UNIVERSITE CLAUDE BERNARD LYON-I
43, Boulevard du 11 Novembre 1918
69621 VILLEURBANNE

Diplôme d'Etudes Supérieures Specialisées

informatique documentaire

* NOTE DE SYNTHESE

TRAITEMENT AUTOMATIQUE DES LANGUES

ET DOCUMENTATION

AUTEUR : Paulette BERNHARD

DATE : mai 1979

UNIVERSITE CLAUDE BERNARD (LYON - 1)

D.E.S.S. D'INFORMATIQUE DOCUMENTAIRE

T R A I T E M E N T A U T O M A T I Q U E

D E S L A N G U E S

E T D O C U M E N T A T I O N

Note de synthèse

présentée par

Paulette BERNHARD

mai 1979

T A B L E D E S M A T I E R E S

Considérons une collection de 20.000 documents composés chacun de 400 caractères, soit au total 20.000 x 400 = 8 millions de caractères, ou 160.000 lignes, une ligne étant formée de 50 caractères.
Si un individu lit en moyenne 1 ligne à la seconde, il lui faudra approximativement 40 heures pour lire la totalité des documents, tandis que l'ordinateur accomplira cette opération de lecture en moins d'une minute.

note extraite de l'ouvrage de L. BOURELLY et E. CHOURAQUI sur le système documentaire SATIN 1.

N O T E P R E L I M I N A I R E

Je suis bien consciente du fait que le sujet de cette note de synthèse reste vaste, malgré les limitations successives apportées.

Il m'a semblé important, puisque j'avais réuni une informatio assez diversifiée, de tenter de faire plutôt une présentation sommaire des différents systèmes recensés que leur étude approfondie, qui pourrait cependant faire l'objet de travaux - et de mémoires - ultérieurs.

L'étude qui suit n'est donc ni exhaustive ni même très poussé Elle pourrait être considérée comme une espèce de "dossier" sur quelques aspects du traitement automatique des langues en documentation.

Je voudrais enfin remercier toutes les personnes - et elles sont nombreuses - qui m'ont permis d'accéder aux informations, ainsi qu'à l'ensemble des documents cités dans la bibliographie.

I PRESENTATION ET ESSAI DE DELIMITATION DU SUJET

I PRESENTATION ET ESSAI DE DELIMITATION DU SUJET

Le traitement automatique des langues relève à la fois des Mathématiques, de l'Informatique et de la Linguistique; son but central, en documentation, est de permettre l'analyse automatique des documents, mais il peut déboucher aussi bien sur des traitements d'édition automatique d'index que sur des processus de traduction automatique, ou encore sur la construction automatique de thesaurus, et bien d'autres applications...

Dans chacun de ces types d'opérations entrent en jeu des procédures plus ou moins complexes, à partir de textes ayant déjà subi une préparation d'entrée, ou à partir de textes en langue naturelle; ces procédures peuvent être uniquement statistiques (fréquences, cooccurences), uniquement linguistiques (aspects morphologiques, syntaxiques, sémantiques), ou encore combinées.

Le domaine du traitement automatique des langues, en particulier des langues naturelles, dans les systèmes documentaires, fait encore très largement l'objet de recherches et d'expérimentations.

Les recherches bibliographiques

Le champ bibliographique couvert au départ était très vaste. Les recherches ont porté :

- sur le dépouillement pour l'année 1978 d'un certain nombre de bibliographies spécialisées (Bulletin signalétique 101, Information science abstracts, library and information science abstracts, computing rewiews)

- sur une consultation plus complète en particulier des revues Documentaliste et TA Informations

- sur une interrogation d'un fichier de l'ASE (Agence spatiale européenne) ,qui n'a malheureusement pas donné d'informations immédiatement utilisables

- sur les bibliographies citées dans des ouvrages plus généraux (1) ou dans des articles de synthèse (2) concernant l'analyse documentaire

- enfin, au fur et à mesure de leur consultation, sur les bibliographies des différents ouvrages ou articles concernant chacun des systèmes ou aspects particuliers.

C'est ainsi qu'une centaine de références avaient été collectées, principalement à partir des concepts suivants (en français et en anglais) : ANALYSE AUTOMATIQUE / LANGAGE NATUREL / INDEXATION AUTOMATIQUE / TRADUCTION AUTOMATIQUE / TRAITEMENT AUTOMATISE .

Dans un premier tri, je n'ai retenu que celles concernant l'entrée des informations dans les systèmes documentaires, y compris celles incluant l'indexation et la traduction automatique, en éliminant d'une part les méthodes purement statistiques et, de l'autre, celles propres à la recherche ou à la formulation des questions : il restait une quarantaine de références.

