte, or une requête renvoie d'habitude à un signification contextuelle (l'utilisateur a besoin d'une information particulière, dans un contexte spécifique, celui de la tâche pour la résolution de laquelle cette information lui fait défaut).

Aussi, pour être »informatifs », les descripteurs doivent-ils représenter des syntagmes et non des unités lexicales isolées. Il s'ensuit donc qu'il est nécessaire de doter le système avec des techniques de TAL capables de reconnaître et d'analyser ces syntagmes (Lallich-Boidin 1986 ; Eymard, 1992). Cependant, le problème n'est pas ainsi entièrement résolu ; des études (Paganelli, 1997) ont montré que, souvent, les requêtes ne portent pas seulement sur des concepts et ne sont pas exprimées uniquement par des syntagmes nominaux. Elles concernent également des modes opératoires formulés verbalement (surtout dans le contexte des applications techniques). Par conséquent, il est nécessaire d'élargir le modèle de traitement linguistique de sorte qu'il puisse prendre également en compte les syntagmes verbaux. Ainsi, l'analyse doit-elle porter sur un énoncé élémentaire (syntagme verbal + syntagmes nominaux). Comme cette articulation est la »brique de base » de tout discours, on peut dès lors faire le pari que le système peut être conçu de manière à servir aussi bien à des fins de recherche d'information que d'extraction automatique de connaissances.

— la définition des unités pertinentes pour la réponse : quelle serait l'unité textuelle que le système devrait retourner à l'utilisateur suite à une requête ? le syntagme ? la phrase ? le paragraphe ? le chapitre ? le texte tout entier ? Le problème est ici de trouver un juste milieu entre la pertinence de la réponse (le fait qu'elle doit contenir toute l'information que l'utilisateur exige) et son efficacité (il est clair que si le système retourne à l'utilisateur un chapitre, voire le texte tout entier, il y a plus de chances que l'information soit complète, mais, en même temps, la réponse n'est point optimale en termes d'efficacité, l'utilisateur étant obligé de dépenser beaucoup de temps pour sélectionner lui-même ce qui l'intéresse exactement). Une bonne solution de compromis entre les deux contraintes serait de choisir le paragraphe comme unité de réponse (Mounier, 1996 ; Ouerfelli, 2001).

— nombre de documents ne contiennent pas uniquement du texte mais aussi des images, graphiques, schémas, etc. L'indexation des unités non-textuelles de ce type pose plusieurs problèmes complexes : quels sont les marqueurs de surface permettant de relier ces unités aux unités textuelles correspondantes ? comment indexer les unités mixtes (par exemple, les schémas qui contiennent également du texte) ? comment présenter la réponse lorsqu'il n'y a pas de lien immédiat entre le paragraphe textuel pertinent et l'unité non-textuelle qui pourrait le compléter? etc.

À travers cette énumération, on peut déjà constater que l'utilisation des outils plus complexes de traitement de la langue devient de plus en plus nécessaire, même dans le cadre des systèmes classiques de recherche d'information. Mais, notre hypothèse va plus loin, en considérant que le développement de tels outils pourrait approcher les systèmes de recherche d'information et les systèmes à base de connaissances.

Un système à base de connaissances est un dispositif complexe censé accompagner l'utilisateur dans la résolution d'un problème. Les systèmes experts, apparus dans les années soixante, comptent parmi les plus répandus aujourd'hui (Parsaye et al., 1988 ; Hayes-Roth et al., 1994 ; Bradley, 1994). Un tel système est composé d'une base de connaissances et d'un moteur d'inférences. L'utilisateur, confronté à un problème spécifique, interroge le système (d'habitude en langue naturelle) sur les solutions possibles. Il lui fournit un certain nombre d'informations ponctuelles à partir desquelles le système peut caractériser le problème concret qu'il doit résoudre. Le système fait ensuite appel aux connaissances fondamentales dont il dispose dans la base afin de réduire cette situation particulière à des cas type. Lorsque les raisonnements automatiques qu'il met en œuvre aboutissent, le système propose les solutions les plus plausibles et, si l'utilisateur le désire, les raisons qui justifient ces choix.

Un système »expert » est donc une sorte »d'assistant intelligent ». »Expert », parce qu'il est spécialisé dès sa conception et possède beaucoup de connaissances sur un domaine

limité ; »assistant », parce qu'il ne prend pas de décision, il propose des solutions à l'utilisateur qui s'en sert, donc, sauf cas exceptionnel, il n'est pas autonome ; »intelligent » parce qu'il est capable de proposer des solutions adaptées à des problèmes spécifiques, à partir d'une masse de connaissances et de raisonnements généraux.

