

Analyse du Langage Naturel à partir de Connaissances acquises par Apprentissage automatique

Boulakradeche Mohamed
Université Badji Mokhtar Annaba
Faculté des Sciences
Département d'Informatique

Résumé : Le but de ce travail est l'analyse automatique du langage naturel. Aussi nous avons commencé notre rapport par un rappel des mécanismes élémentaires d'AALN, puis d'une présentation de l'architecture du système proposé ainsi que des systèmes d'apprentissage en général. Enfin nous terminons par l'implantation du système d'apprentissage choisi dans notre système d'AALN.

Mots clés : traitement du langage naturel, arbre syntaxique, système expert, acquisition de connaissance, induction automatique, apprentissage.

1. INTRODUCTION

Dans ce rapport nous nous intéresserons à la description d'un système d'analyse automatique du langage naturel.

En dotant un système d'une interface langage naturel, on peut en accroître sensiblement la convivialité et l'efficacité. En général, il est plus facile d'exprimer des énoncés en langage naturel que dans un langage déclaratif spécialisé. Par exemple, dans un système d'interrogation de bases de données, l'énoncé « Quels sont les enseignants de l'institut d'informatique ? » est plus simple et plus concis que la requête correspondante en SQL. De plus, le langage naturel est notre moyen de communication habituel, son utilisation permet d'éviter l'apprentissage d'un langage technique.

Dans un premier temps nous proposons la présentation d'une méthode illustrant quelques mécanismes élémentaires d'analyse automatique du langage naturel. Puis, après une présentation de l'architecture de notre système d'Analyse Automatique du Langage Naturel (AALN), nous parlerons des systèmes d'apprentissage en général. Enfin nous terminons par la description de notre système d'AALN à partir de connaissances acquises par apprentissage automatique

2. MECANISMES ELEMENTAIRES D'AALN

La première étape du traitement du langage consiste à identifier la catégorie lexicale de chaque mot. Les connaissances concernant la nature des mots figurent dans un lexique ou dictionnaire. Celui-ci contient en outre des connaissances morphologiques permettant d'identifier les diverses variations d'un même mot, notamment de savoir que 'à mordu' correspond au verbe 'mordre', conjugué à la troisième personne du passé composé et à la voix active. Pour traiter la phrase « Le chien de ma voisine a mordu Ali » les connaissances suivantes sont nécessaires:

Mot	Nature	Connaissances	Morphologiques
le	article défini	masculin singulier	le
de	préposition	de	
ma	adjectif possessif	1 ^{re} s., féminin singulier	mon
chien	nom commun	masculin singulier	chien
voisine	nom commun	féminin singulier	voisin
Ali	nom propre	masculin singulier	Ali
a mordu	verbe transitif	3 ^e p.s, passé comp., voix active	mordre

La deuxième étape consiste à identifier des structures syntaxiques. Ces structures sont décrites en termes de catégories lexicales et permettent de décrire une grande variété de phrases. Voici quelques règles qui définissent une grammaire élémentaire:

- 1) article + nom commun => groupe nominal
- 2) adjectif possessif + nom commun => groupe nominal
- 3) nom propre => groupe nominal
- 4) article + nom commun + prép. + groupe nominal => groupe nominal
- 5) verbe transitif + groupe nominal => groupe verbal
- 6) verbe intransitif => groupe verbal
- 7) groupe nominal + groupe verbal => phrase

La règle 1 peut se lire comme suit "un article puis un nom commun forment un groupe nominal". La règle 2 reconnaît 'ma voisine' comme groupe nominal. La règle 3 identifie 'Ali' comme étant un groupe nominal. Par contre, si on remplace 'Ali' par 'mon Ali', aucune de ces règles n'identifierait cette réponse comme étant grammaticalement correcte. On peut intégrer dans ces règles des contraintes morphologiques afin de rejeter une réponse telle que 'mon voisin'. Après avoir appliqué la règle 2 à 'ma voisine', il est possible d'appliquer la règle 4 au groupe de mots 'Le chien de <groupe nominal>'. La figure 1 illustre la réduction progressive de la structure initiale jusqu'à l'identification d'une structure de phrase élémentaire (règle 7).