Par la suite, compte tenu des objectifs de l'étude (aboutir à une "note de synthèse" en un temps assez court), ainsi que des documents qu'il m'a réellement été possible d'obtenir et de consulter, le travail qui suit se limite à la présentation de quelques systèmes - opérationnels ou en cours d'expérimentation - partagés entre :

---

(1) en particulier les ouvrages de BELY, BORILLO, VIRBEL - CHAUMIER et COYAUD, SIOT DECAUVILLE

(2) les articles de COURRIER et SPARK JONES

(3) Cependant, l'aspect "traduction automatique des langues naturelles" a été volontairement négligé dans cette étude (il correspond à une grande partie des références éliminées). On peut signaler toutefois que les travaux les plus importants menés dans ce domaine en France ont été conduits par le GETA (M. VAUQUOIS, Grenoble), qui continue ses recherches en particulier sur des applications possibles à la documentation.

- des systèmes qui exigent, à l'entrée, des types plus ou moins élaborés d'expressione de relations et qui aboutissent d'une part à l'édition automatique d'index et, de l'autre, à des traitements plus spécifiques des informations que dans les systèmes automatisés traditionnels

- des systèmes qui traitent le texte intégral en langage naturel et dans lesquels les composantes linguistiques sont intégrées dans les programmes de traitement eux-mêmes.

Le lien entre ces différente réalisations est la présence de composantes linguistiques (en particulier grammaticales) intervenant soit au moment de l'introduction des données soit au niveau du traitement.

L'ensemble des systèmes décrits est le suivant :

/ comme exemples d'édition automatique d'index

- OLPI (On-Line Phrase Input) - Chemical Abstracts Service
- NEPHIS (NEsted PHrase Indexing System) - University of Western Ontario
- Relational indexing - University of Western Ontario
- PRECIS (PREserved Context Indexing System) - British Library

/ comme exemples de systèmes exigeant l'introduction de descriptions élaborées (utilisation d'un métalangage)

- SATIN 1 - Unité de recherche Analyse documentaire et calcul en archéologie. CNRS.
- TITUS II - Institut textile de France

/ comme exemples et expérimentations du traitement du texte intégral

- SYNTAXEME - Direction des Recherches et Moyens d'Essais (DRME)
- Système à l'étude à l'Institut national des Sciences et Techniques Nucléaires (INSTN/CEA) et au Centre d'Informatique Juridique (CEDIJ)
- PIAFDOC (Programmes Interactifs d'Analyse du Français appliqués à la DOCumentation) - Université de Grenoble et Documentation française
- CONDOR - (Communication in Natürlicher Sprache mit Dialog Orientierten Retrieval Systemen) - Société Siemens

II SITUATION DANS LA CHAINE DOCUMENTAIRE

II SITUATION DES DIFFERENTS SYSTEMES RETENUS DANS LA CHAINE DOCUMENTAIRE

Lee étiquettes 1 à 4, placées sur le schéma de la chaîne documentaire qui suit, désignent les différents points où interviennent les systèmes décrits; il s'agit en particulier :

- de celui de la caractérisation du contenu , où les systèmes aussi bien d'édition automatique que de descriptions à l'aide d'un métalangage élaboré réclament un travail humain d'analyse et de décision

- de celui des traitements à proprement parler automatiques de la langue, qui se font sur le mode conversationnel et où souvent sont prévues des interruptions pour les nombreux cas (d'erreurs, d'ambiguïtés, de mots nouveaux à introduire, etc.) que doit résoudre, là encore, le documentaliste ou l'analyste.