Malgré l'engouement scientifique et commercial qu'ils ont suscité dès les années soixante-dix (pour des statistiques d'utilisation, voir Hayes-Roth et al., 1994), la plupart des systèmes experts sont relativement encombrants, principalement pour deux raisons :

— leur spécialisation poussée. En effet, l'efficacité des systèmes experts dépend étroitement de la limitation du domaine d'application (même s'il existe aujourd'hui des plate-formes générales qui peuvent être adaptées à plusieurs domaines).

— la quantité considérable de travail humain nécessaire pour la mise à jour des bases de connaissances ou pour l'adaptation de la plate-forme générale à un nouveau domaine d'application. Ce travail ne se limite pas à des tâches opérationnelles et routinières qui pourraient être confiées à de simples exécutants, mais implique également l'identification des connaissances fondamentales, donc la participation d'experts humains.

Si les résultats de ces systèmes peuvent être intéressants, leur coût est donc considérable : aux coûts initiaux de l'achat d'une plate-forme générale, s'ajoutent les coûts d'adaptation à un domaine d'application spécifique et les coûts de mise à jour. Cela pour un système dont l'autonomie est limitée et qui, s'il est sans doute utile pour conforter l'expert humain dans ses prises de décision, ne saurait le remplacer.

Cependant, ce type de système pourrait avoir un avenir intéressant si des solutions sont trouvées pour réduire certains de ses désavantages, comme son autonomie limitée. Ainsi nous il nous semble que des applications plus larges sont envisageables si on parvient à une construction et à une mise à jour automatiques de la base de connaissances, en réduisant au minimum le travail humain. Certes, cette automatisation permettrait l'élargissement du domaine d'application, et le système ne serait plus en l'occurrence un système »expert ». En revanche, il trouverait ainsi une application tout aussi profitable à la recherche d'information.

Comme la clé de cette reconversion nous semble le développement d'un outil approprié de traitement de la langue, nous allons y consacrer la section suivante.

L'UTILISATION DU TAL POUR LA RECHERCHE D'INFORMATION ET LE TRAITEMENT DES CONNAISSANCES

Les premières tentatives en matière de traitement automatique des langues remontent aux années cinquante. L'enthousiasme suscité par les premiers automates, capables de résoudre des problèmes mathématiques ou logiques de manière relativement efficace, s'est alors emparé d'autres champs d'application, notamment linguistiques (pour un panorama de l'Intelligence Artificielle, voir Barr et al., 1981-1989 ; pour une introduction, voir Haugeland, 1987). Les premières tentatives visaient la traduction automatique, qui apparaissait comme un enjeu à la fois important dans le contexte de la guerre froide (les premiers systèmes étant bilingues, anglais-russe/russe-anglais) et plus facile à atteindre. Cette impression de facilité s'appuyait certes sur une conception assez simpliste de la langue (qui, il faut le dire, n'était pas à l'époque propre uniquement aux mathématiciens et logiciens mais également à bon nombre de linguistes) et qui explique en partie les résultats mitigés de ces premières tentatives : le système typique était doté d'un dictionnaire bilingue et se limitait à substituer les mots d'une phrase en anglais par les mots correspondants en russe (et vice-versa), opération suivie d'un éventuel ré-ordonnancement grammatical, afin d'éviter des structures syntaxiques trop lourdes. Vue la technique utilisée, on comprend aujourd'hui pourquoi les résultats ont pu décevoir. Mais, au-delà de leurs effets immédiats

sur les recherches en traduction automatique (concrétisés souvent par une réduction des crédits, voir le rapport Alpac, 1966), ces échecs ont eu le mérite de contribuer à une prise de conscience générale concernant la complexité des langues naturelles et la difficulté de les modéliser (Bar Hillel, 1964 ; Dreyfus, 1984).

Cette prise de conscience et la nécessité de fonder les traitements automatiques sur une compréhension plus profonde de la langue ont été renforcées à chaque nouvelle tentative de création de systèmes à base de TAL. L'élaboration des grammaires morphologiques et syntaxiques a mis en évidence la nécessité du recours, même partiel, à la sémantique, les tentatives de formalisation du sens ont montré l'interdépendance entre la sémantique et la pragmatique, etc. (pour une présentation thématique et historique, voir Sabah, 1990). Ainsi est-on passé d'une vision partielle et fragmentée de la langue à une vision plus large, mettant en avant l'interdépendance des différents niveaux (morphologie, syntaxe, sémantique, pragmatique) et la nécessité d'étudier des phénomènes linguistiques particuliers non pas dans le cadre strict de chacun de ces niveaux, mais dans une perspective transversale.