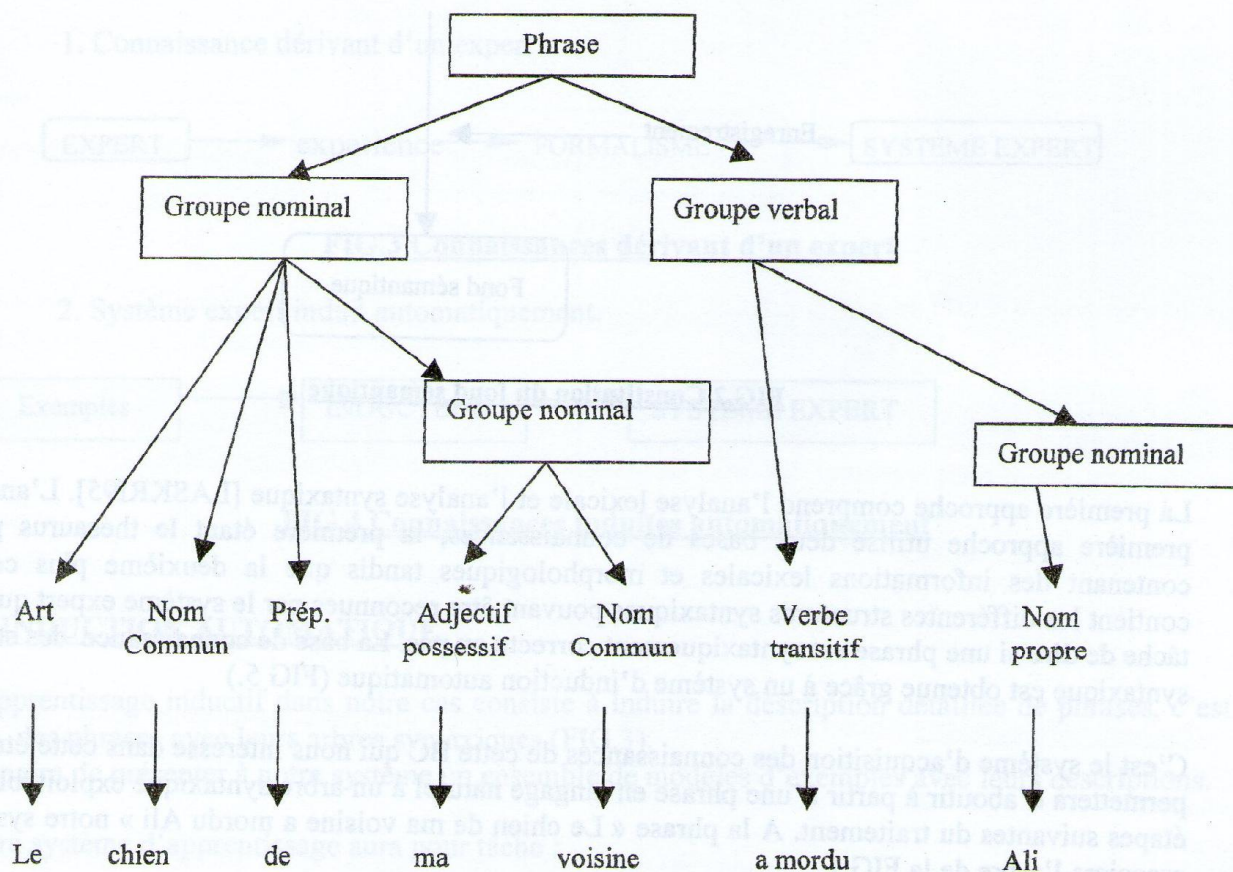


FIG.1 Arbre syntaxique

3.ARCHITECTURE DU SYSTEME D'AALN.