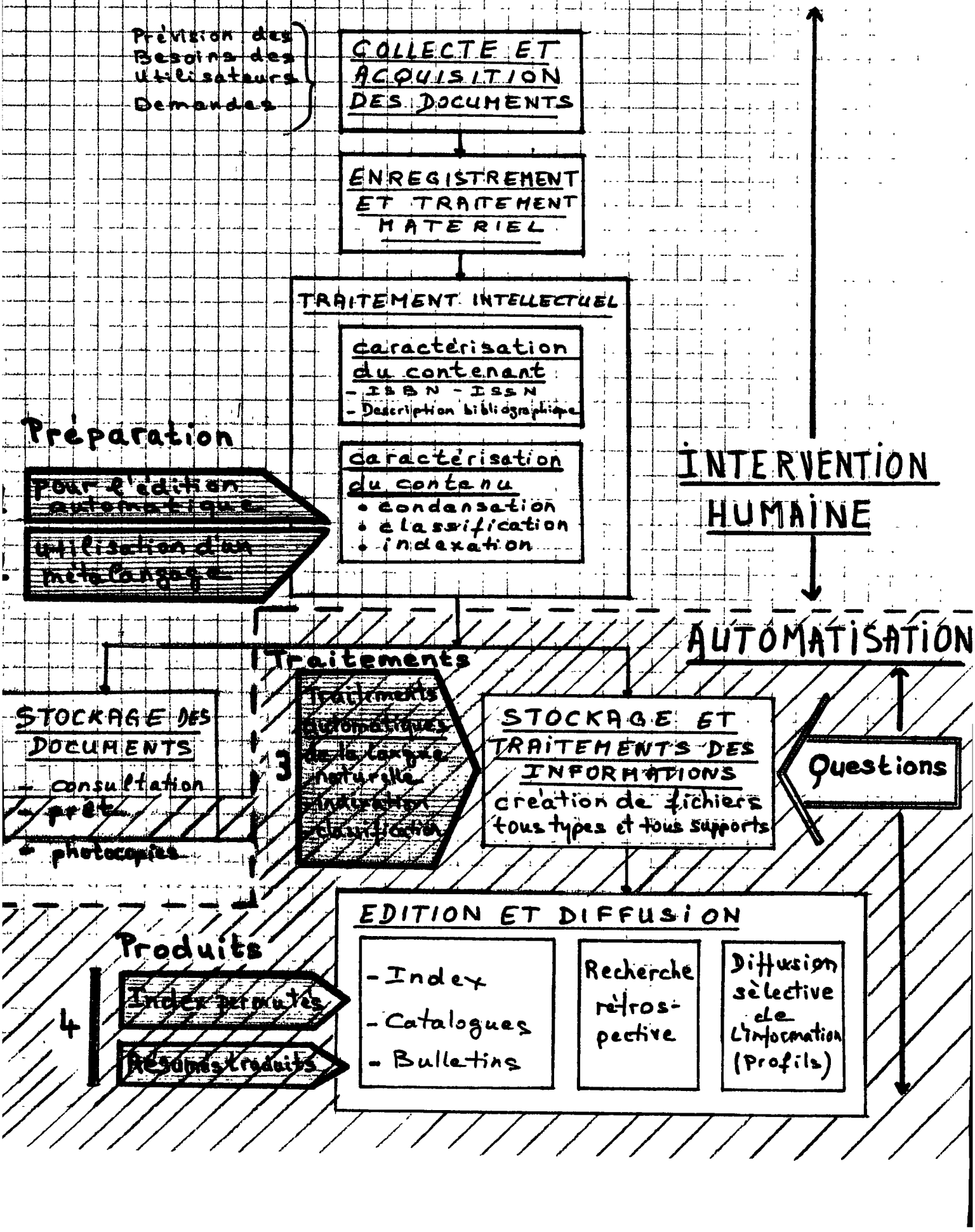
- de celui, qui semble le plus performant mais n'est que le résultat des phases de description antérieures, de l'édition automatique d'index et des produits de la traduction automatique

II semble résulter de cette rapide présentation que la notion de "traitement automatique des langues" ne se traduit encore que bien rarement - en entier - dans la réalité et que l'intervention humaine reste relativement prépondérante.

Une évolution se dessine cependant, car il y a une importante différence entre le travail intellectuel que demande par exemple l'indexation "traditionnelle" et les interventions de choix que réclament les programmes; dans ce cas, une grande partie de l'analyse est déjà "digérée" par le système (mais la mise au point des programmes est un travail très lourd et très coûteux).

Schéma simplifié de la chaîne documentaire

et domaines d'application du traitement automatique des langues



III EDITION AUTOMATIQUE D'INDEX

III EXEMPLES DE SYSTEMES D'EDITION AUTOMATIQUE D'INDEX PERMUTES

Les informations recueillies ont permis de recenser quatre systèmes visant à l'édition automatique d'index permutés et nécessitant une présentation plus ou moins compliquée des données d'entrée.

Les deux premiers, OLPI et NEPHIS, sont très simples et ne font pratiquement pas intervenir de relations entre les termes à permuter.

Les deux suivants, RELATIONAL INDEXING et PRECIS, plus élaborés, utilisent une série d'indicateurs de relations ou d'opérateurs de rôle.

Les pages suivantes présentent les grands traits de ces systèmes; pour plus de précision, chaque description est complétée par des schémas et des exemples.

III. 1 O L P I (On-Line Phrase Input)

Ce système est mis en oeuvre, à titre expérimental, par le Chemical Abstracts Service, pour la génération d'index "articulés" à partir de phrases en langage naturel.

Il est caractérisé par un processus de traitement interactif, qui fait intervenir un jeu de 12 fonctions et une suite d' "écrans programmés", chaque écran correspondant au choix, par l'analyste, d'un mot ou d'un groupe de mots dans la phrase introduite.

Le seul type de relation, ou plutôt d'inclusion, qui est prévu par le système est celui aboutissant à la création de spécifiques d'un terme alors qu'ils n'apparaissent pas dans la phrase initiale (cf les écrans 5 et 6 de la page suivante).

L'expérience décrite proposait deux types de procédures :

- soit le système propose à l'analyste une suite d'écrans sur lesquels il désigne le ou les mots à sélectionner comme entrées (voir l'ensemble des schémas de la page suivante)

- soit il propose lui-même une série de mots "candidats", que l'analyste accepte ou refuse ou modifie, par l'intermédiaire d'un "écran de transaction".

La seconde procédure n'aboutit pas exactement aux mêmes résultats que la première et nécessite encore quelques "ajustements".