Toute tentative de conception d'un système à base de TAL ne saurait négliger aujourd'hui cet acquis historique : l'efficacité même du système dépend du modèle de la langue sous-jacent. Le rôle du modèle est de permettre un découpage dans la masse des connaissances linguistiques, d'identifier et de structurer celles qui sont nécessaires pour le fonctionnement du système. Ces connaissances sur la langue (ou méta-connaissances) doivent être représentées d'une certaine manière pour être accessibles à la machine (dans un certain langage) et organisées de sorte qu'elles soient opérationnelles (par exemple, sous forme de règles, lexiques, heuristiques etc.). Certes, selon les buts précis du système, la vision qu'on veut avoir de la langue peut être plus ou moins globale, le modèle plus ou moins puissant, la nature et la complexité des mécanismes linguistiques retenus plus ou moins larges. Ainsi, pour certaines tâches (comme l'indexation automatique), la prise en compte des aspects morpho-syntaxiques est-elle considérée dans la plupart des cas suffisante, les phénomènes sémantiques n'étant considérés que *a minima*. En revanche, pour d'autres (comme le traitement automatique des connaissances) il est indispensable d'avoir une vision plus complète de la langue : en occurrence, si la morpho-syntaxe est certes nécessaire, elle n'est cependant que le premier pas vers la prise en compte de la sémantique et de la pragmatique, dont la modélisation semble la véritable clé pour avancer.

Deux problèmes complémentaires se posent donc dès le départ : celui du modèle linguistique à utiliser (qu'est-ce qu'il faut retenir de la langue ?), et celui de sa représentation formelle (comment représenter et faire manipuler par la machine les connaissances linguistiques ainsi délimitées ?). Les deux questions sont étroitement liées, car si le modèle linguistique à utiliser est certes choisi pour son adéquation avec les objectifs fixés pour le système, il l'est également (sinon surtout) pour sa compatibilité avec les possibilités de représentation formelle dont on dispose. Le problème qui se pose en fin de compte est celui de parvenir à une vision de la langue qui soit à la fois proche de sa complexité inhérente et opérationnelle pour des traitements automatiques, c'est-à-dire exprimable sous forme de règles et grammaires.

Un compromis difficile à réaliser, d'autant plus difficile lorsqu'il s'agit d'une application au traitement des connaissances et à la recherche d'information, ceci pour les raisons suivantes :

— le traitement des connaissances et la recherche d'information à base de connaissances nécessitent des outils de formalisation du sens. Les limites des tentatives précédentes d'une sémantique générale (Lenat, 1990, 1995 ; Lenat et al., 1995) montrent que cette formalisation doit s'appuyer sur les traitements de surface. La première difficulté est donc de

l'ordre du modèle de la langue utilisé : celui-ci se doit d'identifier de manière »claire et distincte » les mécanismes linguistiques qui permettent un traitement du sens à partir de la surface de la langue. En supposant que ces mécanismes aient été identifiés, apparaît une deuxième difficulté, liée à leur formalisation sous forme de grammaires ;

— fonder les traitements sémantiques sur la surface de la langue implique aussi le développement des outils (grammaires) appropriés de traitement morpho-syntaxique. Les tentatives classiques (Fillmore, 1968) ont montré qu'une correspondance biunivoque entre les grammaires morpho-syntaxiques et les grammaires sémantiques n'était pas la solution la plus appropriée. Se pose donc ici la question d'un ancrage plus souple de la sémantique dans la morphologie et la syntaxe, mais également celle d'un traitement morpho-syntaxique complet de la phrase : la seule analyse des syntagmes nominaux ou verbaux (pris séparément) n'étant susceptible de rendre compte que d'une partie des mécanismes de transmission du sens. Cependant, un traitement complet de la phrase pose un certain nombre de problèmes difficiles à résoudre : l'identification des syntagmes complexes (surtout lorsqu'ils sont éclatés, c'est-à-dire lorsqu'il n'y a pas de proximité immédiate entre les différentes formes qui les composent), l'identification de la place syntaxique occupée par ces syntagmes de manière à pouvoir inférer par la suite leur rôle sémantique, le traitement des structures anaphoriques ou des formes ayant une fonction de fléchage (surtout lorsque le référent linguistique apparaît dans une autre phrase), le traitement des structures imbriquées, la reconstitution des structures élidées ou »écrasées » (comme certaines nominalisations déverbales, par exemple, qui, de point de vue morphologique, sont des noms, mais qui, de point de vue sémantique, renvoient à des actions), etc. ;

— enfin, il apparaît indispensable de disposer d'un modèle de représentation des connaissances parfaitement compatible avec les outils de traitement linguistique, afin de faciliter le stockage des connaissances dans une structure formelle permettant des raisonnements automatiques.

