Ecole Nationale [illegible] de Bibliothécaires

Diplôme Supérieur de Bibliothécaire

Université Claude Bernard Lyon I

DESS Informatique Documentaire

Note de synthèse

LES APPLICATIONS DU
TRAITEMENT AUTOMATIQUE DU LANGAGE NATUREL

JM.CHALON
Sous la direction de JP. LARDY
U.R.F.I.S.T. LYON

1991

LES APPLICATIONS DU TRAITEMENT AUTOMATIQUE DU LANGAGE NATUREL

JM. CHALON

RESUME:
Différentes approches du traitement automatique du langage naturel français . Les systèmes d'application sont classés en fonction du degré de complexité de leur analyse linguistique indépendamment du domaine d'application.

DESCRIPTEURS:
Langage naturel; Traitement automatisé; Analyseur langage; Analyse morphologique; Analyse syntaxique; Analyse sémantique; Linguistique appliquée; Francais; Logiciel.

ABSTRACT:
Different approaches of natural french language automated processing. Application systems are classified by the degree of complexity of their linguistic analysis, independently of the application field.

KEYWORDS:
Natural language; Automated processing; Language analyzer; Morphological analysis; Syntactic analysis; Semantic analysis; Applied linguistics; French; Software.

S O M M A I R E

M E T H O D O L O G I E

I PRESENTATION DE LA RECHERCHE

Avant toute recherche, il convient d'en préciser le plus finement possible les orientations et les limites . Notre directeur de recherche a exprimé le besoin de faire le point sur les applications issues des travaux sur le traitement automatique des langues naturelles. L'objectif de notre recherche est donc de recueillir des informations sur les différentes applications du traitement automatique des langues naturelles (T.A.L.N.) et sur les bases théoriques de celles-ci.

Le domaine a été préalablement limité au français écrit. De plus, notre demandeur a écarté les secteurs de la traduction automatique ou assistée par ordinateur (T.A.O.) et des correcteurs orthographiques.

II LA RECHERCHE

Notre sujet, à l'intersection de la linguistique, de l'informatique et des sciences de l'information fera l'objet d'une recherche automatisée complétée par une recherche manuelle.

A La recherche automatisée

1) Choix et description des bases de données

Après consultation du Répertoire des banques de données professionnelles 1990 de l'ADBS, et parmi les bases de données accessibles, nous avons sélectionné pour leur domaine de couverture, Pascal, Francis, Inspec, Social Scisearch et MLA Bibliography dont suivent les caractéristiques.

PASCAL

Produite par le CNRS-INIST (Institut National pour l'Information Scientifique et Technique) depuis 1973, Pascal couvre les domaines des sciences et techniques.

La base est composée d'une partie multidisciplinaire, PASCAL M, et de 11 banques sectorielles. Les données proviennent d'articles de périodiques français et étrangers, de rapports scientifiques, de thèses, de comptes-rendus de congrès.

La mise à jour de la base est mensuelle. Disponible sur QUESTEL ou sur IRS-ESA, entre autres serveurs, PASCAL compte environ 7 500 000 références auxquelles la mise à jour ajoute 430 000 données par an.

FRANCIS

Autre base importante de l'INIST-CNRS, Francis est composée de 20 domaines couvrant essentiellement les sciences humaines. Nous avons choisi d'interroger Francis Sciences du langage dont les données issues d'articles de périodiques (31%), d'ouvrages, de rapports, de comptes-rendus de congrès, de travaux universitaires sont relatives aux principaux domaines de la linguistique, incluant la linguistique appliquée. Environ 60 000 références ont été entrées depuis 1972. La mise à jour de la base est trimestrielle et son volume augmente de 3 200 données par an. Francis est servi par QUESTEL.

INSPEC

Base de données britannique produite depuis 1969 par l'Institution of Electrical Engineers (IEE), Inspec couvre les domaines de la physique et de l'informatique, incluant un secteur sciences et technologies de l'information et de la communication. Les 3500 000 références provenant de la littérature mondiale sont pour 80% des articles de périodiques et pour 15% des actes de congrès. Disponible sur de nombreux serveurs dont DIALOG, Inspec accroit son volume de 220 000 données par an grâce à une mise à jour mensuelle.

MLA BIBLIOGRAPHY

Base de données américaine du Modern Language Association, MLA Bibliography dépouille près de 3 000 périodiques et séries concernant la linguistique et la littérature. Depuis 1966, la base contient environ un million de références. Servie par DIALOG, MLA Bibliography est mise à jour bimensuellement, ce qui augmente son volume de 100 000 données par an.

SOCIAL SCISEARCH

Produite aux Etats Unis par l'Institute for Scientific Information (ISI), cette base couvrant les sciences sociales et humaines concerne la linguistique et la communication. Disponible sur DIALOG, Social Scisearch contient depuis 1972, 2 600 000 références extraites principalement de périodiques. La mise à jour mensuelle en accroît le nombre d'environ 120 000 par an.

2). L'interrogation des bases de données

Nous avons effectué l'interrogation en deux phases.

Une première recherche par QUESTEL pour les bases françaises PASCAL et FRANCIS. Questel permet de rappeler les mêmes équations de recherches en passant d'une base à l'autre.