Le système est présenté comme satisfaisant du point de vue opérationnel; une moyenne de 3,31 "entrées" ont été générées par phrase, 96,5 % d'entre elles étant considérées comme "acceptables".

Entrée du texte sur l'écran 1 (où seul le n° d'identification apparaît)

écran 2

FLAG HEADING
76678546X
effect of sodium thiocyanate on glycoproteins secretion in
1 2 3 4 5 6 7 8
tracheal mucus
9 10
HDG: [3,4]
NUM: [1]
CLASS:

choix des mots 3 et 4 →

écran 3

FLAG HEADING
76678546X
effect of sodium thiocyanate on glycoproteins secretion in
1 2 3 4 5 6 7 8
tracheal mucus
9 10
HDG: [6]
NUM.
CLASS:

choix du mot 6 ↓

écran 4

FLAG HEADING
76678546X
effect of sodium thiocyanate on glycoproteins secretion in
1 2 3 4 5 6 7 8
tracheal mucus
9 10
HDG: [10]
NUM.
CLASS:

choix du mot 10 →

écran des résultats correspondant à 2, 3 et 4 :

RESULTS
76678546X
effect of sodium thiocyanate on glycoproteins secretion in tracheal mucus

Sodium thiocyanate
glycoproteins secretion in tracheal mucus in relation to

Glycoproteins
secretion, in tracheal mucus, sodium thiocyanate effect on

Mucus
tracheal, glycoproteins secretion in, sodium thiocyanate effect on

---

Entrée de termes spécifiques n'apparaissant pas dans le texte initial :
termes spécifiques de "glycoproteins"

écran 5

SURROGATE
76678546X
effect of sodium thiocyanate on glycoproteins secretion in
1 2 3 4 5 6 7 8
tracheal mucus
9 10
CTH: GLYCOPROTEINS
SURROGATE: [Avidins]
NUM: [1]
HEADING WORD NUMBERS: 6

ajout du mot "avidins" ↓

écran 6

SURROGATE
7667854X
effect of sodium thiocyanate on glycoproteins secretion in
1 2 3 4 5 6 7 8
tracheal mucus
9 10
CTH: GLYCOPROTEINS
SURROGATE: [GLYCOPHORINS]
NUM:
HEADING WORD NUMBERS: 6

ajout du mot "glycophorins" →

écran des résultats correspondant à 5 et 6 :

RESULTS
7667854X
effect of sodium thiocyanate on glycoproteins secretion in tracheal mucus

Avidins
secretion, in tracheal mucus, sodium thiocyanate effect on

Glycophorins
secretion, in tracheal mucus, sodium thiocyanate effect on

III. 2 N E P H I S (NEsted PHrase Indexing System)

Ce système, mis au point à l'University of Western Ontario, vise à la production d'une "indexation permutée par matières assistée par ordinateur". Il permet aussi bien la permutation de phrases courtes (titres) que celles d'ensembles plus longs (par exemple des notices complètes).

Dans ce système, le jeu des relations entre les mots ou groupes de mots est exprimé par des emboitements successifs, que dirigent quatre fonctions de commande (utilisation des caractères spéciaux suivants : <, >, ?, @). Les mots ou groupes de mots sont alors remplacés par un tiret dans les différentes permutations.

Les caractères spéciaux ? et @ marquent les chaînes de caractères (en particulier les prépositions et les débuts de phrases non significatifs) qu'il convient ou non de permuter et d'éditer, selon les cas.

On verra, dans le schéma de la page suivante, deux exemples de présentation des textes à permuter, avec en regard les permutations obtenues.

Il est prévu, dans une exploitation ultérieure du système en mode conversationnel, de regrouper les "entrées" par têtes de chapitre.

Pourvu que l'on évite d'employer des verbes, que l'on utilise de préférence des substantifs et des prépositions ainsi que, si possible, un vocabulaire contrôlé, ce système est présenté comme permettant d'assurer, en des permutations élégantes, la description de sujets compliqués - et cela à un coût très raisonnable (en place et en temps de traitement).

Exemple de présentation de la phrase d'entrée :

Teaching of <Information Science> at <University of <Strathclyde>>

permutations générées par le programme :

Teaching of Information Science at University of Strathclyde

Information Science, Teaching of–at University of Strathclyde

University of Strathclyde. Teaching of Information Science at–

Strathclyde, University of–. Teaching of Information Science at–

Traitement d'une notice complète

Computer Analysis of Library Postcards (CALP)?/ <@Norman D <Stevens>>?. – in <Journal? of the <American Society for Information Science> 25 (1974), 332>?: <Humour? on <@Analysis? of <Postcards? Depicting <Libraries?. Depictions –>>? by <Computers>>?: >

permutations obtenues :

Computer Analysis of Library Postcards (CALP)/ Norman D Stevens. – in Journal of the American Society for Information Science 25 (1974), 332: Humour on Analysis of Postcards Depicting Libraries by Computers

