Indexation automatique et la Recherche D'information dans les documents

MAHMOUDI Seyed Mohammad
Université de Téhéran
mahmoudi@ut.ac.ir

1- Introduction : Indexation automatique et le TAL

Le développement des systèmes d'informations automatiques, lié d'une part à l'accroissement de la technologie de l'information (TI), et d'autre part au développement des autoroutes de l'information, a considérablement favorisé le développement des systèmes d'informations spécialisés pour la recherche d'information dans un univers infiniment vaste (MAHMOUDI,2002,PP.171-194).

L'une des perspectives ouvertes par l'informatique, particulièrement fascinante, est la conception de "systèmes" et de "machines" qui comprendraient notre propre langue. En fait, un tel objectif est très mal défini et ses applications sont souvent mal connues.

Traditionnellement, le traitement automatique des langues naturelles (TAL) est l'un des domaines essentiels de l'intelligence artificielle (IA). Au vu de nombreux résultats incomplets et des tâches inachevées on constate, en effet, que le traitement automatique de documents écrits en langue naturelle dont de nombreux informaticiens et spécialistes de la transmission de l'information rêvent maintenant depuis près de trente ans n'en est encore qu'au stade de la petite enfance" (METZGER,88,P.13). De nombreuses méthodes et approches de natures très variées ont été mises en œuvre : chacune de ces méthodes s'applique à un ensemble de problèmes très limités, aucune de ces méthodes ne représente une solution universelle.

L'objectif du TAL est la conception des systèmes et des programmes informatiques capables de traiter d'une façon automatique des données linguistiques (écrites ou orales) exprimées dans une langue dite « naturelle » – où *langue naturelle* qui s'oppose à *langage artificiel* (informatique, mathématique, logique, etc.) –, en vue de la compréhension et la production des informations ou des connaissances nouvelles.

Les programmes informatiques que conçoit le TAL et leurs applications dans des tâches précises ne constituent toutefois que la partie émergée du domaine. L'informatique est ici au service de linguistique, comme support pour la description des phénomènes linguistiques (DAL, HATHOUT, NAMER,2004,P.3). En effet, l'élaboration de ces programmes suppose un travail en amont qui peut se situer sur plusieurs domaines comme : *linguistique, logique, mathématiques, statistiques* et *informatique* (FUCHS,1993,P.13).

Les applications dans le domaine du traitement automatique des langues naturelles se sont considérablement étendues et cette tendance s'accroît continuellement. Ces applications peuvent être rangées en deux grandes catégories :

- celles qui nécessitent une *"analyse automatique "* de textes en vue d'une reconnaissance partielle ou complète des unités qui le constituent : mots, phrases, concepts. On trouvera ici, par exemple la recherche d'information, la correction orthographique et syntaxique, la documentation et la traduction automatique, l'interrogation en langues naturelles de bases de données scientifiques et techniques (interface et dialogue homme/machine en langage naturel), ou de bases de données à la formation et enseignement assisté par ordinateur (EAO), le traitement de la parole, le contrôle de systèmes informatiques ou automatiques (commande de robots) : dialogue avec un système expert ou un robot, l'extraction d'information ou indexation automatique de document

- celles qui comportent une *"génération automatique"* de formes linguistiques (construction d'une forme syntaxique ayant un sens), pour un système de traduction automatique, par exemple, ou bien pour formuler des réponses à des questions et l'élaboration automatique de résumés de textes. On peut aussi citer la résolution de problèmes en langue naturelle (programmation d'ordinateur, à la limite), la production automatique de lettres (par exemple, réponse aux lettres de réclamation).

Parmi les nombreuses applications du TAL, on peut accorder une place privilégiée à l'indexation automatique. Cette approche, qui est basée, à priori, sur une analyse linguistique, est féconde et permet des solutions importantes au repérage du contenu du texte.

La fiabilité et le bon déroulement de l'indexation automatique, en vue de la reconnaissance automatique du contenu de documents, impose au préalable la nécessité d'une réflexion théorique sur les opérations qui en constituent les composantes. La mise en place d'un système de représentation du contenu du texte est donc au centre de toute application automatique en langue naturelle. Ce système doit permettre le cheminement de trois étapes majeures et étroitement liées :

1- Définir un univers référentiel de la représentation du "contenu du texte". Qu'est ce que le contenu du texte ?

2- Définir une stratégie de traitement qui permettrait, à l'aide de certains formalismes et des outils théoriques et techniques, le repérage du contenu du texte. Il s'agit en fait du passage de la forme originale du texte, en langue naturelle, à la représentation de son contenu, en langage documentaire.

3- Constituer une base de données spécifique qui contiendrait l'ensemble des éléments informatifs répertoriés dans le texte. Cette base de données doit en principe permettre aux chercheurs une consultation pertinente sur les informations contenues dans le texte.

Dans les lignes qui suivent, nous allons d'abord présenter brièvement quelques approches de la définition de représentation du contenu du texte, puis nous décrirons certaines méthodes d'analyse les plus fréquentes, surtout la méthodologie de la conception d'un analyseur MS capable de traiter le contenu du texte.

2- La Représentation du contenu de texte

L'indexation automatique s'effectue par une analyse documentaire. C'est à dire par un processus plus ou moins formalisé qui permettrait "l'extraction du sens des documents" (GARDIN,74,P.120). Autrement dit l'analyse documentaire devrait aboutir à la représentation du contenu d'un document.

En fait, une telle définition paraît bien générale et bien ambiguë : qu'est ce que le "sens" d'un document ? Peut-on "extraire" le sens d'un document ? Etc.

Il n'existe malheureusement pas encore de procédure qui conduise systématiquement à une définition encyclopédique et universelle du "contenu", -i.e. le "sens" d'un document. La plupart des chercheurs définissent selon leurs propres préoccupations et en fonction de leurs objectifs particuliers ce qu'ils entendent par représentation du contenu d'un document.

Afin d'avoir une idée plus précise de la notion du contenu, nous présentons ici certaines définitions terminologiques les plus fréquentes.

2.1- Recherche d'information

La recherche d'information (RI) est un ensemble de techniques et d'outils informatiques dont la finalité initiale était bibliographique : il s'agissait d'aider les usagers à trouver, dans des fonds documentaires, les références concernant un thème particulier. L'amélioration des capacités de stockage des ordinateurs a changé la nature du problème, qui n'est désormais plus d'exploiter des notices bibliographiques mais de conserver et d'accéder directement aux informations textuelles contenues dans les documents qui constituent les fonds (DAL, HATHOUT, NAMER,2004,P.14).

Les techniques de la recherche d'information qui été longtemps réservée aux spécialistes de la documentation est aujourd'hui très largement utilisées dans la catégorisation et classification des documents multimédias, le catalogage des documents et références bibliographiques et l'accès à leur contenu, résumer des textes, et surtout l'analyse de contenu et accès aux informations recherchées à travers les moteurs de recherche sur Internet.

Le problème le plus important de la recherche d'information dans un système d'indexation automatique est de distinguer dans un ensemble de documents volumineux et complexes ceux qui contiennent des informations de ceux qui ne les contiennent pas. L'opération d'indexation automatique est donc particulièrement difficile dans la mesure où elle pose de problème de l'interprétation et la représentation du sens du texte.

Les systèmes de recherche d'information doivent ainsi disposer d'un modèle de la représentation des informations contenues dans les documents, et d'une procédure permettant de déterminer leur pertinence comme réponses à une requête particulière.

Un tel objectif est malheureusement très difficile à atteindre en TAL, étant donné la nature et la structure complexe de la langue qui répond mal aux formalismes bien structurés de l'informatique. Les systèmes de recherche d'information peuvent donc au mieux analyser et repérer approximativement le sens des informations, et évaluer leur proximité avec celui de la requête, de façon à classer les documents en fonction de leur pertinence comme réponses à la requête.

Pour concevoir un système de recherche d'information documentaire le plus adapté, les approches les plus utilisées à l'heure actuelle relèvent de techniques comme l'élaboration d'un analyseur MS permettant de repérer des expressions particulière ou des SN, qui sont en fait des éléments les plus informatifs d'un document. Nous verrons *infra* comment on peut concevoir un analyseur MS.

Les réponses, souvent médiocres, des systèmes de recherche d'information comme les moteurs de recherche montrent que des progrès importants restent à faire. L'utilisation de représentations linguistiquement motivées est l'une des pistes qui pourrait conduire à l'amélioration des résultats des systèmes de recherche d'information. Bien que de nombreuses expériences d'intégration d'outils de TAL dans des systèmes de RI n'aient pas été concluantes, la question de l'apport du TAL à la RI reste ouverte (DAL, HATHOUT, NAMER,2004,P.15).

2.2- Les mots clés

Pendant longtemps, l'analyse documentaire a été réduite à un ensemble d'opérations systématiques par lesquelles on constituait une base de données qui contenait quelques enregistrements dont chacun contenait un certain nombre de champs et un ensemble de mots dits "mots clés" repérés dans le texte au cours de l'analyse.

Dans une application documentaire, les mots clés s'employaient souvent comme des références pour désigner le contenu du texte. Rappelons que dans de telles analyses, les mots vides comme des articles et des prépositions ainsi que les caractères non alphabétiques (chiffres, signes de ponctuation, trait d'union etc.) ne sont pas pris en compte comme des éléments d'indexation.

Ce type d'indexation utilisé en particulier par IBM dans STARIS (Storage And Information Retrieval System) a pour avantage sa très grande simplicité de mise en œuvre informatique. Il présente néanmoins de nombreux inconvénients, notamment en ce qui concerne la porté informative très faible de ces mots. Il est donc nécessaire de préciser que les mots clés ne sont que des mots de la langue, c'est à dire les mots du lexique. Chaque mot du lexique peut en fait représenter autant de prolifération qui est prévue dans une entrée lexicale. Voici on présente quelques inconvénients de l'indexation automatique basée sur les mots clés :

1- Il n'y a pas de normalisation des mots. Les deux mots « système » et « systèmes » sont différemment perçus.