Une seconde recherche dans SOCIAL SCISEARCH, INSPEC et MLA Bibliography avec la procédure "one search" de DIALOG qui permet l'utilisation des mêmes équations pour les trois bases.

a.) *Equations de recherches:*

PASCAL et FRANCIS

Nous interrogerons sur le basic index en vocabulaire libre. Nous croiserons "langue ou langage naturel" avec "traitement automatique ou informatique" et nous réduirons le domaine au "français".

Base : PASCAL

Qu.	Reponses		
1	67649	SCIENCE INFORMATION/FG	
2	1971	(LANGAGE? OU LANGUE?) 1AV NATURE+/T	
3	32345	AUTOMATISATION? OU INFORMATISATION?	
4	43147	TRAITEMENT? AV (AUTOMATI+/T OU INFORMATI+/T)	
5	74158	3 OU 4	
6	36199	FRANCAIS??	
7	222	1 ET 2 ET 5	
8	18	7 ET 6	<-----
9	165	1 ET 5 ET 6	
10	147	9 SAUF 8	
11	7	10 ET LINGUISTIQUE	<-----

Nous avons sélectionné les 18 et 7 réponses des questions 8 et 11.

Base : FRANCIS

Les questions sont les mêmes avec un nombre de réponses différent.

A la question 11 nous obtenons 14 réponses que nous sélectionnons.

SOCIAL SCISEARCH, INSPEC et MLA BIBLIOGRAPHY

Nous reprenons la même stratégie de recherche avec le vocabulaire anglais correspondant.

File 2 : INSPEC

Set	Items	Description
S1	3912	NATURAL(-)LANGUAGE
S2	1959	AUTOMAT?(W)(TREATMENT? OR PROCESS?)
S3	34	S1 AND S2
S4	21	S3 AND PY>1980 <-----

Nous sélectionnons les 21 réponses de la question S4.

La procédure "one search" de DIALOG nous permet de poser les mêmes questions à Social Scisearch et à MLA Bibliography. Nous obtenons respectivement 10 réponses pour chaque base.

b.) *Résultats*

Nous examinerons la pertinence des résultats obtenus avec chaque base de données et vérifierons la présence éventuelle de doublons.

- Nous avions retenu pour PASCAL 18 et 7 soit 25 références parmi lesquelles 11 sont antérieures à 1980: nous avions omis de préciser la date de publication dans l'équation de recherche. Sur les 14 références restantes, 9 paraissent pertinentes. Les 5 autres références traitent de traduction automatique ou sont strictement théoriques.

- Nous rencontrons le même problème avec FRANCIS où 10 références sur 14 ont une date de publication postérieure à 1980. Parmi celles-ci, 7 semblent pertinentes et différentes des précédentes. Pour ces deux bases nous retenons donc 16 données traitant de théorie et des applications.

- Sur les 21 références proposées par INSPEC, seulement 5 traitent du français dont 2 de l'oral. Nous ne retenons donc que 3 articles.

- Parmi les 20 références des deux autres bases, 2 sont pertinentes et peuvent être retenues. Nous n'obtenons donc que très peu de données (5) par Inspec, MLA et Social Scisearch. Celles-ci sont différentes des premières : il n'y a pas de doublons.

Au total, notre recherche automatisée nous a fourni 21 références qui sont pour la plupart des articles de périodiques. Ce résultat peut paraître peu important mais il s'explique par la caractéristique de notre sujet: il y a dans les revues dépouillées par les bases de données un grand nombre d'articles théoriques sur le traitement automatique des langues naturelles, mais beaucoup moins sur ses applications à l'étude ou commercialisées.

Nous complèterons donc notre recherche manuellement.

B. La recherche manuelle

Devant le petit nombre de références obtenues par l'interrogation des bases de données, nous avons complété notre recherche grâce à la documentation de notre demandeur et aux renseignements et documents obtenus à l'ENSB.

Nous avons consulté d'abord les derniers numéros de Pascal Théma T 205 (publication de la base Pascal sciences de l'information) sans résultat.

Un parcours sérieux de la dernière année de divers périodiques (On Line Review, Soft & Micro, Science et Vie Micro) nous a fourni des informations sur différents produits commerciaux. Nous y avons trouvé deux articles intéressants qui n'étaient pas encore retenus.La consultation du catalogue des logiciels *Bureautique, PAO et gestion documentaire* de CXP et du *Répertoire des produits et services de traitement automatique de la langue française* (x) nous a permis de recenser les diverses applications de traitement automatique des langues naturelles commercialisées.A notre demande, certaines sociétés de "linguisticiel" nous ont adressé les documentations de leurs produits.

III. RESULTATS

Notre recherche automatisée ne nous a donné que peu de références, ce qui s'explique par les caractéristiques de notre sujet. D'une part les bases de données anglaise et américaines n'ont qu'une faible couverture du domaine du français, d'où le peu de résultats satisfaisants pour ces bases. D'autre part, les domaines d'application du traitement automatique des langues naturelles sont assez variés et peuvent dans certains cas se rapprocher de l'intelligence artificielle. Ainsi nous aurions pu compléter notre recherche par l'interrogation de base telle que Artificial Intelligence produite par EIC/Intelligence Inc.

Notre recherche manuelle a permis de compléter l'aspect commercial de ces applications.

Une part importante des articles signalés dans les bases se trouvait sur place:

- dans la documentation du demandeur
- à l'ENSB
- à la Bibliothèque Universitaire

De plus nous avons obtenu de certaines sociétés une documentation sur les applications commercialisées.