Computers. Analysis of Postcards Depicting Libraries. Humour: Computer Analysis of Library Postcards (CALP)/ Norman D Stevens. –in Journal of the American Society for Information Science 25 (1974), 332

Stevens. Norman D –. Computer Analysis of Library Postcards (CALP). – in Journal of the American Society for Information Science 25 (1974), 332: Humour on Analysis of Postcards Depicting Libraries by Computers

Humour on Analysis of Postcards Depicting Libraries by Computers: Computer Analysis of Library Postcards (CALP)/ Norman D Stevens. – in Journal of the American Society for Information Science 25 (1974), 332

American Society for Information Science. Journal 25 (1974), 332. Computer Analysis of Library Postcards (CALP)/ Norman D Stevens: Humour on Analysis of Postcards Depicting Libraries by Computers

Libraries. Depictions–Postcards. Analysis by Computers. Humour: Computer Analysis of Library Postcards (CALP)/ Norman D Stevens. –in Journal of the American Society for Information Science 25 (1974), 332

Journal of the American Society for Information Science 25 (1974), 332. Computer Analysis of Library Postcards (CALP)/ Norman D Stevens: Humour on Analysis of Postcards Depicting Libraries by Computers

Postcards Depicting Libraries. Analysis by Computers. Humour: Computer Analysis of Library Postcards (CALP)/ Norman D Stevens. – in Journal of the American Society for Information Science 25 (1974), 332

III. 3 RELATIONAL INDEXING SYSTEM

C'est également à l'University of Western Ontario que ce système a été expérimenté, sur 1000 puis 3000 abstracts : il tente d'introduire les mécanismes de la "psychologie de la pensée" dans l'expression de relations entre les sujets et se base sur des combinaisons par paires de mots ou groupes de mots, selon neuf "catégories de relation", associées à neuf opérateurs.

Ces relations peuvent être représentées dans des diagrammes à deux dimensions, où l'on observe des "directions" (de haut en bas et de gauche à droite) signifiant que le mot du bas ou de droite est subordonné à celui du haut ou de gauche.

Ces diagrammes présentent une analogie avec les formules des structures de composants en chimie organique et peuvent, comme elles, être transcrites dans une "table de connection".

Cette opération - qui paraît relativement compliquée - est ensuite traduite en "format d'entrée" (dont on trouvera un exemple à la page suivante).

Les signes de relation sont remplacés, à l'édition, par des prépositions ou des phrases prépositionnelles.

Les auteurs assurent que cette méthode de représentation des relations n'entraîne pas de distorsions de sens lors des permutations.

Liste et signification des catégories de relation :

	Awareness	Temporary Association	Fixed Association
Concurrent	/0 Concurrent	/* Self-activity	/; Association
Not distinct	/= Equivalence	/+ Dimensional	/(Appurtenance
Distinct	/) Distinctness	/- Action	/: Functional dependence

leur traduction numérique :

/0 1	/* 4	/; 7
/= 2	/+ 5	/(8
/) 3	/- 6	/: 9

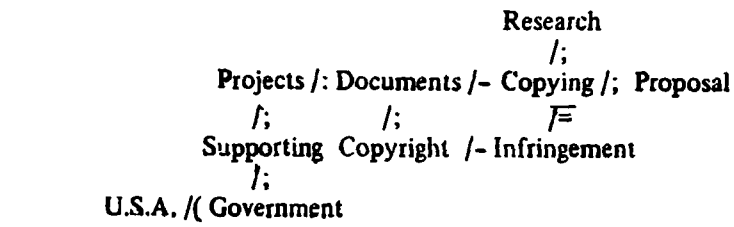
Phrase indexée :

"A proposal that copying for research should not be an infringement of copyright of documents from projects supported by the government of the U.S.A " (J. Amer. Soc. Inf. Sci., 1974, *25*, 145).

Format d'entrée :

```
n=25145
v=1;s=projects
v=2;s=supported
v=1;s=government
v=1;s=USA
v=1;s=documents
v=2;s=copying
v=1;s=research
v=2;s=infringement
v=1;s=copyright
v=1;s=proposal
w=1;p=*;r=7;p=*;w=2
w=2;r=7;p=*;w=3
w=4;r=8;l=1;w=5
g=1;w=1;r=9;l=1;w=5
w=5;r=6;l=1;w=6
g=7;w=7;r=7;p=for;l=1;w=6
w=6;r=11;w=8
w=9;r=6;p=of;w=8
g=4;w=5;r=7;l=1;w=9
w=6;p=:;r=7;w=10
```

Diagramme correspondant :



Présentation d'une des permutations générées :

Government
of USA supported projects. : documents
copying for research not as infringement
of copyright : proposal.