Dans la section suivante, nous indiquerons les lignes générales d'une démarche d'utilisation du TAL pour le traitement des connaissances et la recherche d'information qui répond à quelques-unes de ces difficultés.

PROPOSITIONS POUR UN SYSTÈME DE RECHERCHE D'INFORMATION ET DE TRAITEMENT DES CONNAISSANCES

Le travail que nous menons au sein de l'équipe Cristal du laboratoire Gresec trouve son origine dans quelques constats que nous avons déjà indiqués dans les sections précédentes, mais qu'il conviendrait peut-être de rappeler brièvement :

— l'augmentation des possibilités de stockage électronique des documents (Bertrand, 1991) appelle la création de systèmes avancés en matière de traitement de l'information et des connaissances. En effet, il est aujourd'hui possible de stocker une masse impressionnante de documents de nature diverse, et il semble que les systèmes dont on dispose n'exploitent que partiellement cette capacité technique. Les systèmes d'indexation automatique permettent déjà une réduction du travail humain, mais ils sont loin de l'éliminer. Les systèmes de recherche en texte intégral exigent une participation importante de l'utilisateur qui se traduit par un effort de sélection et une perte de temps considérables (Paganelli, 1997). Les réponses sont souvent peu pertinentes. Ces observations sont également valables pour les systèmes de type »expert », où la présence humaine reste essentielle pour la création et la mise à jour des bases de connaissances.

— nombre de connaissances produites et échangées prennent une forme textuelle (langue naturelle écrite) et, en parallèle, nombre de besoins informatifs concernent des

informations textuelles. Les systèmes de recherche d'information et de traitement des connaissances font appel à des outils de traitement automatique de la langue, mais il nous semble que leur utilisation pourrait être améliorée et élargie.

Ces constats nous ont amené à l'hypothèse qu'il serait possible de concevoir un système complexe de recherche d'information et de traitement des connaissances pour des textes écrits en français. Ce système est censé permettre :

— un traitement automatique des connaissances à partir d'une base de données textuelles. Des outils linguistiques spécifiques permettraient d'extraire les connaissances et de les organiser automatiquement dans une base, ce qui réduirait considérablement le travail humain ;

— une recherche d'information en langue naturelle, effectuée directement sur la base de connaissances. Cela permettrait à l'utilisateur d'avoir des réponses plus précises qu'une recherche classique, puisque des connaissances dispersées dans le texte seraient représentées ainsi dans le même objet. Alors qu'une recherche classique permet au meilleur des cas d'accéder aux diverses séquences textuelles susceptibles de véhiculer ces informations, grâce à l'interrogation directe de la base, l'utilisateur aurait un accès plus rapide à des connaissances synthétisées. L'interrogation en langue naturelle lui permettrait également de formuler librement sa requête, sans passer par des index prédéfinis ;

— une recherche d'information classique (mais améliorée) en texte intégral. Classique, parce que les unités de réponse seraient des morceaux du texte d'origine. Améliorée, parce qu'un module de traitement linguistique permettrait un appariement plus flexible entre la requête et les index et parce que la définition des index ne se ferait plus hors contexte. Cette possibilité doit être envisagée pour le cas où l'utilisateur aurait besoin du texte d'origine et non d'une représentation de ce texte.

Puisque les deux fonctions centrales du système (recherche d'information et extraction des connaissances) sont conditionnées par la performance des outils de TAL employés, il convient de présenter en quelques mots la démarche générale d'analyse linguistique. La préoccupation principale étant d'aboutir à une représentation du sens, se pose immédiatement une double question : d'un côté, il s'agit de définir ce qu'on entend par sens, de l'autre, de trouver les moyens pour l'extraire à partir de la surface de la langue. Le modèle général de la langue sur lequel nous fondons la conception de l'analyseur linguistique (voir Rouault et al., à paraître) offre des solutions intéressantes à ces deux problèmes.