S Y N T H E S E

INTRODUCTION

Le traitement automatique du langage naturel écrit a intéressé les chercheurs dès l'apparition de l'ordinateur. Les applications sont multiples (accès à des bases de données, consultation de système experts, dialogues homme/machine...) et d'une grande importance économique: ne parle-t-on pas des "industries de la langue" ?

Nous aborderons d'abord les différentes phases mises en oeuvre pour interpréter un énoncé, puis nous examinerons les diverses applications fonctionnelles et leur degré de traitement.

I. LE TRAITEMENT AUTOMATIQUE DU LANGAGE

Toute entreprise de traitement informatique du langage naturel fait des hypothèses sur ce qu'est le langage. En linguistique, il est convenu de considérer cinq niveaux de la langue écrite.

- Le niveau morphologique permet de reconnaître les mots sous les différentes formes (conjugaison, déclinaison,...).

- Le niveau lexical met en correspondance le mot une fois reconnu avec les informations dont on dispose sur ce mot.

- Le niveau syntaxique rend compte de l'agencement des mots dans une phrase.

- Le niveau sémantique fait correspondre des situations du monde réel aux structures reconnues par le niveau syntaxique.

- Le niveau pragmatique interprète ces situations dans le contexte plus général d'un échange d'informations entre l'auteur et le lecteur.

Chacun de ces niveaux suppose un traitement très spécifique permettant l'idée d'un système modulaire en traitement automatique. L'architecture de la plupart des systèmes existants se borne à une exécution séquentielle des traitements allant du morphologique au pragmatique. Les réalisations ayant une approche intégrée sont plus rares. L'informatique ne rencontre pas de grosses difficultés pour les traitements morphologique et lexical, ce n'est pas le cas pour les niveaux supérieurs (syntaxe, sémantique, pragmatique) pour lesquels de nombreux modèles ont été envisagés.

Ainsi l'interprétation d'un énoncé met généralement en oeuvre plusieurs modules:

- analyseur morphologique et lexical
- analyseur syntaxique et sémantique
- module d'inférence

En fonction de la spécificité de leur domaine d'application et du degré d'interprétation souhaité, les programmes de traitement du langage naturel combinent plus ou moins ces différents modules.

Le traitement morphologique consiste à décomposer les mots en radicaux à partir de la forme représentative stockée dans le lexique. Ce traitement permet de ne pas mentionner toutes les formes que peut prendre un mot dans le lexique, et ainsi de le rendre moins volumineux.

L'analyse lexicale fournit des informations sur le mot:
- informations grammaticales (nature du mot, ses flexions)
- informations sémantico-pragmatiques (traits, synonymes)

Cette analyse permet de lever l'ambiguïté lexicale: certains mots ont un sens différent selon le contexte ("suite","avocat"). La vocation des analyses morphologique et lexicale est de fournir les constituants de base des modules syntaxique et sémantique qui suivent.

Les analyseurs syntaxiques utilisent diverses méthodes pour rechercher les différentes façons de regrouper les mots. Certains procèdent du haut vers le bas: ils cherchent des phrases vraisemblables dès le début de l'analyse. D'autres procèdent du bas vers le haut en essayant les différentes combinaisons locales de mots et faisant marche arrière chaque fois qu' une combinaison de mots est inacceptable. Certains analyseurs utilisent des formalismes: les " réseaux de transition augmentés " expriment la structure des phrases et des groupes de mots sous forme d'une suite explicite de changements d'état à suivre ; les " grammaires fonctionnelles lexicales " établissent pour chaque phrase une structure fonctionnelle dans laquelle les fonctions grammaticales sont étroitement associées aux mots qui les remplissent.

Les analyses ainsi obtenues constituent les données de la quatrième partie du programme: les analyseurs sémantiques. Ils transforment la forme syntaxique en une forme "logique". Ces analyseurs peuvent utiliser le formalisme du calcul des prédicats.

Enfin la dernière étape de compréhension des langues est l'analyse pragmatique c'est à dire l'analyse du contexte; en effet toute phrase est prise dans un réseau de correspondances: à un moment donné par une personne donnée. Cette analyse essaie de lever les ambiguïtés contextuelles par un codage des connaissances du monde environnant permettant de tirer des inférences.

Nous allons voir un exemple de succession des analyses que pourrait mettre en oeuvre un programme de compréhension du langage naturel (Winograd, 1984).