2- Du fait de l'absence d'une analyse plus profonde, comme l'analyse syntaxique, et en défaut de certain mots vides, chaque élément des syntagmes nominaux et des mots composés et des expressions variées serait isolé et interprété indépendamment de leur contexte ; ce qui fausse effectivement le résultat d'une analyse et introduit inutilement du bruit.
Par exemple, le mot composé "Royaume-Uni" devient :
"Royaume" (uni terme) et "Uni" (uni terme).
Ainsi le syntagme nominal "chemin de fer", après élimination des mots vides, devient :
"chemin" (uni terme) et "fer" (uni terme).
Rappelons que dans des systèmes plus élaborés, les mots-clés sont déterminés automatiquement par l'élimination des mots fonctionnels (mots supposés « non informatifs » comme l'article, la préposition, l'auxiliaire, etc.) de l'ensemble des mots du corpus. Ceux qui restent après filtrage obtiennent le statut de mot-clés au sens du mot « informatif ».

3- Les synonymes et les homonymes ne sont pas distingués.

4- Les ambiguïtés morphologiques des mots ne sont pas résolues. Dans le cas du français, 30% en moyenne des mots pris isolément sont ambigus. L'amélioration des systèmes d'indexation automatiques passe par la définition des procédures plus rentables que les mots-clés et la sélection de termes de manière plus fine rend l'indexation plus efficace (SIDHOM,2002,P.37).

2.3- Les descripteurs

On a souvent employé la synonymie généralement admise entre "mot clé" et "descripteur" pour désigner les éléments de la langue considérés comme des descripteurs d'un fonds documentaire. En fait une telle confusion est très fâcheuse car les systèmes documentaires n'ont pas en principe pour finalité de fournir à l'utilisateur des renseignements sur les mots mais sur les choses.

En effet, dans la pratique documentaire, il est inadmissible de confondre les "descripteurs" et les "mots clés". Alors que les mots clés ne sont que les mots de la langue, c'est à dire des symboles, des éléments du lexique, "le descripteur, quant à lui, signifie une unité, une substance au sens de philosophie d'Aristote. Le descripteur ne peut donc pas être considéré, à l'instar des mots de la langue, comme un symbole sans référence. On pourrait vouloir le caractériser comme le mot de la langue actualisé dans le discours, mais cela n'est pas tout à fait suffisant. (...) Le descripteur n'est donc pas mot de la langue, mais syntagme du discours, ou plus exactement syntagme-type par rapport aux syntagmes-occurrences des énoncés" (LE GUERN,91,PP.23,24).

2.4- Les syntagmes nominaux

Un syntagme est un groupe de mots formant une unité à l'intérieur de la phrase (Ex. le plus beau jour ; passant par là) (Petit Robert). Pour Saussure (cours de linguistique générale), le syntagme se compose toujours de deux ou plusieurs unités consécutives (re-lire ; contre tous ; s'il fait beau temps, nous sortirons). Plusieurs grammairiens, avec Charles Bally, estiment que tout syntagme peut être considéré comme binaire, c'est-à-dire formé de deux éléments : un déterminé et un déterminant. Pour la grammaire dite «nouvelle», toute phrase verbale est constituée par deux syntagmes : Le syntagme nominal et le syntagme verbal sont solidairement unis : *Les petits ruisseaux /font les grandes rivières* (extrait de M. GREVISSE : Le bon usage,2001,p.27).

Pour Michel Le Guern "le syntagme nominal est la plus petite unité nominale de discours susceptible de servir de base à une relation référentielle autonome qui permet de désigner un objet. Dans un système d'information automatisé, les descripteurs sont les syntagmes nominaux des documents constituant le corpus" (LE GUERN,91,P.24).

Ainsi, le caractère référentiel du SN, pour distinguer les unités de nature syntagmatique et non syntagmatique, est au centre de toute considération sur les SN. À ce propos Jean Paul Metzger écrit :

"au sein d'un discours (≈ un texte), certaines unités peuvent être distinguées par leur caractère référentiel : elles représentent des éléments -des objets, des phénomènes, ...-d'une réalité -un espace? -non linguistique dans laquelle le discours se développe, réalité dans laquelle se situe le locuteur (l'auteur) et sur laquelle il formule un certain nombre de propositions.

Ces unités sont essentiellement des occurrences de syntagmes nominaux (/le petit chien du voisin/), de noms propres (/Paul/,/Paris/), de pronom (/moi/,/je/,/nous/,...) mais aussi de certains "adverbes"(/ici/,/hier/,/tout de suite/,...) et de certaines marques flexionnelles du verbe qui relèvent de la notion de temps.

C'est autours de ces unités, de nature syntagmatique et non lexicale, que se forment ensuite les énoncés du discours sous la forme de propositions, de phrases, de "périodes",...qui ont pour fonction d'apporter de l'"information" sur la réalité ainsi référencée" (METZGER,88,P.26).

Chercher à décrire le contenu d'un document exige donc la mise en évidence des parties du discours à valeur référentielle. "À travers le syntagme nominal, considéré comme la plus petite partie du discours porteuse de référence à la réalité extralinguistique, il est possible de développer une méthodologie de conception de base de données textuelles" (BOUCHÉ,88,P.1).

Le fait d'indexer un document par l'ensemble des syntagmes nominaux qu'il contient, représente de nombreux avantages par rapport aux systèmes documentaires classiques. En effet, cette représentation est riche en information et est représentative du contenu. La non-détection de synonymies entre les syntagmes nominaux est un facteur de silence. La non-prise en compte des relations d'hyperonymie/hyponymie va dans le même sens.

Pour lutter contre le silence dans ce tel mode d'indexation, il est nécessaire de rendre compte des relations de synonymie et d'hyperonymie/hyponymie entre les syntagmes. Dans certains cas, ces relations sont perceptibles à travers l'analyse morpho -syntaxique et dans d'autres cas, il est nécessaire d'intégrer dans le travail un thésaurus en permutant certains termes du syntagme, puis de relancer à nouveau la recherche (SIDHOM,2002,P.77).

3- Les Méthodes d'analyse pour la représentation du contenu

3.1- Méthodes statistiques

Jusqu'à une date très récente, de nombreux systèmes d'indexation utilisaient encore des méthodes statistiques pour sélectionner les éléments "significatifs" d'un corpus donné. En fait dans la méthode statistique, le système en place calcule la fréquence des mots ou des termes significatifs pour évaluer la pertinence de ces mots dans le texte. Ainsi, lorsque la fréquence des mots sont calculée, le système effectuera ensuite une comparaison entre cette fréquence et une table de fréquence moyenne pour un domaine donné. Dans ce contexte, les termes qui ont un taux d'apparition supérieur à la table de fréquence sont retenus comme des termes d'indexation. Ces termes pourraient représenter le contenu du texte.

Les difficultés rencontrées par cette approche sont essentiellement dues aux faibles fréquences observées. Les hapax (mots n'apparaissant qu'une fois) sont en effet très fréquents, surtout en français. La situation devient encore plus impossible si on essaye de prendre en compte des expressions, leur fréquence ne pouvant être que plus faible (BOUCHÉ,88,P.2).

Aujourd'hui la plupart des chercheurs sont d'accord pour estimer que les méthodes statistiques ne sont guère applicables sans une analyse linguistique préalable et approfondie.

3.2- Méthode de projection

Il s'agit ici de projeter le document sur un univers référentiel, structuré a priori. Cet univers qui est construit selon un modèle de réseau sémantique représente le même domaine de la connaissance que le document. Dans la phase de projection, les éléments de document qui sont en accord avec la référence sont retenus.

La réalisation d'une relation d'inférence entre les parties d'un document et la représentation référentielle d'un univers, est extrêmement difficile. Cette réalisation nécessite en fait la structuration de nombreuses relations complexes et la mis au point d'un thesaurus et un réseau sémantique, tout en utilisant des techniques de l'intelligence d'artificielle.

Du fait de nombreuses difficultés concernant les caractéristiques propres de la langue naturelle, cette méthode n'est pas, à ce jour, parfaitement réalisable. Les dérivations morphologiques (genre, nombre, flexions verbales, etc.), et les variations syntaxiques (insertion d'adjectif, adverbe, etc.), qui sont à l'origine de nombreuses fluctuations de formes des mots, rendent le problème extrêmement complexe.

3.3- Analyse de bruits de fond

Ce type d'analyse (voir aussi TOWNSEND,88) permet une recherche la plus simpliste. Analyseur balaie la phrase pour retrouver simplement certains mots clés ou balises. Le reste est simplement ignoré ; ce sont les bruits de fond. Les balises activent ensuite les fonctions prévues dans l'ordinateur.

Cette analyse peut être correct pour des environnements restreints impliquant un vocabulaire limité et peu de commandes associées. L'un des inconvénients est que l'analyse ne se réalise pas correctement sur une phrase contenant une demande ambiguë. Le programme par exemple reçoit la syntaxe correcte mais, sans analyse sémantique, il ne peut exécuter la commande convenable : la non-prise en compte du contexte provoque une erreur.

3.4- Analyse morpho-syntaxique

Dans une application en TAL, nous sommes confrontés à un double problème : utiliser une théorie linguistique est très complexe, ne pas utiliser est trop restrictif et aucune des deux démarches n'est réellement satisfaisante si on prend en compte les paramètres fondamentaux que sont la cohérence, la généralité, l'efficacité et la réutilisabilité, mais aussi la gestion de développement des descriptions (BLACHE,2001,P.115). Nous proposons ainsi une approche linguistique explicite basée sur la conception d'un analyseur MS.