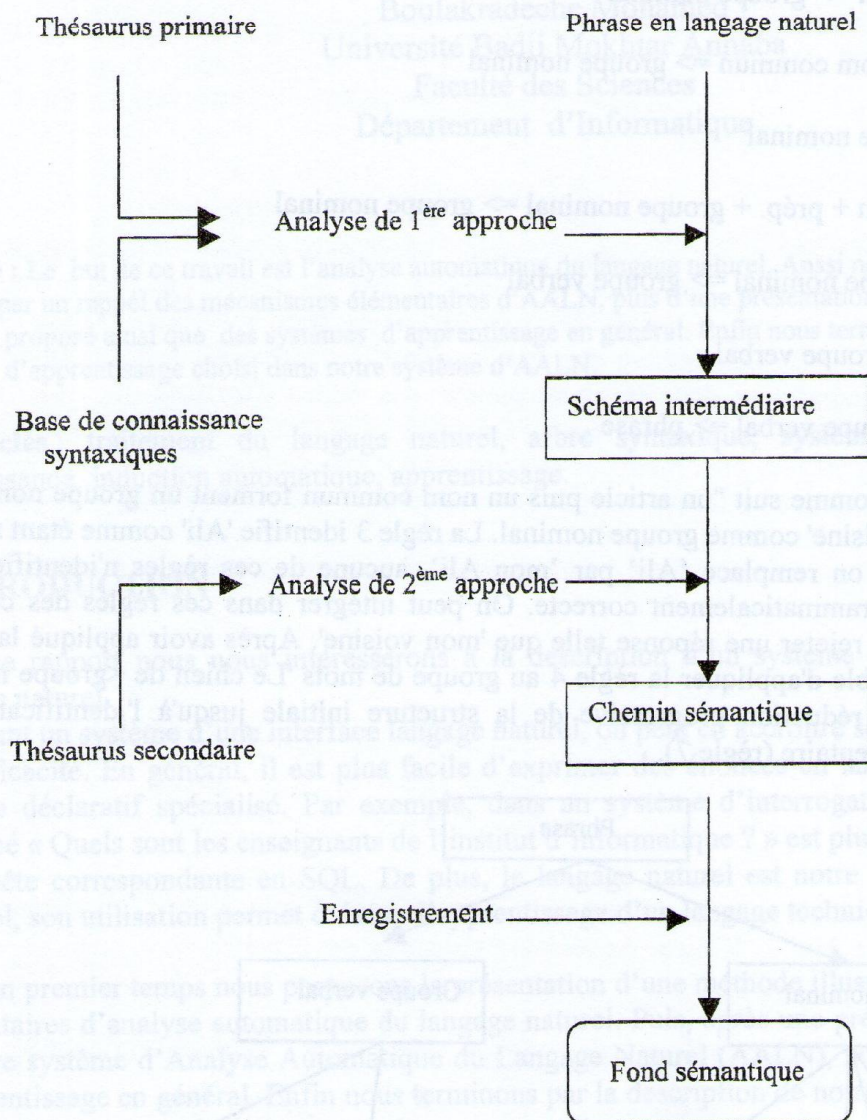


FIG.2 Constitution du fond sémantique

La première approche comprend l'analyse lexicale et l'analyse syntaxique [LASKRI95]. L'analyse de première approche utilise deux bases de connaissances, la première étant le thésaurus primaire contenant des informations lexicales et morphologiques tandis que la deuxième plus complexe contient les différentes structures syntaxiques pouvant être reconnues par le système expert qui aura la tâche de dire si une phrase est syntaxiquement correcte ou pas. La base de connaissance des structures syntaxique est obtenue grâce à un système d'induction automatique (FIG 5.)

C'est le système d'acquisition des connaissances de cette BC qui nous intéresse dans cette étude, cela permettra d'aboutir à partir d'une phrase en langage naturel à un arbre syntaxique exploitable par les étapes suivantes du traitement. A la phrase « Le chien de ma voisine a mordu Ali » notre système lui associera l'arbre de la FIG 1.

La deuxième approche s'attaque au problème de la sémantique qui est une représentation de la signification de la phrase ou du texte. Ce problème délicat n'a pas encore été abordé.

4. SYSTEME D'APPRENTISSAGE

4.1 INTRODUCTION

Il existe plusieurs types de systèmes d'apprentissage dont :

- Simple mémorisation
- Par instruction
- Par analogie
- Par la découverte
- Par les exemples

Cette liste n'est pas exhaustive, il existe d'autres types de systèmes d'apprentissage, et nous allons dans notre cas présent nous intéresser à la dernière citée.

L'apprentissage par les exemples ou apprentissage par induction signifie que celui qui apprend doit généraliser à partir d'exemples qui sont d'habitude préparés par un enseignant. C'est l'une des formes la plus activement étudiée d'apprentissage automatique.

4.2 ACQUISITION DE CONNAISSANCE POUR LES SYSTEMES EXPERTS

Ne perdons pas de vue que le but de notre travail est l'acquisition de connaissances syntaxiques sur le langage naturel pour permettre à notre système expert d'accepter une phrase correcte en langage naturel et dans la mesure du possible d'accepter une phrase comportant des erreurs mais qui reste compréhensible pour un humain.