La réponse à la première question s'inspire de la théorie notionnelle d'Antoine Culioli (Culioli, 1995 ; 1999), concernant le processus de matérialisation des connaissances en discours. Le concept clé par lequel Culioli explique le passage de l'extralinguistique au linguistique est celui de notion, sorte de composite de sens qui permet différentes matérialisations de surface, selon la volonté du locuteur, le type de discours, les contraintes contextuelles, etc. La notion est ainsi une sorte de charnière entre l'extralinguistique et le linguistique : extralinguistique par nature, car de l'ordre de l'idée, linguistique parce qu'elle peut s'exprimer, s'expliciter par le discours, bien qu'évidemment la langue ne soit pas son seul moyen d'extériorisation. La notion est ainsi une sorte de point obligé de passage entre ce qui est de l'ordre du *dicible* et ce qui est de l'ordre du *dit*.

Approcher le sens au niveau notionnel revient donc à le concevoir dans sa forme la plus abstraite (ou la moins énoncée). Pour prendre un exemple, la notion »habiter » est un composite non-différencié de significations potentielles, qui se concrétisent dans la langue par des mots comme : »habiter », »habitant », »habitation »... mais également par des mots *frontière* comme »construire », »voisin », etc. (voir Culioli, 1999). Ainsi, à ce niveau

fondamental n'y a-t-il pas de différence entre prédicatif (»habiter ») et non-prédicatif (»habitant », »habitation »), ni, d'autant moins, entre les différentes formes de la prédication (»habiter », »être_habité »), ni, encore moins, entre ses différentes énonciations possibles (»j'habite », »il habitait », etc.). Si la notion représente le niveau le plus abstrait de la signification, l'énoncé se situe au pôle opposé, celui de l'expression la plus concrète d'une partie de cette signification, dans un contexte particulier et par des mécanismes spécifiques.

Le passage de la notion à l'énoncé est donc à la fois un processus de définition d'un sens particulier et un processus de différenciation. À travers l'expression linguistique dans un contexte discursif, la notion est contrainte de prendre des formes de plus en plus différenciées aboutissant à l'explicitation d'un sens particulier. Ainsi, la notion »habiter » peut-elle donner naissance à un objet individuel de discours (»habitant », »habitation ») ou à un prédicat (»habiter »). L'objet individuel peut renvoyer dans l'extralinguistique à un référent extensionnel (»un habitant ») ou à un référent type (»l'habitant »). Le prédicat peut indiquer une propriété ou bien introduire une action (»L'habitant possède une habitation » *vs* »L'habitant démolit son habitation »). Pour résumer, disons que le sens énoncé peut être de nature non-prédicative (»homme », »maison », »pomme », etc.) ou prédicative (»être », »habiter », »manger », etc.). Et que la combinaisons des deux catégories au sein d'un même énoncé peut avoir deux effets sur la progression du discours : statique, correspondant à l'expression d'une propriété (»L'homme est intelligent ») et dynamique, correspondant à l'introduction d'une action (»L'homme construit une maison »).

Le modèle que nous utilisons repose sur l'idée que ce parcours de concrétisations successives peut être reconstitué à partir des traces qu'il laisse à différents niveaux de la langue (morphologie, syntaxe, sémantique, pragmatique). L'extraction du sens peut donc se faire en identifiant et en exploitant les phénomènes suivants :

— le fonctionnement syntaxique d'un verbe dans un contexte concret, c'est-à-dire en fonction d'un schéma syntaxique actualisé en surface (rang, nombre et ordre des compléments régis). La définition de ces schémas de base s'appuie sur la théorie des rangs des compléments élaborée à partir du régime des compléments de Maurice Gross (Gross, 1975). L'hypothèse est qu'à chaque verbe français on peut associer un certain nombre de schémas syntaxiques prédéfinis. Le schéma syntaxique particulier actualisé en surface peut être identifié à travers l'étude des syntagmes (surtout nominaux) susceptibles d'occuper ses places de compléments ;