L'analyseur MS est probablement le mieux adoptée et le plus universel pour l'indexation automatique des documents. Cet analyseur, qui est le plus puissant, mais aussi le plus complexe, permet ainsi d'extraire les syntagmes nominaux.

Traditionnellement, la conception d'un analyseur MS s'effectue en plusieurs grandes étapes dont les deux premières sont l'analyse morphologique et l'analyse syntaxique. L'analyse syntaxique proprement dite s'effectuera une fois l'analyse morphologique réalisée. Mais du fait de nombreuses ambiguïtés rencontrées au cours de l'analyse morphologique, une analyse syntaxique partielle s'impose avant même l'analyse morphologique proprement dite.

3.4.1- Analyse morphologique

Les premiers objectifs de l'analyse morphologique sont : segmenter le corpus, c'est-à-dire le diviser en segments qui représentent chacun un seul morphème ; classer ces segments en morphèmes et catégories morphologiques, leur attribuer des traits métalinguistiques.

Il est clair qu'une analyse morphologique non ambiguë facilite l'analyse syntaxique qui suit. En d'autres termes, *"l'imbrication des phénomènes morphologiques et syntaxiques introduit la nécessité de construire une grammaire de reconnaissance où les règles sont dépendantes du contexte d'occurrence de chaque constituant. Analyse morphologique et syntaxique s'entre-déterminent donc et devront ainsi être constamment entremêlées"* (Berrendonner, Bouché, Le Guern, Rouault,1981,pp.3-28).

3.4.2- Analyse syntaxique

En règle générale lorsque l'analyse morphologique est réalisée on peut effectuer une analyse syntaxique. Cependant comme Michel Le Guern précise "pour résoudre les ambiguïtés morphologiques, il convient d'avoir préalablement commencé une analyse syntaxique partielle". Il conviendra aussi de

prendre en compte le contexte et le contenu logique de chaque forme dans la phrase. Car "pour des raisons historique et formelles, l'analyse syntaxique entretient d'étroites relations avec la logique" (BLACHE, 2001, p.43).

L'analyse syntaxique, effectuée par des **analyseurs syntaxiques,** consiste à explorer les textes préalablement traités sur un plan morphologique et lexical pour en extraire les règles de construction et d'agencement des mots, de manière à pouvoir analyser la structure des phrases. Diverses méthodes sont utilisées par les analyseurs syntaxiques pour rechercher ces différentes façons de regrouper les mots. Les uns utilisent des démarches formelles, élaborées à partir de 1955 par des linguistes et des logiciens, comme *les grammaires transformationnelles, descriptives, catégorielles,* conçues à l'origine pour aborder avec rigueur les difficultés posées par la traduction automatique, ou encore de toutes nouvelles grammaires, dites à *réseaux de transition augmentés ("augmented transition network") ou fonctionnelles lexicales ("lexical function grammars"),* plus proches du formalisme informatique. D'autres utilisent des démarches probabilistes et statistiques, pour étudier en priorité les formes de regroupement de mots les plus vraisemblables ou les plus probables. Aucune grammaire formelle ne parvient certes à résoudre toutes les difficultés grammaticales et logiques que pose une langue donnée. Ainsi la syntaxe permet de définir des règles de réécriture ou plutôt de combinaison et de ré-agencement de mots et, permet aussi, d'élaborer des programmes informatiques de reconstruction et de recréation de textes (Voir aussi VUILLEMIN,87).

3.5- Analyse logico-sémantique

Le traitement de l'écriture, pour être efficace, doit associer les phases de reconnaissance morphologiques avec les contraintes dues au contexte (OLLIVIER, WEINFED,2000,P.221).

L'approche morpho-syntaxique, qui est souvent utilisée par de nombreux chercheurs en TAL, est basée principalement sur une grammaire hors contexte. Cette limitation de connaissances ne fournira pas tout à fait un traitement adéquat et exhaustif ; il faudra donc envisager d'autres analyses et connaissances, comme l'analyse logico-sémantique, pragmatique et du contexte, pour minimiser l'ampleur de nombreuses ambiguïtés qui sont générées par l'analyseur.

L'analyse logico-sémantique, qui est calquée sur des théories logiques, est cruciale pour le TAL, cette analyse constitue en fait à définir un modèle sémantique, c'est à dire un ensemble de concepts et de règles qui permettent l'interprétation de formes purement syntaxiques.

L'approche logico-sémantique joue un rôle très important pour déterminer le contenu référentiel du texte. Cette approche se démarque des autres analyses (par exemple l'analyse MS) par sa propre définition de "référence".

3.5.1- Les propositions

Comme nous l'avons vu, au paravent, l'approche linguistique et surtout syntaxique, considère les syntagmes nominaux ou les descripteurs comme des unités ou des petites parties de discours qui sont porteuses du sens et références à la réalité extra linguistique du discours ; dans cette analyse on ne tient pas compte, ni de la coordination (C), ni de l'apposition (T), ni du verbe (V).

Dans un formalisme logico-sémantique, s'appuyant sur une réflexion linguistique, les ***propositions*** sont, en fait, considérées comme des énonciations qui dénotent une valeur de vérité (ou une valeur référentielle). Notons que la logique classique analysait toute proposition en sujet, copule et prédicat : S ε P (S est P). Le rôle du sujet est de désigner des objets, le rôle du prédicat est d'exprimer une propriété sur ces objets. Le verbe être (le copule) relie le sujet au prédicat. Ces prédicats sont à rapprocher des fonctions logiques à un argument ou propriété $f(X)$. Plus généralement, on peut extrapoler la notion de prédicat à la partie de la proposition qui "dit quelque chose", par opposition au reste qui est censé désigner. On rejoint alors la notion de fonction logique (extrait de DUPONT,90,P.296).

Exemples : [le chien] [aboie] ou [le chien] (est) [aboyant]
[Paul reçoit une lettre] ou [Paul] (est) [recevant une lettre] ou
[Paul] reçoit [une lettre]
[Marie] envoie [une lettre] à [Paul]

On distingue ainsi des termes (Paul, Marie, une lettre, le chien) et des prédicats ou fonctions logiques (reçoit, envoie-à). On remarque immédiatement qu'un SN a, dans le cadre de telles analyses, le statut d'un terme.

Le caractère référentiel du SN associé à son contenu prédicatif semble donc faire du SN une unité informationnelle centrale.

3.5.2- Les quantificateurs

Il est certain que les propositions contribuent aussi à la représentation de l'information contenue dans le discours, cependant parmi toutes les propositions, certaines sont plus "informatives" que d'autres. La question qu'on se pose c'est de savoir comment les distinguer. La réponse est la suivante :

Le degré d'informatif d'une proposition dépend en fait à la propriété du quantificateur qu'elle possède. Il faudra donc étudier de près les quantificateurs.

L'étude des quantificateurs permet de mettre en place l'opposition centrale entre «**discret**» et «**continu**», et suggère déjà une utilisation de la logique de Montague (LE GUERN préface pour DUPONT,90,P.6).

Les quantificateurs linguistiques, comme leur homologue logique ont la charge de lier une variable. Ici nous avons "extrapolé" cette opération dans le passage du contenu au discret. Les quantificateurs linguistiques sont approximativement représentés par les prédéterminants. Une quantification linguistique, est une opération qui a pour effet de fermer un assemblage ouvert. Explicitée par l'antirelation <continu/discret>, la quantification opère le passage du contenu au discret. Le quantificateur linguistique opère sur un nom commun. Mais un groupe quantifié est aussi susceptible de fermer un assemblage ouvert, c'est à dire de le faire glisser du contenu vers le discret (DUPONT,90,PP.297-298).

Exemples : chien noir -----> continu
son chien noir -----> discret intermédiaire
le chien noir -----> discret

Le degré de ouverture/fermeture, continu/discret, dépend essentiellement du degré de l'individualisation de l'univers de discours. De ce fait nous concluons que plus l'univers de discours est fermé, plus il est discret et plus il est concret et informatif.

3.5.3- Logique extensionnelle et intensionnelle

Ce paragraphe utilise, essentiellement, le document (LAINÉ,LAROUK,VIDALENC,88,PP.1-11).

L'emploi simultané des logiques intensionnelle et extensionnelle en TAL est aussi indispensable. La logique extensionnelle s'oppose à la logique intensionnelle par l'existence ou non d'un univers de référentiel dans lequel sont pris les objets manipulés. Autrement dit cette opposition s'illustre dans l'opposition entre langue et discours.

On peut définir l'extension d'un concept par l'ensemble des objets auxquels ce concept s'applique. Par exemple, l'extension du concept "homme" est la classe des objets dont on peut dire qu'ils vérifient la propriété "homme" dans l'univers de référence.

La logique extensionnelle n'est autre que la logique des classes (i.e. la logique classique). La logique intensionnelle porte sur des prédicats libres simples au moyen de l'opérateur binaire de composition que nous noterons *, pour construire des prédicats libres complexes.

Pour produire un discours, c'est à dire le passage de la logique intensionnelle à la logique extensionnelle, le locuteur construit des éléments à partir des prédicats de la langue. Ces éléments sont des références à des objets de la réalité extra-linguistique. Pour construire l'unité minimale du discours (i.e. le SN), il faut se livrer à une opération logico-sémantique qui est l'opération de fermeture. Par exemple, dans "le cheval", l'article "le" est le quantificateur qui réalise l'opération logique de fermeture.