Un système expert acquiert de la connaissance de deux façons :

1. Connaissance dérivant d'un expert .

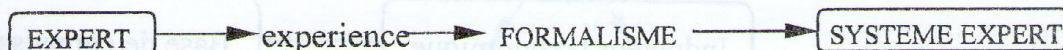


FIG.3 Connaissances dérivant d'un expert

2. Système expert induit automatiquement.

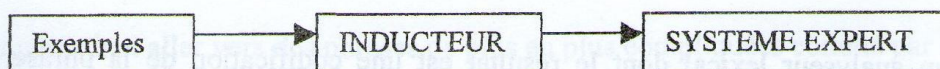


FIG.4 Connaissances induites automatiquement

4.3 INDUCTION AUTOMATIQUE

L'apprentissage inductif dans notre cas consiste à induire la description détaillée de phrases, c'est à dire des phrases avec leurs arbres syntaxiques (FIG.3).

Il s'agira de présenter à notre système un ensemble de modèles d'exemples avec leurs descriptions.

Notre système d'apprentissage aura pour tâche :

- * L'apprentissage inductif de descriptions discriminantes qui fait la différence entre concepts ou phrases totalement différentes par leurs syntaxes.
- * L'induction taxonomiques qui relient les concepts de sous_ensemble et de surensemble dans une hiérarchie.

5 - SYSTEME D'INDUCTION AUTOMATIQUE

Notre système d'induction automatique se présente sous la forme de deux modules (FIG.5) : un inducteur discriminant et un inducteur taxonomique. Le premier génère des connaissances hors contexte alors que le second génère des connaissances basées sur le premier mais tenant compte du contexte. Par contexte on comprend ici base de connaissance.