25145

III. 4 P R E C I S (PREserved Context Indexing System)

Ce système, mis au point dans le cadre de l'automatisation de l'index par matières de la British national bibliography, est également basé sur une indexation codée à l'entrée, suivie de l'édition par ordinateur. Il a été expérimenté par le Département dee Arts et Spectacles de la Bibliothèque nationale (1) et dans le cadre d'EUDISED pour la constitution d'un corpus commun de matériels audio-visuels (2).

Il est original par la méthode d'analyse du sujet qu'il introduit, ainsi que par ses composantes : sémantique (thésaurus) et syntaxique (opérateurs de rôle ou de fonction).

L'observation d'une corrélation entre ces opérateurs, énumérés toujours dans le même ordre, et les "cas profonds" de la langue de départ,qu'il est possible de considérer comme des "universels linguistiques" (2), a amené d'une part l'extension du système à d'autres langues européennes (allemand, français, danois, polonais en particulier) et, de l'autre, l'étude de son "potentiel translinguistique"(3).

La page suivante donne deux exemples de présentation des chaînage d'entrée et des permutations correspondantes, ainsi que la liste des opérateurs de rôle.

Ce système est le seul parmi ceux retenus qui non seulement fasse appel à un thesaurus mais encore contribue à l'élaboration de celui-ci puisqu'il augmente au fur et à mesure que des termes nouveaux sont rencontrés et retenus.

---

(1) cf FERRIER
(2) cf SØRENSEN
(3) cf VERDIER, AUSTIN - On entend "translinguistique" dans le sens suivant :"commutation de chaînes d'entrée de termes dans une langue source en chaînes équivalentes dans une langue cible différente".

Géré par un processus indépendant, le thesaurus admet les trois relatione d'équivalence, générique et aesociative; les renvois correspondant à un terme sont reproduits automatiquement dans l'index si ce terme figure dans un chaînage d'entrée.

PRECIS représente sans doute le système le plus prometteur dans le domaine de l'édition automatique d'index. Il est proche, par l'élaboration qu'il exige des données d'entrée, des systèmee plus complets qui vont être décrits dans le chapitre suivant; mais on pourrait peut-être regretter qu'il se limite à l'édition automatiqus st n'évolue pas vers des traitements plus élaborés.

Chaines d'entrée :

1) Groenland
2) Exploration $v par $w de
3) Rasmussen, knud

Permutations opérées :

GROENLAND
Exploration par Rasmussen, Knud

EXPLORATION. Groenland
par Rasmussen, Knud

RASMUSSEN, KNUD
Exploration du Groenland

cf. SØRENSEN

(1) théâtre
(p) acteur
(p) jeu de l'acteur
(2) enseignement $w du
(s) rôle $v de l' $w dans l'
(3) improvisation

THÉATRE
Acteur. Jeu de l'acteur. Enseignement. Rôle de l'improvisation.
ACTEUR. Théâtre
Jeu de l'acteur. Enseignement. Rôle de l'improvisation.
JEU DE L'ACTEUR. Théâtre.
Enseignement. Rôle de l'improvisation.
IMPROVISATION. Théâtre
Rôle dans l'enseignement du jeu de l'acteur.

cf. FERRIER

LISTE DES OPÉRATEURS DE ROLE

	OPÉRATEURS PRINCIPAUX	
	Environnement du système observé	o Lieu
	Système observé (Opérateurs de base)	1 Élément-clé : objet d'une action transitive; agent d'une action intransitive 2 Action/Effet 3 Agent d'une action transitive; Aspects; Facteurs
A	Données se rapportant à l'observateur	4 Point de vue
	Exemple choisi	5 Échantillonnage de population/Région étudiée
	Présentation des données	6 Public/Forme
	OPÉRATEURS INTERPOSÉS	
	Éléments dépendants	p Partie/Propriété q Membre d'un groupe quasi-générique r Ensemble
	Relations entre concepts	s Définisseur de rôle t association due à l'auteur
B	Concepts coordonnés	g concept coordonné
	OPÉRATEURS DIFFÉRENCIATEURS (préfixés par $)	h Différenciateur de 1er niveau n'apparaissant pas en vedette i Différenciateur de 1er niveau apparaissant en vedette j Différenciateur mis en évidence k Différenciateur de *n*ième niveau n'apparaissant pas en vedette m Différenciateur de *n*ième niveau apparaissant en vedette n Différenciateur entre parenthèses n'apparaissant pas en vedette o Différenciateur entre parenthèses apparaissant en vedette d Date
	CONNECTEURS (Éléments de locutions de liaison préfixés par $)	v terme de descente w terme de remontée
C	LIENS ENTRE THÈMES	x 1er élément d'un thème coordonné y élément consécutif d'un thème coordonné z élément d'un thème commun

IV SYSTEMES UTILISANT UN METALANGAGE

IV EXEMPLES DE SYSTEMES NECESSITANT UNE DESCRIPTION ELABOREE DES INFORMATIONS (UTILISATION D'UN METALANGAGE)

Les systèmes décrits précédemment pourraient être considérés comme marginaux par rapport aux systèmes documentaires automatisés, dans la mesure où ils ne sont pas conçus pour permettre le stockage et le traitement des informations en vue de la recherche, ce que réalisent les deux systèmes dont la présentation suit.