SN -----> quantificateur + prédicat

Dans l'exemple suivant : "chien noir à taches blanches", on n'utilise aucun article, l'expression se situe donc au niveau de la logique intensionnelle. Tandis que "le chien noir à taches blanches", se situe au niveau de la logique extensionnelle. L'article défini "le" nous permet donc le passage de la logique intensionnelle à la logique extensionnelle.

La dualité et la complémentarité de ces approches permettent donc d'aboutir à une représentation de la structure logico-sémantique des textes très pertinente par rapport au contenu effectif du discours.

3.6- Analyse pragmatique.

Analyse pragmatique correspond au dernier niveau d'analyse des programmes de compréhension des textes et des TAL (Les analyses précédentes qui fait passer de la phrase au «sens», est souvent appelée par les informaticiens : «phase de compréhension»). Les analyses morpho-syntaxiques et logico-sémantiques précédentes ayant permis d'aboutir à certaines inférences ou conclusions sur le sens latent d'un énoncé, à partir, en somme d'une analyse intrinsèque de ses significations, les analyseurs pragmatiques sont d'autres programmes de traitement qui vont essayer de comparer ces résultats ou inférences à ce que l'on peut déduire au contraire d'une analyse extrinsèque du contexte de cet énoncé. En effet, toute phrase écrite par un auteur ou dite par un locuteur est prise dans un réseau de relations et de correspondances avec ce qui a été écrit ou dit et avec ce qui sera dit ou écrit, implicitement ou explicitement. Seule l'étude de ces contextes peut lever ces ambiguïtés, sous réserve de pouvoir définir d'autres règles d'analyse, dites pragmatiques, déductives et inférentielles, capables de parvenir à ce résultat (Voir aussi VUILLEMIN,87).

4- La Conception d'un analyseur Morpho-Syntaxique

Décrire le contenu d'un document c'est avant tout dégager les partie du discours à valeurs référentielles. Dans ce sens, les SN peuvent constituer la base pour représenter des entités du discours.

L'indexation automatique suppose donc la conception d'un analyseur MS capable de repérer automatiquement les SN qui, dans une perspective de recherche documentaire, sont considérés comme les éléments les plus informatifs.

Dans cette partie, nous allons d'abord présenter les différentes étapes de la conception d'un analyseur MS pour la reconnaissance automatique des SN dans les textes rédigés en français, puis nous essayerons de donner quelques solutions informatiques.

1ère Étape : Le choix du corpus

Le repérage du contenu du texte nécessite au préalable la définition d'un univers du travail, c'est à dire le choix d'un corpus particulier qui correspond aux applications envisagées. Le corpus doit réunir un ensemble des textes variés mais homogènes, de point de vue de la forme et du style (littéraire, familier etc.). À ce titre, il est important de préciser que tout style de la langue ne se prête pas à une analyse automatique correcte ; dans certains styles littéraires dont l'emploi de métaphores est très fréquent il y a une impossibilité absolue d'envisager une indexation automatique adaptée. De ce fait la

première tâche du concepteur en TAL consiste à sélectionner dans chaque document les textes qui lui permettront d'effectuer une analyse moins ambiguë.

Le choix d'un corpus adapté pour l'analyse automatique de la langue française dont la diversité des styles et la richesse font le charme, est une tâche très difficile. À ce propos il est important de préciser que parmi les trois styles principaux de la langue, utilisés dans les expressions différentes, comme la langue littéraire, classique et familiale, la langue classique ou standard (contemporaine) qui est la langue des journaux, de la radio, des conférences, des livres scientifiques, de l'enseignement des écoles et des universités se prête mieux au TAL. Bien que cette langue n'échappe pas complètement à la tentative grandissante des autres styles pour exercer leur influence lexico-syntaxique, d'où de nombreuses irrégularités pour réaliser une analyse correcte de la langue.

2ième Étape : Analyser les contraintes et les ambiguïtés

Dans une perspective du TAL, les ambiguïtés de la langue française sont très nombreuses. La désambiguïsation des différentes formes d'ambiguïtés nécessite une recherche indépendante et approfondie. Dans le cadre de cet article nous n'avons pas la possibilité d'analyser et lever toutes les ambiguïtés. Cependant nous analysons ici un certain nombre de contraintes et d'ambiguïtés les plus fréquentes.

a- Les propriétés générales :

Certaines propriétés caractéristiques de la langue française, qui se retrouvent d'ailleurs dans toutes les langues, indépendamment des réalisations syntaxiques, rendent son traitement automatique délicat. Ces propriétés correspondent à trois aspects essentiels suivants :

1- La polysémie (homonymie, homotaxies) est le fait qu'à un mot donné peuvent correspondre plusieurs sens distincts (homonymie) ou qu'à une forme de phrase donnée correspondent des interprétations diverses (homotaxie).

Pour montrer un cas "homotaxie" l'exemple suivant, qui est devenu un exemple canonique, peut être intéressant : "le pilote ferme la porte". Cet exemple peut amener la machine à une double analyse :

1- $(\text{le pilote})_{SN}$ ferme_{V} $(\text{la porte})_{SN}$;

2- $(\text{le pilote ferme})_{SN}$ la_{Y} porte_{V} ;

2- La paraphrase (synonymie, allotaxie, définition) correspond au phénomène symétrique qui fait qu'un même concept peut être énoncé de façons diverses : plusieurs mots possèdent le même sens (synonymie), des phrases différentes recouvrent la même idée (allotaxie) et des équivalences existent entre un mot et une phrase (définition) ;

3- Le rapport au contexte, qui englobe anaphore, implicite, métaphore, repérage, trope (la catachrèse, la métonymie, la synecdoque, la litote, l'antiphrase, l'hyperbole), implique des références à l'ensemble du discours et à la situation du locuteur pour rechercher les antécédents des pronoms ou reconnaître ce que désigne telle ou telle expression (anaphore), pour mettre en évidence ce qui est resté implicite (non dit ou glissement de sens d'un mot par rapport à son sens propre, ce qu'on désigne du terme général de trope), ou pour s'adapter à l'interlocuteur, comprendre ses buts et situer ce qu'il dit par rapport à sa situation particulière au moment de l'énonciation (repérage).

À ces problèmes il faudra probablement en ajouter d'autres qui assurément ne peuvent trouver de solution que par une analyse plus profonde comme l'analyse du contexte et logico-sémantique.

b- Les mots composés

Définir le mot dans le contexte de TAL n'est pas chose facile. Isoler un mot simple ne présente pas de grandes difficultés; au contraire du langage parlé, le langage écrit fait apparaître explicitement les séparateurs (espace, ponctuation), mais l'identification des formes composées est un problème compliqué. Comment peut-on distinguer un verbe composé constitué d'un nom et d'un infinitif alors qu'ils se sont séparés tous les deux par un espace (blanc)? Une forme est en fait une unité de traitement

qui est la suite de caractères comprise entre deux blancs et ne comportant aucun blanc. Un texte est défini comme une suite de formes dont chaque forme ne peut s'approprier qu'une seule catégorie grammaticale ou morphologique.

Devant une telle situation, il semble actuellement clair qu'on ne peut pas effectuer une segmentation cohérente et correcte, tant qu'on n'a pas régularisé manuellement un certain nombre de faits problématiques au préalable. De ce fait pour ce qui concerne les mots composés nous proposons les solutions suivantes pour éviter des ambiguïtés éventuelles :

- Lorsque le texte est saisi au terminal, au fur et à mesure de la frappe, il est possible d'enregistrer les mots composés sans aucun espace entre leurs éléments constitutifs. En outre, il est aussi possible d'intercaler simplement un trait d'union entre les mots composés; le trait d'union peut souvent jouer un rôle d'union.

Ex.

|valeur*absolue| =====> |valeurabsolue| ;

ou

|valeur*absolue| =====> |valeura-bsolue| ;

Certains chercheurs comme A. Eyango et M. De Brito, proposent dans leurs thèses, pour la langue française, un schéma relationnel pour la reconnaissance des formes composées. Dans ce modèle chaque élément du mot composé considéré comme une unité minimale du discours sera intégré dans le lexique; le repérage des mots composé s'effectuera donc selon des règles de combinaisons des mots composés, en constituant d'un analyseur morphologique, dans lequel un schéma relationnel est introduit, et en consultant le lexique. Pour ce faire, l'analyseur vérifie si la forme cherchée est susceptible d'être le premier terme d'une séquence. Pour cela, comme écrit A. Eyango, il faut adjoindre à chaque graphie simple un indice d'appartenance (Ind-App) ou non à une séquence de termes. Il existerait donc une dépendance fonctionnelle du type : graphie -----> indice App-.

- En ce qui concerne les séquences de mots de hautes fréquences, comme les locutions diverses, dont les éléments constitutifs sont en principe séparés par un espace, pour éviter tout découpage abusif et les démarches embarrassantes, comme par exemple l'insertion des traits d'union etc.., on peut les ranger dans un lexique spécifique comme des unités indépendantes; dans ce cas les espaces internes ne seront pas tenues en compte. Lexique contient donc toutes les chaînes de caractères non segmentables comportant des espaces.

c- Les formes contractées ou affixées

Dans la plupart des langues du monde, il existe un certain nombre de formes ambiguës, "contractées" ou "affixées" qui se prêtent mal à une analyse et à une classification morphologique directe. Il est donc nécessaire d'envisager une phase de pré-traitement ou de régularisation morpho-syntaxique qui précède l'analyse MS proprement dite pour repérer et régulariser ses formes.

Le pré-traitement morpho-syntaxique doit, en effet, remplir les tâches et les objectifs suivants :

- Décomposer, en des séquences équivalentes de formes catégorisables, certaines formes qui ne pourraient être intégrées dans la classification;
- Dissocier les deux fonctions syntaxiques portées par un seul mot;
- Éliminer de la surface toutes les formes résultant d'un amalgame;
- Réduire le nombre de mots à classer ;
- Ramener une séquence exceptionnelle à une séquence plus générale et régulière;
- De ne pas avoir à introduire une catégorie particulière.