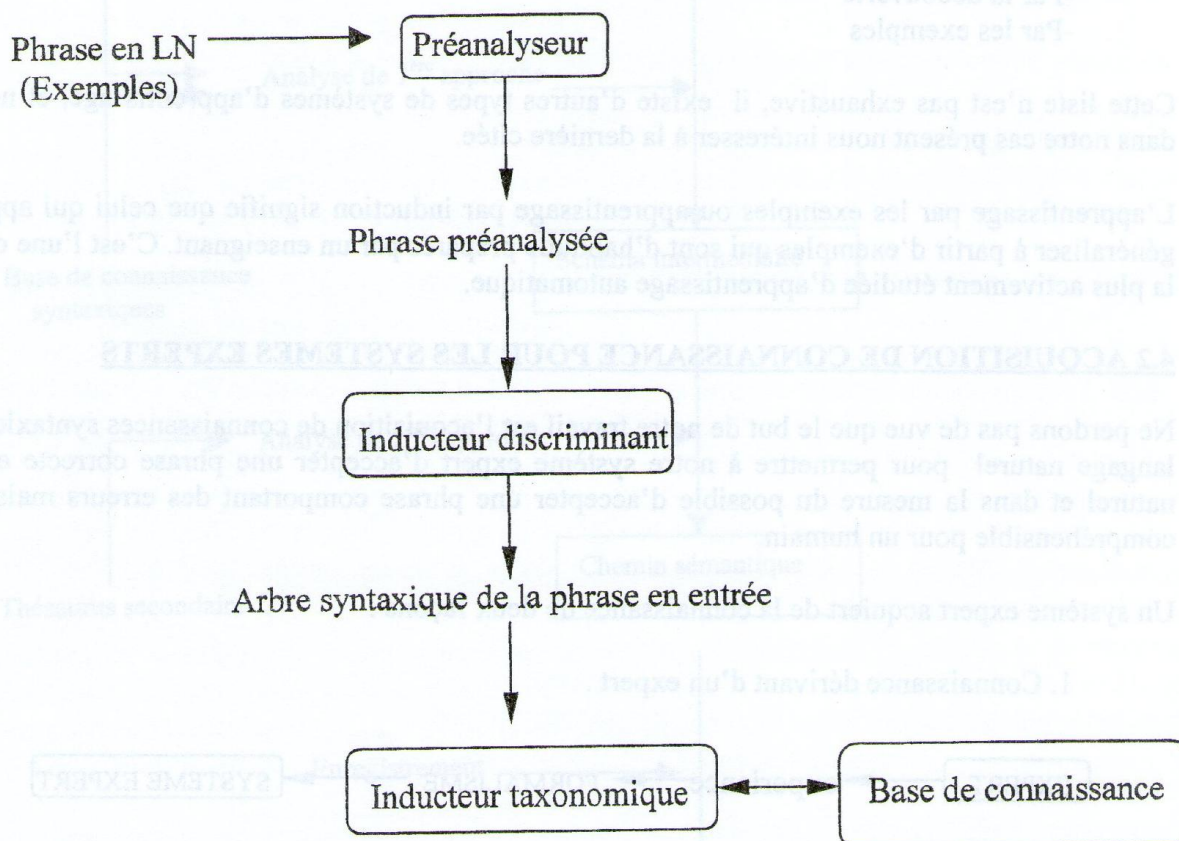


FIG.5 Système d'acquisition de connaissances

Préanalyseur : C'est un analyseur lexical dont le résultat est une codification de la phrase. La codification est un remplacement des mots ou groupe de mots par leurs valeurs lexicales.

Inducteur discriminant : Il génère une structure d'arbre avec une racine dont la valeur est PHRASE (PH) et dont les feuilles ou noeuds terminaux contiennent les mots de la phrase en langage naturel en entrée. Entre la racine et les noeuds terminaux on trouvera les valeurs lexicales associées à la phrase en cours d'analyse (FIG.1).

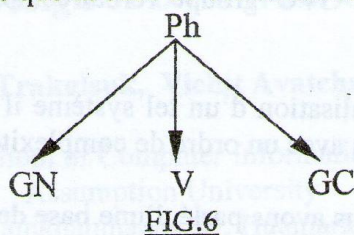
Inducteur taxonomique : Son rôle est d'insérer le résultat de l'inducteur discriminant dans la base de connaissance sans toucher à son intégrité. Souvent cette insertion est une modification de la base de connaissance ou plus exactement une mise à jour.

La base de connaissance : C'est l'ensemble des connaissances syntaxiques que notre système expert va utiliser pour reconnaître ou pas une phrase en langage naturel en entrée. En fait seuls les arbres syntaxiques seront gardés dans la BC au moment de l'acquisition quand à l'analyse elle se fait par unification avec le contenu de la base de connaissance.

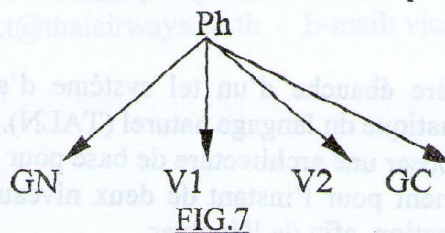
L'avantage d'une telle méthode pour remplir une BC est premièrement de ne pas faire appel à un expert, deuxièmement la BC aura une taille illimitée théoriquement.

6 - EXEMPLES D'ACQUISITION

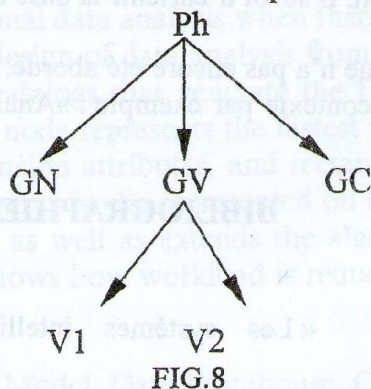
Au départ on suppose que la base de connaissance est vide.



Une phrase ayant la structure de la FIG.6 (groupe nominal suivi d'un verbe et d'un groupe complément) résultat de l'inducteur discriminant sera insérée telle quelle dans la B.C.