— le fonctionnement sémantique du verbe en relation avec les actants qui occupent ses places d'argument. Il s'agit ici de catégoriser les verbes français en fonction de leur contribution possible à la progression du discours (Fuchs, 1991 ; Manès-Gallo et al., 1992). Deux grandes classes se détachent : les états et les processus. Un verbe catalogué »état » a une fonction sémantique statique, il sert surtout à introduire des propriétés (»Paul est grand »), alors qu'un verbe catalogué processus a une fonction dynamique, il modifie les objets du discours (»Paul mange une pomme/la pomme est mangée »). L'inconvénient est qu'il y a peu de verbes qui peuvent être classifiés a priori comme »état » ou »processus », la plupart pouvant changer de fonction sémantique selon le contexte. Il convient donc de différencier le type de procès référent du verbe (hors contexte), du type de procès référent de l'énoncé (en contexte). Les verbes seront ainsi catalogués a priori comme étant plutôt »état » ou plutôt »processus » en fonction de leur fonctionnement majoritaire. Quant au type de procès de l'énoncé, celui-ci dépend évidemment du type de procès référent du verbe, mais également du schéma syntaxique particulier où il apparaît (le verbe »manger », par exemple, peut porter un processus s'il est suivi d'un complément d'objet – »L'homme

mange la pomme » – ou un état s'il a un fonctionnement absolu – »L'homme mange pour vivre »). Entrent également en jeu la nature sémantique des constituants nominaux qui saturent les places de complément/argument dans un contexte particulier, les marques énonciatives, etc. ;

— ces informations sont représentées sous forme de schéma sémantique. Un schéma sémantique type comprend les places suivantes : une place de relateur, destinée à accueillir les informations concernant le fonctionnement sémantique du verbe prédicat ; trois places d'argument saturées par les constituants nominaux du verbe prédicat (la première place recueille les informations concernant la notion source de l'énoncé – ou sujet grammatical ; la deuxième correspond à la notion but – pour les verbes transitifs, il s'agit du nom occupant la place de complément d'objet ; la troisième est destinée à la notion bénéficiaire, qui peut correspondre, en syntaxe, au datif) ; une place d'énonciation destinée à recueillir des informations comme le temps, la modalité, etc. ; une place globale où seront enregistrées les valeurs d'ensemble de l'énoncé (le type de procès de l'énoncé, par exemple). Le schéma sémantique recueille ainsi les informations extraites à partir de la surface de la langue et facilite le passage à la représentation des connaissances (Berrendonner et al., 1992).

En effet, la construction des objets de connaissances peut se faire de manière automatique, en exploitant les informations recensées dans le schéma sémantique. Ainsi, les places d'arguments sémantiques sont-elles transformées en objets individuels (ou concepts), alors que la place du relateur engendre un objet prédicatif. Les informations sémantiques concernant les constituants nominaux enrichissent les objets individuels ; les informations sur le verbe prédicat et celles sur l'énoncé global enrichissent les objets prédicatifs. De cette manière, on parvient à une représentation des différents types de connaissances que le discours peut véhiculer : on y retrouve la distinction entre non-prédicatif et prédicatif (à travers la constitution des objets individuels et des objets prédicatifs), et, à l'intérieur des expressions prédicatives, celle entre l'expression d'une propriété (caractéristique statique du discours) et l'expression d'une action (caractéristique dynamique), dichotomie préservée grâce à la typologie des procès.

Cette démarche d'extraction du sens est suivie pour la réalisation des deux fonctions centrales du système : le traitement des connaissances et la recherche d'information. Lors de l'extraction des connaissances, l'analyseur linguistique travaille sur la base de données textuelles et permet de reconstituer les connaissances prédicatives et non-prédicatives véhiculées par ces textes dans une base de connaissances, sous forme d'objets individuels et d'objets prédicatifs. Lors de la recherche d'information, la demande, exprimée en langue naturelle, subit un traitement similaire dont le résultat est la création d'un objet requête. La réponse consiste dès lors dans un appariement entre l'objet requête et les objets créés à partir de la base de données textuelles.

Selon le type d'utilisateur (expert, novice, etc.) et selon ses besoins, deux solutions de présentation de la réponse sont envisageables : une présentation directe de l'objet (ou des objets) de connaissance susceptible(s) de satisfaire la requête et une présentation sous forme d'énoncés en langue naturelle. La première forme de présentation serait plus difficile à interpréter par un utilisateur novice (car elle nécessiterait un minimum de connaissances sur la structuration de la base et sur le formalisme utilisé), mais elle aurait l'avantage d'être plus complète et de réduire le risque d'erreur. La deuxième présuppose la participation d'un générateur linguistique capable de traduire les objets de connaissance en discours ; elle serait certes plus adaptée à un utilisateur novice, mais un expert risquerait de la trouver plus contraignante (en termes de temps de réponse et de complétude). Il est évident que ces deux solutions ne s'excluent pas mutuellement. Comme le principal souci

de ce type de système devrait être le confort cognitif de l'utilisateur et la qualité de la réponse, il nous semble que les deux solutions devraient être envisagées.