Dans cette perspective, les solutions que nous proposons pour la langue française consistent à réaliser trois types d'opérations de régularisation de la surface des textes :

La décomposition d'amalgames orthographiques
La décomposition d'amalgames morphologiques
L'analyse des mots en /qu-/

L'amalgame orthographique concerne toute forme de surface que l'on peut considérer comme la simple concaténation de plusieurs mots. L'opération de régularisation consiste alors à découper une chaîne de caractères. Ex.

lequel ---------> le + quel
duquel ---------> du + quel

Un amalgame morphologique se distingue de l'amalgame orthographique par le fait que n'y sont pas apparentes ses formes sous-jacentes. Un amalgame morphologique n'est pas segmentable comme un amalgame orthographique.

La procédure de régularisation consiste, alors, à substituer à une forme, une suite de plusieurs formes. Chaque amalgame sera ainsi remplacé par la suite des formes qui lui correspond. Ex.

au ---------> à le
des ---------> de les

d- Les amalgames syntaxiques

En effet, la langue française est aussi riche d'amalgames syntaxiques, c'est-à-dire de formes qui résultent du regroupement de deux ou plusieurs formes primaires, chacune ayant un rôle syntaxique propre. La solution adoptée consiste à se donner un nombre très restreint de catégories syntaxiques, chacune ayant un comportement distributionnel bien défini {V,F,Y,D,P,Q,C,W,T}. Le prétraitement de nature morpho-syntaxique précède brièvement l'analyse morphologique dans le but de détecter, dans les séquences de formes, une propriété syntaxique quelconque. Par exemple l'occurrence de la forme {/ce/ + relatif} est de nature pronominale et non prédéterminative (SIDHOM,2002,P.73).

***3 ième Étape : Définir une méthodologie de la* segmentation du texte**

L'une des premières étapes importantes de l'indexation automatique présentée ici est la segmentation du texte concerné pour l'identification d'éléments textuels utiles à l'indexation. Ces éléments textuels, qui sont répertoriés ultérieurement dans un dictionnaire spécifique, peuvent être des termes, des suites des mots, des phrases, des thèmes, des unités logiques, etc.

Pour découper un texte, il est important de prendre en considération plusieurs caractéristiques spécifiant le document et favorisant l'"éclatement" d'un texte telles que la nature du texte étudié (technique, scientifique, littéraire, etc.), ainsi que le mode d'organisation du discours (narratif, argumentatif, descriptif, etc.) ou encore la structure physique du document (les attributs typographiques, polices, espaces, etc.) (AKRIFED,2000,P.374).

La détermination des unités minimales de traitement, au sein d'un texte correct, est l'une des tâches la plus importante de l'analyseur morphologique qui pourrait être élaborée de divers points de vue. Nous pouvons par exemple mentionner, d'une manière distincte, les critères de "sens" et les critères de "forme". Il n'est peut être pas inutile de souligner ici qu'il y a une distinction nette entre les mots de la langue (mots lexicaux) et les mots du discours qui sont en fait des descripteurs ou les syntagmes nominaux.

Les mots de la langue, en tant qu'ils sont mots de la langue, ne signifient que des propriétés, jamais des entités ; ils signifient des attributs, et non des substances, tant qu'ils ne sont pas mis en œuvre dans le discours (Le Guern,1991,p.23).

Le choix d'une stratégie de segmentation du texte dépend étroitement aux objectifs du traitement. Dans notre analyse MS nous pouvons uniquement segmenter le texte en phrases et en mots, car ils constituent le support nécessaire pour la reconnaissance automatique des SN.

Un texte est en fait un ensemble cohérent de phrases. De point de vue des solutions informatiques, *"le découpage en phrases est une opération simple et réalisable directement sur la chaîne d'entrée si les frontières des phrases sont repérables en dehors de toute analyse morpho-syntaxique : il suffit de parcourir la chaîne de caractères représentant le texte et d'y repérer les marques de fin de phrase (/. / /? / /! / /;/). Le produit de cette opération sur le texte est un ensemble totalement ordonné de phrases"* (Metzger,1988,p.64).

La phrase est constituée par des mots, c'est-à-dire des séquences de caractères formant des unités autonomes susceptibles d'être utilisées dans les diverses combinaisons des énoncés. Le mot est en fait l'unité libre minimale du discours qui dans une perspective informatique correspond à un caractère ou une séquence de caractères inclus entre deux espaces ou entre un espace et une ponctuation.

Le repérage des espaces et des autres signes de ponctuation, qui jouent le rôle de frontière du syntagme au sein des phrases, permet donc de segmenter chaque phrase en mots.

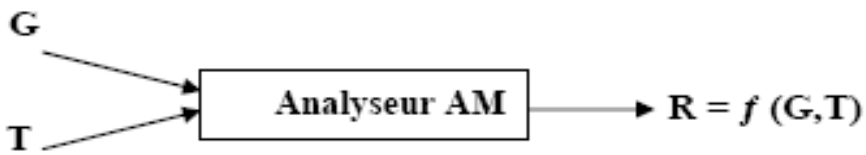
En raison des opérations ultérieures à effectuer sur les mots, chacun de ces mots doit être représenté comme un ensemble de morphèmes. La reconnaissance des morphèmes au sein d'un mot est une opération très importante. Sans cela on ne pourra pas déterminer la catégorie morphologique d'un mot affixé.

3ième Étape : Catégorisation

Lorsque le programme de la segmentation du texte est élaboré, il faudra concevoir un automate à états finis qui essayera d'assigner à chaque unité minimale du texte (mots ou signes de ponctuation) une valeur symbolique, appelée aussi la catégorie morphologique de l'unité. Pour que l'automate, qui est un analyseur morphologique, puisse mettre en œuvre un modèle linguistique, il doit être en consultation quasi permanente avec un ensemble de règles et un lexique spécifique.

L'automate est une sorte de machine théorique qui lit un symbole à la fois et, en fonction du symbole, découpe le mot à analyser en sous-chaînes répertoriées dans le lexique, contrôle l'association de ces sous-chaînes au moyen d'un ensemble de règles de fonctionnement établies sur des classes de sous-chaînes et restitue la ou les classes morphologiques associées au mot analysé, ainsi qu'un certain nombre de variables.

On peut ainsi tenir R, "l'analyse" produite, pour une fonction à deux variables : Le texte T d'une part, et ce que nous appellerons une "**grammaire de référence**", G. G est constituée d'un ensemble d'hypothèses portant sur la langue du texte, et définissant un système de catégories dans lesquelles ses unités peuvent être classées, ainsi qu'un système de traits pertinents qui en déterminent diverses sous-catégories, notamment flexionnelles (BERRENDONNER,83).



En consultant le lexique et les règles de formation des mots, l'analyseur fournit à chaque mot une seule catégorie morphologique respective. Dans certain cas et en fonction du contexte, l'analyseur ne peut pas se décider, il est en fait confronté à une analyse double. Dans ce cas, on dit que l'analyse est ambiguë. Par exemple, le mot "variable" qui est désigné, dans le lexique, comme "nom" et "adjectif" peut entraîner une ambiguïté morphologique dans le discours.

Le lexique contient à la fois les mots, les règles de formation des mots, et surtout les catégories morphologiques de chaque mot donné dans le texte. La définition des catégories morphologiques, qui nécessite une analyse morphologique approfondie, peut être réalisée uniquement par des activités de recherches d'un analyste expert humain dans ce domaine.

Les catégories que nous présentons dans ce travail différent sensiblement de l'organisation de la grammaire traditionnelle enseignée aujourd'hui dans les écoles. Il ne s'agit pas l'intention d'innover à tout prix, il s'agit plutôt d'une adaptation et simplification de l'organisation traditionnelle des parties du discours pour adopter des catégories qui correspondent essentiellement aux applications prévues pour le traitement des syntagmes nominaux. Pour la détermination des catégories lexico-syntaxiques nous sommes inspirés pour l'essentiel par les suggestions de Michel Le Guern et les travaux de Jean Paul Metzger et A. Berrendonner à la suite de nombreuses discussions au sein du groupe SYDO (Système Documentaire) de Lyon pour la langue française.

Liste des catégories :

F : noms-adjectifs
D : prédéterminants
P : prépositions
C : coordonnants
W : adverbes
T : signes de ponctuation
V : verbes
Y : pronoms préverbaux
Q : subordonnants
I : interjections
M : modalisation de phrases

Ajoutons que l'analyse en constituants des syntagmes nominaux nécessite parfois une classification plus fine que celle induite par les «catégories principales» : certaines règles de réécriture ne s'appliquent que pour certains éléments d'une catégorie. Certaines catégories peuvent ainsi être subdivisées en sous-catégories ; sous-catégorisation qui peut être établie, aussi, sur des bases distributionnelles (Voir aussi METZGER,88).
Par exemple :

a- Les sous-catégories de la catégories "F" sont des noms (NOM), des noms-adjectifs (NAN), des adjectifs (ADJ);
b- Les sous-catégories de la catégories "D" sont : DEF, NUM, IND;
c- Les sous-catégories de la catégories "W" sont : AAJ, QUA, PRO, TAM;
d- Les sous-catégories de la catégories "Y" sont : IN1, IN2, INN;

4ième Étape : Constitution d'une base de données textuelle.

Toute application en TAL, est une affaire de mémoire (stockage) et d'intelligence; la mémoire, c'est la base de données; l'intelligence c'est les règles et les programmes.