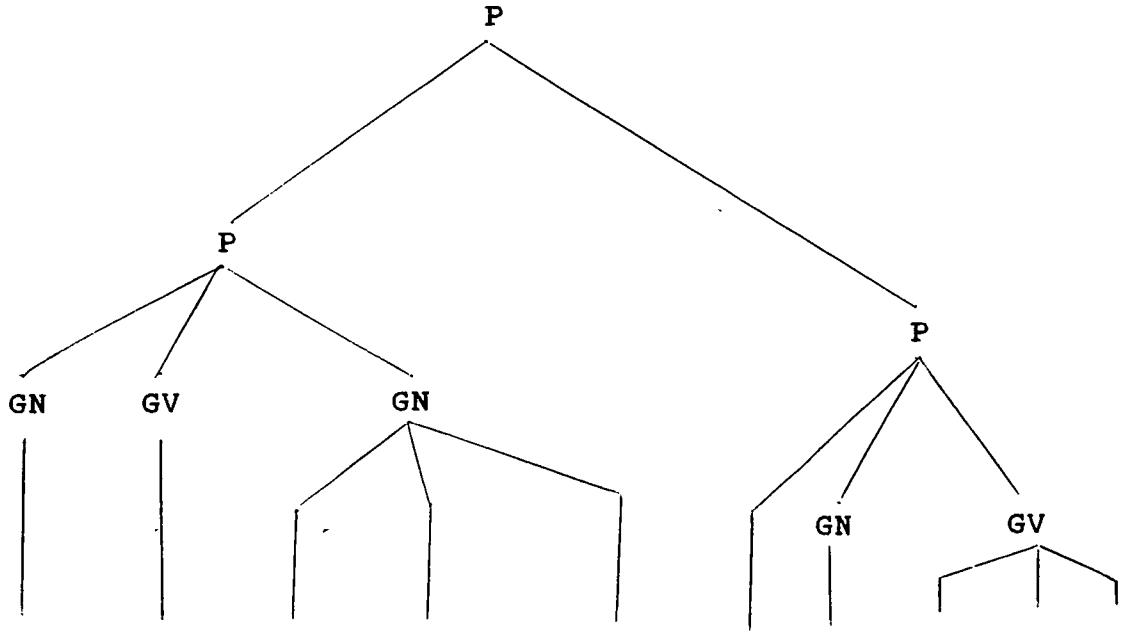
La phrase analysée est : *Elle apprécie les brillants écrivains que nous avons vus hier.*

Le résultat des deux premières analyses est une liste des mots de la phrase accompagnés de leur radical, de leur catégorie lexicale et de leurs traits.

Analyse morphologique et lexical

Mot	Racine	Catégorie	Traits
Elle		pronom	3e pers.féminin singulier
apprécie	apprécier	verbe transitif	présent 3e pers.singulier
les	le	déterminant	défini pluriel
brillants	brillant nom	adjectif	masculin plur. concret masc. plur.
écrivains	écrivain	nom	concret masc. plur.
que		pronom relatif	
nous		pronom	1e pers.plur.
avons	avoir avoir	verbe auxil. verbe transitif	1e pers. plur. présent 1e p. plur.
vus	voir	verbe transitif	participe masc. plur.
hier		adverbe	

Ces données sont soumises à *l'analyseur syntaxique* qui représente la structure de la phrase.



Elle apprécie les brillants écrivains que nous avons vus hier.

L'analyse s'achève par la conversion de la structure syntaxique de la phrase en une forme permettant de tirer des inférences. Notre conversion se fera par calcul de prédicats: le module d'*analyse sémantique* du programme hypothétique représente le contenu logique de la phrase par des symboles du type: *x* est écrivain. Nous aurons ainsi: écrivain (x), brillant (y), voir (z,y,t2), apprécier (y,x,t0), après (t2,t0).

Finalement, le module d'*analyse pragmatique* indique tout ce qui est connu au sujet des variables *x, y, z, t0, t2.*

- x= variable quantifiée
- y= identité déterminée par le contexte
- z= locuteur et d'autres personnes non spécifiées
- t0= moment de l'énoncé du texte
- t2= moment passé décrit par "hier"

La variable x, par exemple est "quantifiée", c'est à dire qu'elle affirme l'existence de quelque chose et identifie des objets particuliers. Le groupe "les écrivains" est défini, en revanche, z reste ambigu car il représente le pronom ambigu "nous".

II. APPLICATIONS DU TRAITEMENT AUTOMATIQUE DU LANGAGE

Face à la masse croissante d'information de toute nature à traiter et pour faciliter le dialogue homme/machine, les équipes de recherche et les "industries de la langue " développent des applications intégrant les techniques de traitement automatique du langage naturel. Les principaux domaines sont la traduction, le classement et la recherche d'information ainsi que le dialogue homme/machine. Comme nous l'avons déjà précisé nous nous attacherons à la langue écrite et nous ne traiterons pas le domaine de la traduction.

Les différents systèmes font appel à un type de traitement plus ou moins complexe généralement en fonction de l'étendue de leur domaine d'application. Nous essaierons de dégager plusieurs niveaux de systèmes en fonction de leur degré de "raisonnement" qui permettra de lever les ambiguïtés du langage telles que:

- ambiguïté typographique (le caractère "E" peut s'interpréter comme "e","é","è","ê","ë")
- ambiguïté morpho-lexicale ("micro ordinateur" peut représenter "micro" et "ordinateur" ou "micro-ordinateur")
- ambiguïté morpho-syntaxique ("président" peut être un nom au masculin ou un verbe conjugué)
- ambiguïté de ponctuation (le point peut marquer une fin de phrase ou une abréviation en milieu de phrase)
- ambiguïté contextuelle telle que: " la petite brise la glace" (--> elle a froid *ou* --> elle va se couper) ou encore " il veut épouser une Danoise qui parle italien" (= la personne qu'il a l'intention d'épouser est danoise et parle italien *ou* = la personne qu'il a l'intention d'épouser devra être danoise et parler italien) (Menon, 1988).

Nous traiterons donc les différentes applications non pas par domaine mais en fonction de la "puissance" de leur traitement que nous classerons en trois niveaux:

- Niveau 1: traitement statistique et lexical
- Niveau 2: traitement morpho-lexical syntaxique
- Niveau 3: traitement morpho-lexical syntaxique et sémantique. Bien entendu, cette classification n'est pas rigide: les limites entre les différents niveaux sont floues. Elle tente seulement de refléter les degrés de "compétence" des divers systèmes.