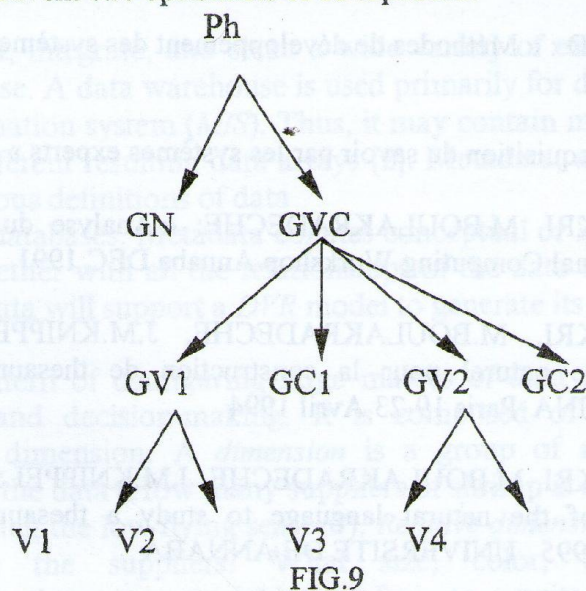


Par contre une phrase ayant la structure de la FIG.7 (même structure que la phrase de la FIG.6 mais avec deux verbes au lieu de un) ne sera pas insérée telle quelle. Pour l'instant notre B.C.=FIG.6, La répétition du verbe va donner une B.C.=FIG.6+FIG.7 ce qui nous donne FIG.8



Notre système d'analyse va pouvoir accepter des phrases de type FIG.6 et FIG.7 c'est à dire du type FIG.8 avec V2 (le deuxième verbe) optionnel.

On peut ainsi aller vers des phrases de plus en plus compliquées comme par exemple celle de FIG.9 ou des éléments de l'arbre peuvent être optionnels et/ou répétitifs.



Ph=phrase; GN=groupe nominal; GVC=groupe verbal complément; GV=groupe verbal; GC=groupe complément; V=verbe.

Remarque : Pour faciliter la réalisation d'un tel système il est important que les phrases en entrée pour l'inducteur soient présentées avec un ordre de complexité syntaxique croissant.

Tout au long de ces exemples nous avons parlé d'une base de connaissance ayant la structure d'arbre. Elle peut tout aussi bien avoir la forme d'un ensemble de règles de la forme : **Si** condition(s) **Alors** action(s).

7 - CONCLUSION .

Ce travail étant une première ébauche d'un tel système d'acquisition de connaissances dans le domaine du traitement automatique du langage naturel (TALN).

Notre démarche a été de proposer une architecture de base pour système de compréhension du langage naturel composé principalement pour l'instant de deux niveaux d'analyse lexicale plus syntaxe. Ce système est en cours de réalisation afin de l'évaluer.

L'analyse de première approche permet de générer une représentation interne (schéma intermédiaire), elle facilite la prise en compte d'une variété assez importante de phrases en langage naturel tout en ayant la possibilité d'insérer le traitement d'une nouvelle forme de phrase sans aucun apport supplémentaire de programmation. Il suffit d'enrichir la base de connaissances avec la nouvelle forme

Le problème d'analyse sémantique n'a pas encore été abordé. C'est un de nos projets néanmoins nous pensons qu'il doit être lié à un contexte par exemple : »Analyse des réponses dans une situation de dialogue «.

BIBLIOGRAPHIE

[BLACK 88] W.J.BLACK : « Les systèmes intelligents basés sur la connaissance » MASSON 1988

[BOULAK 93] M.BOULAKRADECHE : « Analyse de première approche à base de connaissances » Journée d'étude sur le TALN - Annaba 26-27 Avril 1993

[BOULAK 94] M.BOULAKRADECHE: « Analyse du langage naturel à base connaissances » Thèse de magister en informatique - Annaba NOV.1994

[BRIAND 88] R. BRIAND : « Méthodes de développement des systèmes experts » Editions Eyrolles, Paris 1988

[HART 88] A.HART: « Acquisition du savoir par les systèmes experts » Masson, Paris 1988

[LASKRI 91] M.T.LASKRI, M.BOULAKRADECHE: « Analyse du langage naturel à base de connaissances » International Computing Workshop Annaba DEC 1991

[LASKRI 94] M.T.LASKRI, M.BOULAKRADECHE, J.M.KNIPPEL: « Analyse à base de connaissances du langage naturel pour la construction de thesaurus » Colloque international CONSENSUS EX MACHINA Paris 19-23 Avril 1994

[LASKRI 95] M.T.LASKRI, M.BOULAKRADECHE, J.M.KNIPPEL: « A knowledge base for the first approach analysis of the natural language to study a thesaurus support system » Revue SYNTHÈSE N°00 JUIN 1995 UNIVERSITE DE ANNABA

[REMY 90] C.REMY: « L'IA : nature, langage, outil » Micro-système, Mai 1990