En effet, SATIN 1 et TITUS II sont des systèmes complets permettent, par la finesse et l'élaboration - en particulier linguistique - des données d'entrée, d'aboutir à des traitements et à des produits très diversifiés.

IV. 1 LE SYSTEME SATIN 1

Il est extrêmement difficile de tenter de résumer le contenu très détaillé des deux volumes décrivant le système documentaire SATIN 1.

Deux articles, cependant, se consacrent au point central de la description des données, tandis qu'un troisième expose le détail d'une application dans le domaine biologique (1).

Le logiciel SATIN 1, mis au point dans le cadre du CNRS par l'Unité de recherche Analyse documentaire et calcul en archéologie

(1) voir la bibliographie

de Marseille, également utiliaé par le Centre de documentation pour l'urbanisme, e'étend au domaine agronomique (INRA : recherches sur les nématodes) et paraît particulièrement adapté à des centres documentaires très spécialisée et orientée vers la recherche.

En effet, ses caractéristiques principales sont de permettre une représentation très fine de diverses populations de "documents" (textes, enquêtes, objets, etc.) et de fournir en sortie, outre les possibilités habituelles de recherche documentaire, des produits supplém ntaires facilitant l'interprétation des résultats (tableaux synoptiques ...) ou représentant des résultats intermédiaires pour de nouveaux traitements (calculs statistiques, tris, tracés...).

Le système est basé sur un langage documentaire composé d'un vocabulaire très structuré et d'une syntaxe. Les informations d'entrée, ou "documents" (1), sont décrites en deux parties distinctes (thématique et descriptive) qui gouvernent également deux types de phrases (voir p. 2 des schémas):

- les phrases de "type texte", qui regroupent les informations d'identification (auteur, titre, etc.) et de type thématique et dont les éléments sont appelés des "CHAMPS".

  exemple de phrase de type texte :
  ⟨code verbe⟩ ⟨champ 1⟩ ..... ⟨champ n⟩

- les phrases de "type descripteur" qui font plus particulièrement référence à l' "analyse descriptive" et dont les éléments sont appelés des "DESCRIPTEURS"

  exemple de phrase de type descripteur :
  ⟨code verbe⟩ ⟨chaîne descriptive 1⟩ ..... ⟨chaîne descriptive n⟩

---

(1) Un "document" est l'ensemble des informations organisées jugées pertinentes pour représenter une ou plusieurs unités physiques (cf. l'article paru dans Automatisme - voir aussi p. 1 des schémas)

LE LEXIQUE

Le lexique correspondant est formé de catégories lexicales désignées par des termes appelés "VERBES", et qui peuvent introduire les deux types de phrases.

Cependant, seules les catégories de type "DESCRIPTEUR", qui peuvent être hiérarchisées (en des arborescences allant jusqu'à 30 niveaux) ou non hiérarchisées (exprimées linéairement), sont enregistrées dans le système et exploitées (voir p. 3 des schémas). Ainsi, à chaque "descripteur" est associé un "nombre niveau" et un code numérique.

Les éléments des structures hiérarchisées ont entre eux un certain nombre de relations (générique/spécifique, appartenance à un niveau, contiguité, succession); les structures peuvent être explorées à l'aide d'opérateurs lexicaux comme "CHEMINEMENT ASCENDANT" ou "DESCENDANT", tandis que l'opérateur "COINCIDENCE" interdit ce type de consultation du lexique.

LA SYNTAXE

Les relations syntaxiques entre les descripteurs sont établies à posteriori; elles peuvent être orientées ou non, et comprendre ou non un indicateur de liaison.

Les descripteurs, enfin, peuvent être caractérisés par des valeurs qui les qualifient (valeurs sémantiques des classes syntaxiques) ou les quantifient (valeurs algébriques).

Le système génère, par application de règles de transformation, le "document" lui-même, ainsi que les deux types de phrases: TEXTE et DESCRIPTEUR, à partir du lexique et des relations syntaxiques. Il opère, à l'enrégistrement, une analyse automatique de la

syntaxe du métalangage, qui entraîne le rejet de tout document comportant une erreur.

Les traitements aboutissent à des fichiers de recherche, thématique et descriptif, auxquels on peut accéder soit globalement soit partiellement (entrée statique ou dynamique). La consultation se fait au moyen d'un langage d'interrogation qui permet de manipuler des opérateurs logiques (6 types d'opérateurs booléens), de comparaison (5 types d'opérations), lexicaux (cheminement et coïncidence) et spécifique (opérateur MEM : identité entre les classes syntaxiques).

Cet essai de description, très rapide, du lexique et de la syntaxe qui régissent le métalangage du système SATIN 1, donne une idée de la complexité d'élaboration à la fois des outils eux-mêmes et des informations d'entrée.