La mise au point d'un analyseur MS suppose en permanent la consultation de trois organisations référentielles et informatives qui sont intégrées dans une base de données textuelles :

1- Les données (le corpus) et les résultats, en particulier les SN qui sont repérés automatiquement dans le texte à partir d'une analyse MS, sont stockés, organisé, et représentées dans la base de données ;

2- Un ensemble de règles décrivant le fonctionnement de la langue; Il s'agit des catégories syntaxiques ou des règles de réécriture pour la constitution des SN ;

3- Un ensemble de connaissances factuelles : **le lexique**.

Le lexique de l'analyseur fournit les éléments d'information nécessaires dans l'ordre de la morphologie et de la syntaxe. Pour un *item* donné, il indique la catégorie et les valeurs prises par les variables pertinentes, qui ont été classées en variables syntaxiques, variables flexionnelles, et variables lexicales. Les variables lexicales portent pour la plupart sur les contraintes combinatoires, à la frontière entre la syntaxe et la sémantique ; leur prise en compte diminue considérablement le nombre des analyses ambiguës.

Le lexique contient donc :

- L'ensemble des mots constituants du corpus, et des formes minimales du discours qui peuvent être ramener à leurs bases grâce à des règles de flexions;
- Les catégories morphologiques de chaque mot donné dans le texte ;
- Les règles de la formation des mots et des formes canoniques.

Chaque entrée lexicale est le lexème lui-même dont ses arguments représentent les informations linguistiques associées au lexème. Nous rappelons ici que les unités minimales du texte ou du discours sont des formes, tandis que les unités du lexique sont les unités de la langue, les lexèmes.

Quand le texte qu'on veut étudier est entré dans la machine, le système commence à examiner, un à un, tous les mots du texte. Il consulte d'abord une liste des mots grammaticaux et des mots de haute fréquence qui ne nécessitent pas un examen détaillé. La liste des mots est en effet préalablement intégrée à la mémoire de la machine c'est à dire dans un lexique spécifique.

Aujourd'hui, les mémoires d'ordinateur sont suffisamment vastes et les temps d'accès suffisamment court. De ce fait un lexique organisé en entrées, lemmes, profils, doit permettre l'enregistrement de l'ensemble des mots d'un corpus textuel important, tels qu'ils apparaissent dans le discours; La fréquence élevée d'apparition de certains mots ou l'irrégularité de leur formation font qu'il est certainement plus efficace de les faire figurer dans le lexique plutôt que de leur faire subir une analyse morphologique qui les "ramène" à une base ou un ensemble de formants.

5ième Étape : Élaboration des règles de réécriture pour la reconnaissance des SN

1- Reconnaissance des SN

Dans la phase de l'analyse syntaxique, la reconnaissance «des syntagmes nominaux» est au centre des applications de grande envergure, puisqu'elle apparaît comme préalable et indispensable à l'interprétation sémantique des phrases.

Effectivement, la reconnaissance des SN, impose a priori une segmentation correcte et cohérente du texte, en phrases et en formes. Ainsi lorsque les phrases sont segmentées en mots, la catégorisation doit être effectuée d'une manière adaptée et correcte, tout en minimisant les ambiguïtés morphologiques qui en résultent.

Dans une mesure très large, le repérage des SN Simples, étant des séquences "continues" de mots peut s'effectuer généralement par la reconnaissance "des marques" du début (la tête) et de la fin des SN.

La reconnaissance du début des SN ne posera pas en principe de grande difficultés, il suffit de parcourir le segment syntagmatique (phrases, clauses etc.) et de vérifier si le(s) premier(s) mot(s) analysé(s) est(sont) un(des) élément(s) de la tête ou début d'un SN. Les éléments qui jouent le rôle du début de SN ont déjà été répertoriés et rangés dans le lexique comme des mots grammaticaux. Le début d'un SN peut en général regrouper les catégories suivantes : des déterminants, des démonstratifs, des numéraux, des prépositions, des adverbes.

Pour repérer le début d'un SN, nous allons mettre en place un système de règle des combinaisons possibles entre tous les éléments constituant le début d'un SN. Ces règles nous permettront de repérer toutes les séquences de mots qui interviennent entre le début de syntagme et le noyau en position de la tête ou le début du SN.

Dans une représentation syntaxique, tous les éléments d'une phrase ne peuvent pas entrer dans le processus de reconnaissance des SN, certains termes et expressions variées sont moins "informatifs" que d'autres, il faudra donc diminuer avant tout le nombre des structures syntaxiques. De ce fait, il semble possible de supprimer un certain nombre d'éléments "non-informatifs ou parasites" afin de n'obtenir que des phrases simples et de ne retenir que les responsables directes de la formation d'un SN.

Ainsi, dans un premier temps nous allons nous attacher à déterminer tous les éléments inutiles au processus de reconnaissance, ce qui suppose une analyse morphologique des mots qui entourent le SN, puis une analyse syntaxique proprement dite capable de mettre en évidence les groupements de mots responsables directes de la formation des SN, afin de les coder et de pouvoir les comparaître avec des règles de réécriture les concernants. C'est cette étape qui est la plus délicate du programme, car elle demande l'entrée de plusieurs lexiques.

2- Système de règles.

Définir une méthodologie d'analyse et de conception pour élaborer un analyseur MS, c'est, du même coup, choisir une ***grammaire de référence***, c'est à dire un système générateur de **règles** remplissant une tâche centrale à la théorie linguistique. La grammaire de référence est en fait un schéma conceptuel formel pour la spécification de phrases autorisées dans le langage indiquant les règles pour la combinaison de mots dans des phrases et des clauses.

Parmi les différentes catégories des règles utilisées en TAL les règles transformationnelles et les règles de réécriture sont les plus importantes.

- **Les règles de réécriture,** dites aussi les règles syntagmatiques (ou PS, par abréviation de l'anglais phase structure);
- **Les règles transformationnelles**. Une règle dite «transformationnelle» si son applicabilité à une suite dépend, non seulement de la constitution de cette suite, mais de la façon dont cette suite a été dérivée, ce qui n'était le cas pour aucune des autres règles. Les règles transformationnelles sont donc des règles qui n'opèrent pas sur des suites, mais sur des arbres. Ces règles sont appelées aussi des règles "lexicalisation", qui transforment une catégorie lexicale en un mot du lexique (comme : NOM ----> garçon).

Dans notre système de représentations syntaxiques, l'analyse syntaxique est présentée sous forme de règles de réécriture : cette présentation a été choisie en raison de son adaptation au langage informatique utilisé Prolog.

Pour construire une grammaire générative, on cherche un algorithme particulier qui soit capable de générer, c'est-à-dire d'énumérer automatiquement, les représentations syntaxiques associées par l'analyse en constituant chaque phrase grammaticale de la langue. Ce faisant, l'algorithme énumérera les séquences bien formées, en même temps qu'il produira une simulation. Un tel algorithme s'appelle grammaire de réécriture syntagmatique (BERRENDONNER,83).

Une grammaire de réécriture syntagmatique est un algorithme dont les éléments se spécifient ainsi :

G = {N, T, R, X}

N : ensemble des vocabulaires non terminaux.

T : ensemble des vocabulaires terminaux.

R : ensemble des règles de réécriture.

X : symbole initial de N appelé AXIOME.

Un ensemble de règles de réécriture spécifie les relations permises entre des chaînes formées de symboles de V (vocabulaire total : terminaux ou non). Ainsi par exemple, le fait qu'une phrase (P) puisse être composée d'un syntagme nominal (SN) suivi d'un syntagme verbal (SV) sera représenté par une règle de la forme : P ----->SN + SV.

Où le + est le symbole de la concaténation, et la flèche se lit comme une instruction ordonnant de réécrire le symbole de gauche en utilisant les symboles de droite. Nous noterons alors l'ensemble des règles de réécriture (SABAH,88,p.43) :

$R = \{X \text{----->} Y\}$ avec $X, Y \in V^*$ et $X \neq 0$

V* est l'ensemble des chaînes engendrées sur v

En passant plus d'informations comme arguments dans les règles de réécriture syntaxiques, on peut rendre le programme capable d'une interprétation du contexte limité à la phrase. On peut ainsi déterminer le nombre (par le pluriel des noms, des articles et des adverbes) et conserver l'information pour donner des réponses simples ou multiples. Comme les définitions grammaticales sont analysées de façon **récursive**, l'information interprétée en tout point peut être passée plus haut ou plus bas dans la chaîne d'interprétation. On peut aussi qualifier ou quantifier les variables. Les verbes intransitifs peuvent ainsi permettre un contrôle de procédure sans recherche d'objets.

Les règles de réécriture appelées aussi les règles de production sont un moyen de représentation des connaissances qui s'inspire directement de la logique (logique des propositions et logique des prédicats), ces règles permettent de rendre de la structure interne de la phrase et spécifie les relations permises entre des chaînes formées de symboles terminaux et non-terminaux.

Une règle de réécriture peut traduire une relation, une information sémantique ou une action conditionnelle qui contient un granule de connaissance. Ces règles sont indépendantes les unes des autres : la modification d'une règle n'a pas d'effet sur les autres. Cette règle est composée dans certains systèmes par la présence d'un module chargé de vérifier à tout moment la cohérence de l'ensemble des règles. Ce module est lui-même une structure basée sur la connaissance. Les systèmes construits sur cette logique comparent les **faits observés** et **prémisses** de règles (filtre). Un progrès consiste à attribuer un coefficient de vraisemblance aux règles. À titre de comparaison, on note qu'en informatique algorithmique structurée, chaque procédure est un granule. Une règle de réécriture est une expression de la forme :

Si Ⓐ **Alors** Ⓑ

La prémisse "A" exprime les **conditions** d'application de la règle. Elle peut contenir une conjonction de propositions logiques ou de relations. Les hypothèses doivent être vérifiées pour que l'on puisse tirer la **conclusion** "B" qui peut être une action à effectuer ou une assertion à ajoutent dans la base des faits.