PERSPECTIVES

Cette tentative de combinaison de la recherche d'information avec le traitement des connaissances, basée sur des outils de traitement de la langue, a des avantages considérables (comme celui d'offrir une gamme plus large de possibilités de recherche, de permettre en même temps l'accès au texte intégral et à la base de connaissances, de réduire l'intervention des experts humains, de diversifier les possibilités d'expression de la requête en langue naturelle, le confort cognitif et l'interaction avec l'utilisateur, etc.), mais, en même temps, elle rencontre quelques difficultés de taille.

Il y a d'abord toute la série des problèmes posés par l'analyse des langues naturelles. Le modèle général que nous avons présenté propose une approche intéressante à l'extraction du sens et notre travail d'application de ce modèle à une catégorie spécifique de textes écrits (les textes techniques) nous a permis de vérifier son potentiel opérationnel et, en même temps, d'identifier certaines de ses limites. Si nos efforts se sont concentrés pour l'instant sur la résolution des problèmes posés par l'analyse des énoncés simples, il faudra dans l'avenir envisager le traitement des énoncés complexes, ce qui présuppose la résolution des questions comme les liens inter-énoncés, les reprises anaphoriques, les structures imbriquées, etc.

Se pose ensuite le problème de la définition opérationnelle des mécanismes d'inférence mais également de la réalisation des raisonnements automatiques plus proches de ceux présents dans la langue naturelle. Un premier pas dans cette direction a été déjà fait par la modélisation des raisonnements abductifs (Rouault, 1998), dont la mise en œuvre reste cependant problématique. Pourtant, c'est un chantier d'avenir parce que la pertinence et l'efficacité des opérations d'enrichissement automatique de la base de connaissances en dépendent largement.

En troisième lieu, la présentation de la réponse en langue naturelle, impliquant la participation d'un générateur linguistique, soulève la question de son adaptation avec le formalisme de la base de connaissances. Tout comme l'analyse, la génération nécessite la conception d'un outil original, cohérent avec les autres composantes du système et exclut toute possibilité d'importation d'un générateur indépendant. Des travaux proches de cette optique sont en cours (Balicco et al., 2000), mais il reste à vérifier concrètement leur compatibilité avec le système dont il est question ici.

Enfin, un problème auquel il faudra s'attaquer dans l'avenir est celui de l'adaptation de ce système pour une recherche sur un flux informationnel. Les contraintes de l'analyse linguistique rendent a priori cette tâche très difficile. En effet, pour effectuer les traitements, un aller-retour permanent entre le système et le texte est indispensable, ce qui semble difficilement envisageable si le texte n'est pas stocké et disponible en permanence. Cependant, vue l'ampleur des échanges informationnels en ligne, leur analyse ne doit pas être négligée. Une solution réaliste semble la conception d'un système approprié de filtres permettant d'identifier les textes susceptibles d'apporter des connaissances importantes pour le domaine d'application et de les enregistrer dans une base de données. Les traitements continueraient ainsi à se faire sur des textes stockés, mais prendraient également en compte la circulation sur les réseaux.

Ces questions restent ouvertes et esquissent autant de perspectives de développement : pour la recherche d'information et le traitement des connaissances en particulier, mais

également pour toute application qui implique le traitement du sens, l'analyse et la génération linguistiques, ou les raisonnements automatiques.

RÉFÉRENCES BIBLIOGRAPHIQUES

Alpac, *Language and Machines : Computers in Translation and Linguistics. Report by the Automatic Language Processing Advisory Committee*, National Academy of Sciences & National Research Council, Washington D.C., 1966

Balicco, Laurence, Ponton, Claude et Pouchot, Stéphanie, « La génération automatique des textes. Une aide à la recherche d'informations », in *Revue d'Intelligence Artificielle*, vol. 14, n° 1-2, Hermès, Paris, 2000

Bangemann, André, *L'Europe et la société d'information planétaire. Recommandations au Conseil européen*, Bruxelles, 1994

Bar-Hillel, Yehoshua, *Language and Information : Selected Essays on their Theory and Application*, Addison Wesley, Mass., 1964

Barr, Avron et Feigenbaum, Edward (sous la direction de), *The Handbook of Artificial Intelligence*, vol. 1-4, Addison Wesley, New York, 1981-1989

Berrendonner, Alain et Fredj, Mounia et Oquendo, Flavio et Rouault, Jacques, « Un système inférentiel orienté objet pour des applications en langue naturelle », in *Coling* (actes du colloque), Nantes, 1992