1.)Les Applications de Niveau 1

Le type de traitement de ces applications est uniquement statistique et lexical. Certains programmes d'analyse de question sont basés uniquement sur un traitement statistique des réponses libres (Lebart, 1982). A partir d'une lecture lettre à lettre, ce type de système va reconnaître puis compter, classer les formes lexicales et les placer sur un graphique en fonction de leur fréquence. Ces graphiques sont soumis aux interprétations de l'utilisateur. Ce type d'analyse utilisé en sociologie peut-il être réellement qualifié de traitement automatique du langage ?

Le traitement statistique peut être associé à un traitement lexical. C'est le cas de LEXINET, logiciel d'indexation semi-automatique (Chartron, 1987). Le système ne dispose pas de dictionnaire de référence au départ mais le construit au fur et à mesure avec l'aide d'un expert. Après un premier découpage, le critère statistique de la variance permet le filtrage des unitermes: le vocabulaire général de variance faible est rejeté dans un antilexique alors que les termes spécifiques de variance forte, sont soumis à l'intervention humaine pour ne conserver que les concepts significatifs. Cependant l'absence de traitement morpho-syntaxique ne permet pas de lever les ambiguïtés de ce type ("acoustique": adjectif ou nom) pour les termes proposés. Une comparaison entre l'indexation manuelle et l'indexation semi-automatique effectuée avec LEXINET sur un même corpus de documents met en évidence les insuffisances et les avantages de chacune mais ne permet pas de trancher (Chartron, 1989).

Le logiciel de dépouillement d'enquêtes QUESTION est aussi basé sur un traitement statistique et lexical. Ce logiciel analyse les réponses aux questions ouvertes à partir de mots clés obtenus par comptage de la fréquence des mots. De plus des matrices d'occurrences simultanées permettent de repérer les mots apparaissant en même temps et ainsi d'effectuer une analyse du contenu à partir de groupes de mots (Jager, 1989).

Ces système de niveau 1 peuvent être satisfaisants dans un secteur précis mais leur limite se fait très vite sentir dès lors qu'on attend un traitement plus profond du langage permettant de lever les ambiguïtés morphosyntaxiques.

2.)Les applications de niveau 2

Les systèmes de niveau 2 ont des analyseurs syntaxiques qui traitent les niveaux morphologique et syntaxique .

Logiciel de recherche documentaire et d'indexation automatique, DARWIN dispose d'un module de traitement morpho-syntaxique. Ce logiciel n'utilise pas de thésaurus, ce qui permet de traiter des sujets nouveaux. Le système utilise quelques connaissances morphologiques de base et des schémas syntaxiques pour attribuer à chaque mot du texte une catégorie grammaticale vraisemblable, autorisant par la suite une analyse syntaxique de la phrase. Ces règles syntaxiques permettent d'isoler les éléments de signification sans utiliser de dictionnaire (Vergne, 1987). Lorsque de nouvelles expressions apparaissent, elles sont immédiatement reconnues. La recherche documentaire s'effectue à partir d'un mot ou d'une expression. DARWIN y répond en deux temps: d'abord, il donne une liste des expressions contenues dans le fonds et proche de la question posée; cette liste est acceptée ou modifiée par l'utilisateur; ensuite, après envoi de la liste des expressions acceptées il propose une liste des textes du fonds documentaire.DARWIN ne permet donc pas une réelle recherche en langage naturel.

De même d'autres logiciels tels que ACCENTS, système d'accentuation automatique de textes français ou LYDIE, analyseur de communication sont basés sur analyse morpho-syntaxique. TEX-NAT est un système d'analyse de texte développé pour TEXTO (Lancel, 1988). Il dispose de traitement lexical et syntaxique. A partir du dictionnaire de base de 50000 termes et d'un vocabulaire technique par domaine, le système effectue l'indexation en n'examinant que les désignés comme significatifs. Chaque terme est assorti de sa fonction dans l'indexation (mot clé, mot vide, mot en attente) et facultativement de son sens. De chaque mot sont extraits les segments non liés aux flexions nominales et verbales. Le dictionnaire est enrichi en cours de processus d'indexation.

Plus complet, SPIRIT est un système d'indexation et de recherche d'information textuelle. Ce logiciel dispose d'un traitement morphologique et syntaxique associé à un traitement statistique. L'analyseur linguistique a les caractéristiques suivantes: - le système est indépendant du domaine et peut analyser des bases de données nouvelles sans nécessiter un travail important.

- il est modulaire: chaque module traite un aspect de l'analyse linguistique.

Le système a pour but de lever les ambiguïtés morpho-syntaxiques et les problèmes de ponctuation (Fluhr, 1983). L'analyse morphologique est réalisée par consultation d'un dictionnaire de 450 000 entrées. Les mots vides sont éliminés. Un autre lexique d'expressions idiomatiques d'environ 1500 entrées permet au système de reconnaître comme une seule entité les locutions telles que "à concurrence de" ou "mettre en oeuvre". L'analyse syntaxique lève de nombreuses ambiguïtés et reconnaît les mots composés. Les mots retenus sont normalisés sous une forme de base du dictionnaire (livre/nom --> livre, livre/verbe ---> livrer).

A chaque forme de mot correspond d'une part la liste de ses propriétés et d'autre part la liste des mots accentués et désaccentués correspondants: les formes fléchies (pluriel, féminin, formes conjuguées) sont mises en relation avec le mot vedette ou lemme. Toute occurrence d'une forme fléchie est systématiquement rapportée à une occurrence du lemme. Par exemple on stockera dans le dictionnaire à la fois *animal* et *animaux*, en indiquant que l'on a affaire à deux formes du même mot.