3- L'ensemble des règles syntaxiques.

Les règles présentées dans ce travail correspondent uniquement aux syntagmes nominaux simples. Intuitivement, on l'appellera Syntagme Nominal Simple une unité syntagmatique avec prédéterminant, dont le noyau a le rôle syntaxique d'un nom, et n'incluant ni relative, ni incise. Ses règles ont été élaborées, pour une grand part, par l'équipe de SYDO de Lyon.

Syntagmes Nominaux :

[1]	N" -----> N"+N"	Le président Jaurès
[4]	N" -----> D'+N'	Le président
[5']	N" -----> NOM-PRO	lui
[5"]	N" -----> NOM-PRP	Jaurès

Syntagmes Adjectivaux

[6]	A" -----> A'+SP^{n}	conseillé par un banquier
[7]	A" -----> A'	assez dynamique

Expression Nominales :

[8]	N' -----> N+SP^{n}	opposant à la loi
[11]	N' -----> N	ministre

Expressions prédéterminatives :

[12]	D' -----> D-DEF+D-NUM	les trois (candidats)
[13]	D' -----> Prep- "de"+D-DEF	de ces (élection)
[13']	D' -----> W-QUA+Prep-"de"+D-DEF	beaucoup de leur (temps)
[13"]	D' -----> W-QUA+Prep-"de"	peu de (résultat)
[14]	D' -----> D	le

Centres Adjectivaux :

[16]	A' -----> A	gentil
[15]	A' -----> W-AAJ+A	particulièrement fidèle
[15']	A' -----> A+EP	teinté de vert (teinté en vert)

Centres Nominaux :

[17]	N -----> N+EP	chef de gare
[18]	N -----> N+A"	projet très populaire
[19]	N -----> A"	(le) tout petit
[20]	N -----> A'+N	grand sportif

Nominaux :

[21]	N -----> F-NOM	ville
[22]	N -----> F-NAN	fenêtre
[23]	A -----> F-NAN	joli
[24]	A -----> F-ADJ	impartial

Syntagmes prépositionnel :

[28]	SP -----> Prep+N"	chef de la gare

Expansion prépositionnelle

[31]	EP -----> Prep+N'	(chef) de gare

6ième Étape : La conception des solutions informatiques

Dans les étapes qui précédent nous avons essayé de présenter les différentes phases de l'élaboration d'un modèle morpho-syntaxique, c'est à dire un ensemble de concepts et de règles qui permettraient le passage d'un système naturel vers un système automatique. Il s'agit maintenant de prévoir quelques programmes informatiques qui seraient capables de repérer automatiquement des SN d'un texte donné en français.

La conception de tels programmes nécessite une recherche indépendante et approfondie. Dans le cadre de cette recherche nous n'avons pas la possibilité de donner toutes les solutions informatiques en détailles, mais en guise de l'introduction nous présentons ici une liste des programmes qui sont nécessaires pour la réalisation de l'analyseur MS :

- Un programme « interface/user » pour saisir et sauvegarder des données, les questions, et les réponses aux requêtes ;
- Un programme pour la segmentation du texte en phrases et en mots ;

- Un automate à états finis pour assigner à chaque unité minimale du texte une valeur symbolique, appelée aussi la catégorie morphologique de l'unité ;
- Un programme pour la constitution d'une base de données textuelles, qui contient le lexique et les règles de formation des mots et des SN ;
- Un programme pour la réécriture des règles syntaxique en langage informatique (par exemple en Prolog). Ce programme permet en effet la reconnaissance automatique des SN.

Parmi de nombreux langages informatiques, il se trouve que le Prolog, qui a été conçu, à l'origine pour la description de système de réécriture, est mieux adapté pour la reconnaissance des formes grammaticales. Un analyseur syntaxique Prolog peut se présenter sous la forme d'un ensemble de clauses, où chaque clause représente une règle de réécriture. On peut, considérer l'interpréter comme l'analyseur et le programme Prolog comme la grammaire, elle même.

Pour élaborer un analyseur MS en Prolog, il faut d'abord commencer par couper les phrases en composants. On commence par le syntagme nominal et le syntagme verbal, puis il faut poursuivre l'analyse en cherchant les constituants de chaque groupe, noms, adjectifs, prépositions, verbes, et autres éléments. L'opération se fait d'une façon descendante c'est à dire de haut en bas et commence par couper la phrase en deux éléments : SN, SV. Ensuite, les groupes (ou les syntagmes) sont étudiés pour leurs constituants propres.

Au sein de l'analyseur on peut introduire les règles morpho-syntaxiques, dites les règles de réécriture qui permettent le cheminement de l'analyseur pour la reconnaissance automatique des formes morpho-syntaxiques différentes. Les règles de réécriture peuvent être facilement traduites en langage Prolog. Par exemple, pour représenter le SN "les deux chats de la concierge", on peut employer les règles suivantes :

Les règles de réécriture	Les règles en Prolog
N"----->D' + N'	N"(_X,_Y) : -D'(_X,_t) & N'(_t,_Y).
D' -----> D_DEF + D_NUM : "les + deux"	D'(_X,_Y) : -D_DEF(_X,_t) & D_NUM(_t,_Y).
D' -----> D_DEF : "les" .	D'(_X,_Y) : -D_DEF(_X,_Y).
D' -----> D_NUM : "deux"	D'(_X,_Y) : -D_NUM(_X,_Y).
N'-----> N + SP (+SP) : "Chats de la concierge"	N'(_X,_Y) : -N(_X,_t) & SP(_t,_Y).

La clause : N"(_X,_Y) : -D'(_X,_t) & N'(_t,_Y) est une façon de montrer qu'il y a un groupe nominal N" situé entre une expression prédéterminative D' et une expression nominale N'. De même, c'est de montrer que l'expression nominale N', est à son tour située entre un centre nominal N et un syntagme prépositionnel SP etc. Prolog est donc utilisé non seulement pour programmer les analyseurs syntaxique et sémantique, mais aussi pour représenter les connaissances nécessaires à la compréhension.

L'avantage d'un tel système est l'analyse et la reconnaissance de la structure hiérarchique de la phrase aux niveaux de la syntaxe et de la morphologie des composants. Les phrases non reconnues sont rejetées. L'analyse est facilement réalisée et implantée en Turbo Prolog. L'inconvénient majeur tient au nombre important de clauses à définir pour décrire toutes les formes grammaticales possibles et obtenir un système parfait. Ces prédicats de définition constituent un dictionnaire des règles syntaxiques de la langue.

Ainsi, lorsque l'analyse MS est achevée et l'organisation des données lexicales est effectuée, l'implémentation de l'analyseur MS et du base de données en Prolog sera réalisée sous forme de clauses Prolog intégrées directement dans le programme. De ce fait, le programme contient au moins trois types de clauses qui constituent le cœur de l'analyseur :

1- Des clauses représentant des règles de réécriture pour la reconnaissance des SN, ainsi qu'un certain nombre restrictions sur la combinatoire des morphèmes;
2- Des clauses de régularisation qui représentent la grammaire flexionnelle;
3- Des clauses sans prémisses pour les entrées de base de données qui fonctionnent comme l'entrée d'un fichier inversé. Précisons ici qu'une clause sans prémisse exprime un fait, alors qu'une clause suivie d'une suite de prédicats à évaluer énonce une règle d'inférence. Il semble que Prolog, malgré sa structure assez performante pour la constitution du système de règles, ne se prête pas bien pour la gestion de système de bases de données lexicales (SGBD). De ce fait nous proposons de confier la tâche de gestion de données lexicales à d'autre(s) système(s) mieux adapté(s), tels que par exemple un SGBD relationnel (Ex. en ORACLE ou en ACCESS). Dans ce contexte, lorsque l'analyseur est écrit en Prolog, la communication entre le lexique et l'analyseur s'effectuera par une interface permettant l'interprétation et la traduction des instructions provenant d'un sens ou de l'autre .

En règle générale la reconnaissance des syntagmes nominaux ne posera pas de grande difficulté. Il suffit de parcourir la phrase et conformément aux règles de réécriture élaborées et donc à partir d'une analyse syntaxique complète repérer les séquences de mots qui sont susceptibles de se ranger parmi les SN.

Après avoir effectué la reconnaissance automatique de toutes formes des SN au sein du texte analysé, le résultat de recherche sont intégrés dans une base de données textuelles, qui est conçu au préalable pour cette application. Rappelons que dans le cadre de l'indexation automatique ces informations sont les seuls éléments pertinents pour décrire le contenu du document concerné, qui doivent permettre, en principe, de retrouver ultérieurement les informations à une question donnée.

5- Conclusion

Les problèmes du linguistique forment le voyageur. Lorsque l'attention voulait se fixer sur le détail d'une sculpture discrète, l'observateur ne pouvait s'empêcher de regarder la totalité de l'édifice. Consciemment ou non, il regardait tout. Mais a-t-il vu beaucoup de choses?" (DUPONT,83,P.461).

Concevoir un modèle de représentation du contenu de documents, susceptible de généralité dans ses applications, suppose avant tout une réflexion profonde de tous les aspects constitutifs de la langue. Cette réflexion qui serait basée sur l'intérêt linguistique du traitement, permettrait de rendre compte des différentes factorisations et continuités de la langue. La tâche essentielle d'un concepteur en TAL ne se limite pas en fait de donner une solution d'informatique à tout prix ; les outils informatiques sont aujourd'hui en voie de perfectionnement, mais les analyses présentées jusqu'à présent sont encore loin de la perfection.