Bertrand, Roland (sous la direction de), *Micro-ordinateur et traitement de l'information*, À Jour Editions, Paris, 1991

Bradley, Allen, »Case-Based Reasoning : Business Applications », in *Communications of the ACM*, vol. 37, n° 3, mars 1994

Castells, Manuel, *La société en réseaux : l'ère de l'information*, Fayard, Paris, 1998

Culioli, Antoine, *Cognition and Representation in Linguistic Theory*, (textes choisis par Michel Liddle), John Benjamins, Amsterdam/Philadelphia, 1995

Culioli, Antoine, *Pour une linguistique de l'énonciation (tome 3) : Domaine notionnel*, Ophrys, Gap, 1999

Dreyfus, Hubert, *Intelligence artificielle : mythes et limites*, Flammarion, Paris, 1984

Eymard, Gilbert, *Traitement documentaire des sommaires : des mots-clés à l'extraction des connaissances. Application à une documentation technique*, thèse de doctorat, université Pierre Mendès France (Grenoble 2), Grenoble, 1992

Fillmore, Charles, « The Case for Case », in Bach, Emmon et Harms, Robert (sous la direction de), *Universals of Linguistic Theory,* Holt, Reinhart & Wilson, New York, 1968

Fuchs, Catherine, *Les typologies de procès*, Klincksieck, Paris, 1991

Gross, Maurice, *Méthodes en syntaxe, régime des constructions complétives*, Hermann, Paris, 1975

Haugeland, John, *Artificial Intelligence. The Very Idea*, Bradford Books, MIT Press, Cambridge/Bradford, 1987

Hayes-Roth, Frederick et Jacobstein, Neil, « The State of Knowledge-Based Systems », in *Communications of the ACM*, vol. 37, n° 3, mars 1994

Lacroix, Jean-Guy et Tremblay, Gaëtan, *The Information Society and the Cultural Industries Theory*, Sage, Londres-Toronto, 1998

Lallich-Boidin, Geneviève, *Analyse syntaxique automatique du français. Application à l'indexation automatique*, thèse de doctorat, université Pierre Mendès France (Grenoble 2), Grenoble, 1986

Lenat, Douglas et Miller, George et Yokoi, Toshio, « CYC, WordNet and EDR : Critiques and Responses », in *Communications of the ACM*, vol. 38, n° 11, novembre 1995

Lenat, Douglas, « CYC : A Large Scale Investment in Knowledge Infrastructure », in *Communications of the ACM*, vol. 38, n° 11, novembre 1995

Lenat, Douglas, « CYC : Toward Programs with Common Sense » in *Communications of the ACM*, vol. 33, n° 1, janvier 1990

Manès-Gallo, Maria-Catarina et Rouault, Jacques, « Schémas d'action et types de procès », in *Intellectica*, n° 13, Association pour la recherche cognitive et CNRS, Nanterre, 1992

Miège, Bernard, *La société conquise par la communication (2). La communication entre l'industrie et l'espace public*, Presses Universitaires de Grenoble, Grenoble, 1997

Mounier, Evelyne, *Étude expérimentale de la segmentation d'un texte en paragraphes*, thèse de doctorat, université Stendhal (Grenoble 3), Grenoble, 1996

Ouerfelli, Tarek, *La segmentation des documents techniques composites dans une perspective d'indexation. Vers la définition d'un modèle dans une optique d'automatisation*, thèse de doctorat, université Stendhal (Grenoble 3), Grenoble, 2001

Paganelli, Céline, *La recherche d'information dans des bases de documents techniques en texte intégral. Étude de l'activité des utilisateurs*, thèse de doctorat, université Stendhal (Grenoble 3), Grenoble, 1997

Parsaye, Kamran et Chignell, Mark, *Expert Systems for Experts*, John Wiley & Sons. Inc., New York, 1988

Rouault, Jacques, « About Abductive Reasoning. Advances in Knowledge Organisation », in *Proceedings of the Fifth International ISKO Conference*, 25-29 August, 1998, Lille, France, Springer Verlag, Berlin, 1998

Rouault, Jacques et Manès-Gallo, Maria-Catarina, *Intelligence linguistique : le couple sémantique-pragmatique et le calcul des énoncés élémentaires* (à paraître)

Sabah, Gérard, *L'intelligence artificielle et le langage*, Hermès, Paris, 1990

Tremblay, Gaëtan, « La société de l'information : du fordisme au *gatesisme* », in *Communication*, vol. 16, n° 2, Québec, 1995