Pour présenter les documents les plus pertinents, le système calcule une proximité sémantique entre la requête et les textes basée à la fois sur des critères linguistiques et statistiques. Par ces traitements, SPIRIT peut identifier les mots clés des questions posées en langue naturelle. Ceci offre aussi la possibilité de poser tout ou une partie de document comme question. Par ces aspects, ce système est à la limite du niveau suivant qui en plus essaie de prendre en compte les ambiguïtés sémantiques et contextuelles.

3.)Les applications de niveau 3

Ces systèmes cherchent à lever toutes les ambiguïtés du langage naturel grâce à une analyse sémantique. Ils exploitent un savoir lié aux mots qui permet de les organiser en fonction de leurs contenus. Plusieurs systèmes essayent de répondre à ces exigences.

SAPHIR est une interface d'interrogation de bases de données en langue naturelle (Normier, 1982). Ce système contient plusieurs modules de traitement. Pour l'analyse morphologique, il dispose d'un dictionnaire en deux parties: 4000 mots standard, et un ensemble de mots et descriptions spécifiques à l'application.

L'analyse syntaxique est de type Réseau de Transition Augmenté ou ATN proposés par Woods (1970) qui se présente sous la forme de graphes récursifs. Chaque mot rencontré est stocké dans une variable dont la nature sera précisée : le lien d'appartenance ou de localisation est défini en fonction de la variable préposition et de la nature du concept de la structure partielle à raccrocher.

Ainsi *Le grand livre bleu de Pierre* sera analysé de la façon suivante.

Au départ, les variables sont initialisées à l'état 0. Le premier mot à analyser déclenche le stockage qui donne: DETerminant=*le*; NATureDETerminant=article.

L'analyse des mots suivants, jusqu'à *de* compris nous donne un résultat partiel: ListeADJectif = *grand, bleu*; Liste des TRAITS = TAILLE, COULEUR; CONCEPT = *livre*; NATURE = objet; PREP = *de*.

Un appel récursif nous donne la structure partielle associée à la fin de la phrase: CONCEPT = *Pierre*, NATURE = *personne*, QUANTITE = 1.

La définition de la liaison (appartenance ou localisation) s'exécute en examinant le lexique dans lequel on aurait recensé les différents usages concernant l'emploi des préposition pour chacun des concepts (Debili, 1982), on trouverait que la séquence CONCEPT1 et CONCEPT2 se traduit dans le cas où la nature de ces objets est respectivement objet et personne, par une relation d'appartenance ou d'origine. Le processus d'analyse syntaxique nous donne finalement:CONCEPT = livre NATURE = objet
TAILLE = grand
COULEUR = bleu
QUANTITE = 1
APPARTENANCE ou ORIGINE: CONCEPT = Pierre
NATURE = personne
QUANTITE = 1

L'ambiguïté entre appartenance et origine nécessite des connaissances sur l'objet et la personne pour être levée.

L'analyseur sémantique de SAPHIR lève les ambiguïtés. De plus, un dialogueur permet de recourir au choix de l'utilisateur dans le cas où le système dispose de plusieurs interprétations probables.

Ce type d'application a donc pour but de lever le maximum de difficultés du langage naturel en prévoyant de:

- faire correspondre au mot traité un faisceau de connaissance

- réduire le bruit engendré par les phénomènes de polysémie, en permettant de privilégier telle interprétation d'un mot ou d'un groupe de mots en fonction d'une appréciation sémantique du contexte immédiat. Il s'agit de caractériser le contexte d'apparition des éléments polysémiques pour que chaque combinaison reconnue oriente vers une interprétation probable. Ainsi pour les mots *administration* et *administrer* on peut distinguer les contextes différents:

objet = médicament ---> ***soins médicaux***
objet = entreprise ---> ***conseil d'administration***
objet = ministère ---> ***administration centrale***
objet =unité géographique---> ***administration locale***

D'autres systèmes appartiennent à ce niveau. ALETH est un noyau logiciel d'analyse automatique de texte qui peut être utilisé pour des applications spécifiques. Des modules morpho-syntaxiques et sémantique permettent d'analyser les textes par extraction automatique des termes descripteurs à partir d'une base de connaissance.

A partir de cet outil sont développés des systèmes d'analyse de contenu et de dialogue naturel à l'intention de l'utilisateur final: des applications documentaires et en particulier une application télématique avec le minitel "intelligent" (Siri, 1986).

ANAGOGE et HIERARCHIE sont des systèmes d'analyse de textes comprenant un traitement syntaxico-sémantique. Les domaines d'application sont l'estimation de l'efficacité d'un texte du point de vue communicatif pour le premier et l'alimentation de bases de données pour le deuxième.

Enfin, outil d'aide à la création de textes promotionnels, BRAIN BOOSTER dispose de traitements lexical, syntaxique, sémantique et phonétique. Il fonctionne comme un dictionnaire de synonymes, de rimes, d'expressions, et peut être interrogé par critère sémantique et phonétique. Le dictionnaire de 35000 mots est complété par un système de champs sémantiques. Les expressions sont obtenues par repérage phonétique (nombre de syllabes, contenance phonétique, consonnes sonorisables) et par accès aux champs sémantiques. La création de termes s'effectue par combinaison de graphes satisfaisant à des critères sémantiques et sonores.