Ainsi pour achever ces quelques lignes conclusives, il convient ici d'évoquer rapidement quelques remarques, dans la mesure où on veut développer et perfectionner notre analyse sur la représentation du contenu de texte.

5.1- Se situer dans le contexte général

Le recours à une approche structurée, incluant les notions de «général vers particulier» est cruciale. Par conséquent, le processus du TAL peut-être considéré comme une activité d'exploration et de définition des problèmes, activité que mènent, au moyen de conversation et transaction, plusieurs acteurs en interaction dans des situations caractérisées par l'ambiguïté, le conflit et la complexités des domaines d'application.

D'une façon générale, les différents traitements que subit un texte sont étroitement imbriqués les uns avec les autres. De ce fait le traitement des SN doit être situé dans un contexte global. Dans cette perspective, lorsqu'on veut par exemple réaliser l'application informatique de la reconnaissance des SN, au sein d'un corpus de textes constitués de phrases complexes, la première condition nécessaire consiste à intégrer dans le système toutes les règles spécifiques et variées de la phrase qui reflètent la complexité et la variété de la structure morpho-syntaxique du texte. Cela nécessite par ailleurs, une catégorisation correcte des formes morpho-syntaxiques, tout en prenant en compte le contexte et la cohérence des règles de réécriture syntaxiques des SN, SV et des phrases entières.

5.2- Détecter les valeurs référentielles du discours

Dans les pratiques documentaires les plus adaptée, l'indexation automatique doit viser principalement les objets, les référents, et non les signifiés. Il est donc indispensable de détecter, dans le discours, tout objet qui a une valeur référentielle.

Dans cette perspective, il n'est peut être pas inutile de préciser ici que certaines distinctions et oppositions, qui existent entre les concepts fondamentaux de la langue, méritent une attention particulière en indexation automatique. Ces distinctions revêtent d'autant plus d'importance qu'elles se heurtent plus fortement à la conception naïve de certains analyseurs. Voici quelques unes des oppositions les plus importantes :

- les données et les informations (les données deviennent informations lorsqu'elles sont porteuses de sens) ;
- la représentation de forme et la représentation de connaissance ;
- la signification lexicale et la signification textuelle ;
- la reconnaissance de la forme et la compréhension ;
- la forme et le contenu, c'est à dire la forme et le sens ;
- la langue et le discours;
- la reconnaissance et la génération.

5.3- Constituer un système de gestion de données textuelle

La mise en place d'un système de reconnaissance automatique des éléments informatifs de documents pourrait être considérée comme une phase préparatoire à la définition d'un schéma de gestion de données textuelles pour les recherches ultérieures. Il s'agit maintenant de savoir comment peut-on exploiter l'ensemble des SN repérés, dans une perspective de recherche d'information.

En fait, les syntagmes nominaux jouent un rôle central dans un système d'information. En soi, la liste de tous les SN du corpus, accompagnés pour chacun de la liste des références de ses occurrences, est déjà utile. Rien ne nous oblige à définir une base de données qui ne serait construite qu'avec les seuls syntagmes nominaux. Cependant, pour une plus grande efficacité de l'outil d'interrogation, il convient d'associer aussi à cet ensemble des SN, qui sont les éléments du discours, des prédicats étant les éléments de la langue. Il s'agit de définir un schéma de gestion de données dans laquelle vont être représentés les deux ensembles informatifs du document (les SN et les prédicats) et les "liens" qui les unissent entre eux et les associent éventuellement à d'autres entités (textes, unités lexicales...).

La définition d'un tel schéma, n'est pas en fait très facile à réaliser, car elle impose l'élaboration d'un modèle relationnel, qui non seulement doit décrire les entités et les phénomènes attestés dans le corpus, il doit aussi représenter les relations logico-syntaxiques qui structurent l'ensemble des SN et des prédicats. Mais peut-on envisager une telle réalisation tant qu'on n'a pas effectué au préalable une analyse MS correcte et complète d'un texte librement en langue naturelle? Voilà une question qui reste à approfondir.

Références

AKRIFED (Fouzia), 2000. Segmentation automatique des textes, l'exemple du logiciel tropes : bilan et perspectives, in CIFED'2000 (Colloque International Francophone sur l'Ecrit et le Document), sous la direction de Hubert Emptoz et Nicole Vincent, INSA Lyon, Juillet 2000, pp. 373-382.

BERRENDONNER (Alain), 1983. Grammaire pour un analyseur : aspects morphologiques, document du travail du groupe de SYDO Lyon.

BLACHE (Philippe), 2001. Les grammaires de propriétés : des contraintes pour le traitement automatique des langues naturelles, Hermès science, Paris.

BONNET (Alain). 2001. L'intelligence artificielle : promesse et réalité. InterEditions Paris, 2001.

BOUCHÉ (Richard), 1988. Valeur référentielle et langage d'indexation dans les systèmes d'information documentaires, communication faite le 28 Novembre 1988 au Colloque "Archives et Temps Réel", organisé à Lille par le CREDO (Université Lille III), L'ADBS Nord, et les Archives du Nord.

CARRÉ (R), DÉGREMONT (J.F), GROSS (M), PIERREL (J.M), SABAH (G). Langage Humain et Machine, Presses du CNRS Paris 1991.

CHOMSKY (N), 1971. Aspects de la théorie syntaxique, Seuil, Paris 1971

DAL (Georgette), HATHOUT (Nabil), NAMER (Fiammetta), 2004. Morphologie Constructionnelle et Traitement Automatique des Langues : le projet MorTAL, in Temple, M., éditeur, Lexique, Volume 16, Presses Universitaire de Lille, 2004.

DE BRITO (Marcilio). 1991. Réalisation d'un analyseur morpho-syntaxique pour la reconnaissance des Syntagmes Nominaux, utilisation des grammaires affixes. Thèse de Doctorat. Université Claude Bernard Lyon I.

DESCLÉS (J.P), 1990. Langages applicatifs, langues naturelles et cognition, Hermès, Paris.

DUPONT (Pierre), 1983. Éléments logico-sémantiques pour une analyse du français, thèse d'État, Université Lumière Lyon II.

DUPONT (Pierre), 1990. Éléments logico-sémantiques, l'analyse de la proposition, publié chez P. Lang (Sciences pour la communication), Bern 1990.

EYANGO (M). 1985. Lexique interactif pou l'analyse automatique du français. Thèse de troisième cycle à l'Université Claude Bernard Lyon I.

FUCHS C. éd. (1993). Linguistique et traitements automatiques des langues, Paris, Hachette supérieur. P.13-18.

HATON (J.P), 2005. L'Intelligence artificielle, Que-sais-je, PUF, Paris.

GARRIER (Claude), 1991. Maîtrise de l'Intelligence Artificielle, Marabout Alleur, Belgique.

GREVISSE (Maurice), 2001. Le Bon Usage, onzième Édition : Grammaire française avec des remarques sur la langue française d'aujourd'hui, DUCULOT 2001.

LAINÉ (Sylvie), LAROUK (Omar), VIDALENC (Isabelle), 1988. Système d'informations textuelles : L'apport des logiques extensionnelles et intensionnelles, Actes du colloque, 11è conférence internationale sur recherche et développement en information retrieval, organisée par laboratoire I.M.A.G, Grenoble 1988.

LE GUERN (Michel), 1991. Un analyseur morpho-syntaxique pour l'indexation automatique, in revue "Le Français Moderne", N° 1 Juin 1991.

MAHMOUDI (Seyed Mohammad), 2002. Le rôle de la technologie de l'information et de la communication dans le reengineering des systèmes, in "Journal of Management Culture" , Qom Campus of University of Tehran", en persan, N° 11, L'hiver 2002. pp. 171-194.

METZGER (Jean Paul), 1988. Syntagmes nominaux et information textuelle : reconnaissance automatique et représentation, thèse D'État Ès Sciences, Université Claude Bernard, Lyon1.

NÉVÉOL (Aurélie), 2004. Indexation automatique de ressources de santé à l'aide d'un vocabulaire contrôlé, in RECITAL, Fès, Maroc, 19-22 Avril 2004.

OLLIVIER (Daniel), Weinfeld (Michel), 2000. Utilisation de rétroaction et de classifieurs adaptatifs pour améliorer les performances d'un système de lecture de montants littéraux de chèques bancaires, in

CIFED'2000 (Colloque International Francophone sur l'Ecrit et le Document), sous la direction de Hubert Emptoz et Nicole Vincent, INSA Lyon, Juillet 2000, pp. 221-229.

ROUAULT (Jacques) 1987. Linguistique automatique : application documentaires, coll. Sciences pour la communication, Peter Lang, Bern, Francfort, Paris 1987.

SABAH (G), 1990. L'intelligence artificielle et le langage : Représentation des connaissances. Paris, HERMES 1988-1990, tomes 1 et 2.

SIDHOM (Sahbi), 2002. Plate-forme d'analyse morpho-syntaxique pour l'indexation automatique et la recherche d'information : de l'écrit vers la gestion des connaissances, Thèse présentée devant L'Université Claude Bernard – Lyon 1, 11 Mars 2002

SIMON (H. A), 1991. Sciences des systèmes sciences de l'artificiel, traduit de l'anglais par J. L. LE MOIGNE, Paris, Bordas, 1991.

SCHMID (Anne-Marie), 1992, in BULAG (n° 18). Conférence faite au département de linguistique de l'Université de Besançon le 17 Mai 1992.

TOWNSEND (Carl), 1988. Turbo Prolog : applications, Sybex. Paris.

VUILLEMIN (Alain). 1987. Informatique et traitement de l'information en lettres et sciences humaines, Masson Paris 1987