Nous avons donc vu qu'un certain nombre de systèmes de niveau 3, aux performances assez inégales intéressent des domaines d'application variés. Le problème majeur reste d'optimiser la compréhension du langage grâce à des modules pragmatiques disposant de représentation des connaissances.

CONCLUSION

Le traitement automatique des langues naturelles a fait l'objet de nombreuses recherches dont les résultats n'ont pas toujours abouti à des applications. Cependant, les enjeux économiques sont importants et l'emploi du terme "industrie de la langue" se généralise avec l'apparition de produits commercialisés.

Nous avons vu que sont réalisés avec fiabilité des systèmes ayant des performances utilisables dans de nombreux domaines où la sémantique est restreinte. Cependant la problématique majeure de la compréhension du langage naturel reste la représentation des connaissances. Ainsi les systèmes actuellement les plus performants intègrent une certaine capacité de raisonnement proche de l'Intelligence Artificielle. Il n'est donc pas impossible d'espérer l'élaboration et la réalisation de systèmes de traitement automatique du langage, certes complexes, mais incorporant suffisamment de connaissances pour être des outils valables dans de nombreux domaines d'application.

B I B L I O G R A P H I E

ARFI, J. 1989. Premiers pas des services vidéothèques intelligents. *Télématique Magazine*, 1989, n° 34.

ARSAC, J., FLUHR, C. 1983. Development of documentary software. *INFODIAL. 2nd International Congress and Exhibition on Data Bases and Data Banks, 1983*, p. 276-281.

BOUCHE, R., GERMAIN N. 1991. Bibliométrie, infométrie et analyse automatique de documents écrits. *Société française de bibliométrie appliquée. Congrès, 1991.*

CHARTRON, G. 1987. Le traitement de l'information, le logiciel LEXINET dans une chaîne de contrôle de flux. *Société française de bibliométrie appliquée. Congrès, 1987.*

CHARTRON, G., et al. 1989. Indexation manuelle et indexation automatique: dépasser les oppositions. *Documentaliste*, 1989, vol. 26, n° 4-5.

COULON, D. 1985. Informatique et langage naturel: présentation générale des méthodes d'interprétation des textes écrits. *Technique et Science Informatique*, 1986, vol. 5, n° 2, p. 103-128.

DEBILI, F. 1986. *Analyse syntaxico-sémantique fondée sur une acquisition automatique de relations lexicales sémantiques.* Thèse, Paris 11.

FLUHR, C. 1983. Le traitement et l'interrogation des bases de données textuelles avec le logiciel SPIRIT. *Bulletin du Centre des hautes études internationales d'informatique documentaire,* 1983, n° 17.

GROSS, M. 1989 Relation entre le sens et la forme. *Bulletin du Centre des hautes études internationales d'informatique documentaire,* 1989, n° 34, p. 9-15.

JAGER, O. 1989. QUESTION: un logiciel de dépouillement d'enquêtes bien adapté au traitement des questions ouvertes et des opérations de marketing téléphonique. *Soft & Micro,* Novembre 1989.

JOHNSON, T. 1985. *Natural Language Computing : The Commercial Applications.* London, Ovum Ltd.

JAYEZ, J.H. 1980. Un survol des recherches sur le traitement automatique du langage naturel. *Linguisticae Investigationes. Revue internationale de linguistique française et de linguistique générale,* 1980, vol. 4, n° 1, p. 39-109.

LANCEL, JM. 1988. TEX-NAT: a tool for indexing and information retrieval. *RIAO 88 (Recherche d'Information Assistée par Ordinateur). Conférence 21 mars 1988,* vol.1, p.369-378.

LAPORTE, E. 1988. La reconnaissance des expressions figées lors de l'analyse automatique. *Langages*, 1988, vol. 23, n°90, p. 117-126.

LEBART, L. 1982. L'analyse statistique des réponses libres dans les enquêtes socio économiques. *Consommation*, 1982, n°1.

LEBART, L. 1986. Analyse statistique des réponses libres dans les enquêtes par sondage. *Revue française du marketing*, 1986, n° 109.

LICHNEROWICZ, A. 1989. *Etude sur l'état actuel de l'évolution des systèmes d'information non structurés*. Ministère de la Recherche et de la Technologie.

MENON, B. 1988. Indexation automatique et intelligence artificielle: quelques questions de stratégie. *Image et intelligence artificielle dans l'information scientifique et technique, INRIA, 1988*, p. 143-175.

METZGER, J.P. 1988. *Syntagmes nominaux et informations textuelles: reconnaissance automatique et représentation*. Thèse d'état, LYON 1.

NORMIER, B., et al. 1982. SAPHIR, système d'interrogation de bases de données relationnelles en langage naturel. *Linx: Revue internationale des applications automatiques du langage*, 1982, n°2.

NORMIER, B. 1987. Vous avez dit: " langage naturel " ? *La lettre de l'intelligence artificielle*, 1987, n° 25.

PITRAT, J. 1981. *Réalisation d'un analyseur-générateur lexicographique général*. Rapport 79-2 du GR22 du CNRS.

PITRAT, J. 1980. Les programmes qui "comprennent" le langage naturel. *Linguisticae Investigationes. Revue Internationale de Linguistique Française et de Linguistique Générale*, 1980, vol. 4, n° 2, p. 395-414.

POGNAN, P. 1984. Vers une "compréhension" automatique des textes scientifiques. Applications éventuelles en documentation. *Brises*, 1984, n° 4.

RADY, M. 1983. *L'ambiguïté du langage naturel est-elle la source du non-déterminisme des procédures de traitement ?* Thèse, PARIS 6.

RASTIER, F. 1987. Entretien sur la sémantique et l'I.A. *Langages*, 1987, n° 22, p. 123-128.

Répertoire des produits et services de traitement automatique de la langue française. Observatoire des industries de la langue. Daicadif, 1989.

SAPHIR: le langage pour SQL et Db2. *Logiciels & services*, 1987, n° 64, p. 37-39.

SALKOFF, M. 1988. Analyse automatique du français. *Traitement des langues naturelles. Congrès 4 juillet 1988*, p. 217-237.

SEYDEN, E., LAUNET, E. 1988. Les artisans de l'informatique linguistique. *Science et technologie*, 1988, n° 1.

SIRI, N. 1986. Minitel guide, l'annuaire intelligent. *La Revue du Minitel*, 1986, n° 8.

VASSEUR, F. 1989. Un système expert pour consulter les annonces du "Monde". *Videotex et RNIS Magazine*, 1989, n°46.

VERGNE, J. 1987. Une méthode structurelle de la reconnaissance des formes pour l'analyse morpho-syntaxique du français sans dictionnaire. *Reconnaissance des formes et intelligence artificielle: 6e congrès AFCET INRIA,1987.* Dunod, tome 2, p. 933-941.

WINOGRAD, T. 1984. Les logiciels de traitement des langues naturelles. *Pour la science*, 1984, n° 85, p. 90-103.

WOODS, W.A. 1970. Transition network grammars for natural language analysis. C. ACM, 1970, vol. 13, n° 10, p. 591-606.

ZARRI, G.P. 1988. Conceptual representation for knowledge bases and intelligent information retrieval system. *Proceedings of the Eleventh ACM International Conference on Research and Development in Information Retrieval.* Presse Universitaire de Grenoble.

A N N E X E I

LOGICIELS DE TRAITEMENT AUTOMATIQUE DU LANGAGE

NOM	TRAITEMENT	DOMAINE	DISTRIBUTEUR
ACCENTS	lexical syntaxique	accentuation des textes	Edit Inc.
ALETH	morphologique syntaxique sémantique	analyse automatique de textes	GSI ERLI
ANAGOGE	syntaxico- sémantique	analyseur de textes	LIDIA S.A.
BRAIN BOOSTER 2	lexical syntaxique sémantique	aide à la création de textes promotionnels	KAOS S.A.
DARWIN 3	syntaxique	indexation automatique recherche documentaire	CORA
HIERARCHIE	syntaxico- sémantique	analyseur de textes (B.D.)	LIDIA S.A.
INFLUX	lexical sémantique thésaurus	gestion documentaire	DATAWARE
LYDIE 2	syntaxique	analyseur de textes	CORA
MACCAO	syntaxique sémantique phonétique	analyseur de texte pour générateur de système d'EAO	SECIA
MICRO MIND	syntaxique statistique	documentation automatique	SCERGIE
QUALITATIVE	syntaxico- sémantique	traitement d'enquête qualitative	LIDIA S.A.

NOM	TRAITEMENT	DOMAINE	DISTRIBUTEUR
QUESTION	lexical statistique	traitement d'enquête analyse de données	STATILOGIE
SAOR	lexical syntaxique sémantique	orientation automatique de dossiers de retraite	COGNITECH
SAPHIR	morphologique syntaxique sémantique	interrogation de bases de données relationnelles	INTELLIGENT SYSTEMS
SELECTEUR DE COMPETENCE 2	syntaxique	analyse de curriculum vitae	CORA
SPIRIT	morphologique syntaxique statistique	gestion et recherche documentaire	SYSTEX
TEX-NAT	lexical syntaxique sémantique	analyse de texte (TEXTO)	CHEMDATA

A N N E X E II

LISTE DES DISTRIBUTEURS

DISTRIBUTEUR	LOGICIEL	ADRESSE
CHEMDATA	TEX-NAT	17, Quai Gillet 69316 LYON CEDEX 04
COGNITECH	SAOR	167, Rue du Chevaleret 75013 PARIS
CORA	DARWIN LYDIE SELECTEUR DE COMPETENCE	93, Avenue de Fontainebleau 94270 LE KREMLIN BICETRE
DATAWARE	INFLUX	95, Bd Sébastopol 75002 PARIS
EDIT INC.	ACCENTS	1253, Av. McGill Colloge H3B 2Y5 MONTREAL
GSI ERLI	ALETH	1, Place des Marseillais 94227 CHARENTON le PONT
INTELLIGENT SYSTEMS	SAPHIR	147, Avenue Ch. de Gaulle 92200 NEUILLY/SEINE
KAOS S.A.	BRAIN BOOSTER	87, Rue Voltaire 92800 PUTEAUX
LIDIA S.A.	QUALITATIVE ANAGOGE HIERARCHIE	6, Rue Jeanne d'Arc 45000 ORLEANS
SCERGIE	MICRO MIND	8, Rue de Saintonge 75003 PARIS
SECIA	MACCAO	67, Rue Archereau 75019 PARIS
STATILOGIE	QUESTION	41, Rue d'Alleray 75015 PARIS
SYSTEX	SPIRIT	Ferme du Moulon 91190 GIF SUR YVETTE