Ce dernier travail est présenté comme "lourd" et "fastidieux" par les chercheurs qui ont appliqué le syatème à l'étude des nématodes (1).

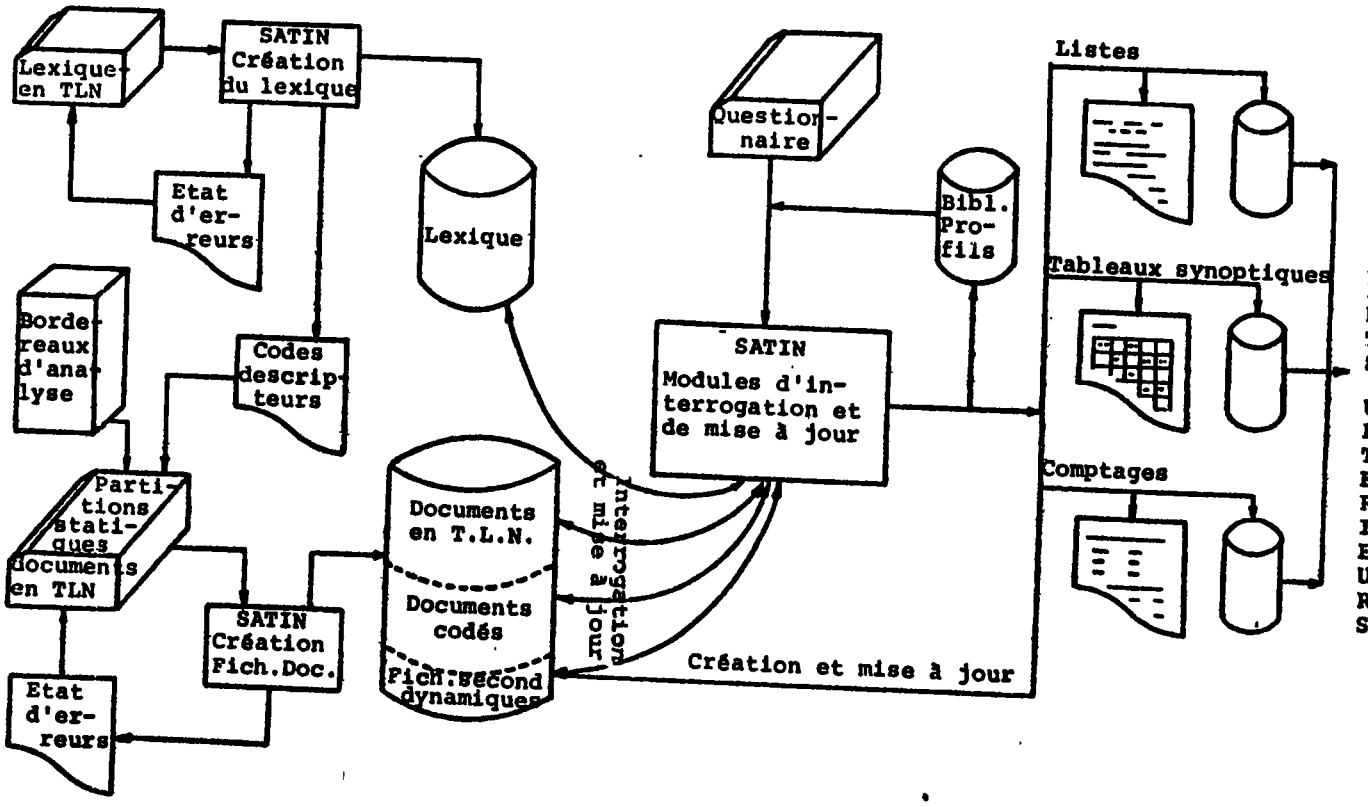
Ils précisent que SATIN 1 demande une "connaissance très fine du corpus de donnée que l'on souhaite analyser" et que son apprentissage est "long" et "délicat".

Ils considèrent toutefois que "c'est un produit adapté à une recherche bien délimitée".

---

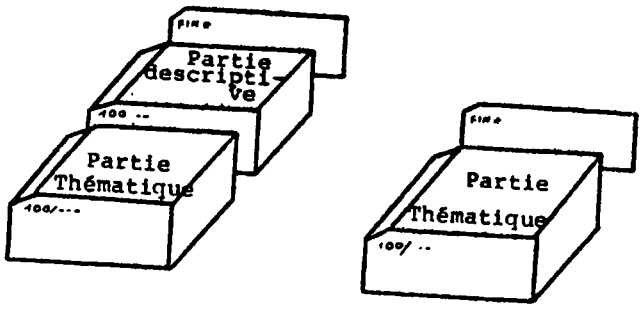
(1) cf. l'article de BRANCA-LACOMBE et TOMASSONE sur la constitution d'une banque de données en biologie.

Les grandes fonctions du système SATIN 1



(TLN : Termes du langage naturel)

(cf. le ouvrage et article de CHOURAQUI et BOURELLY)



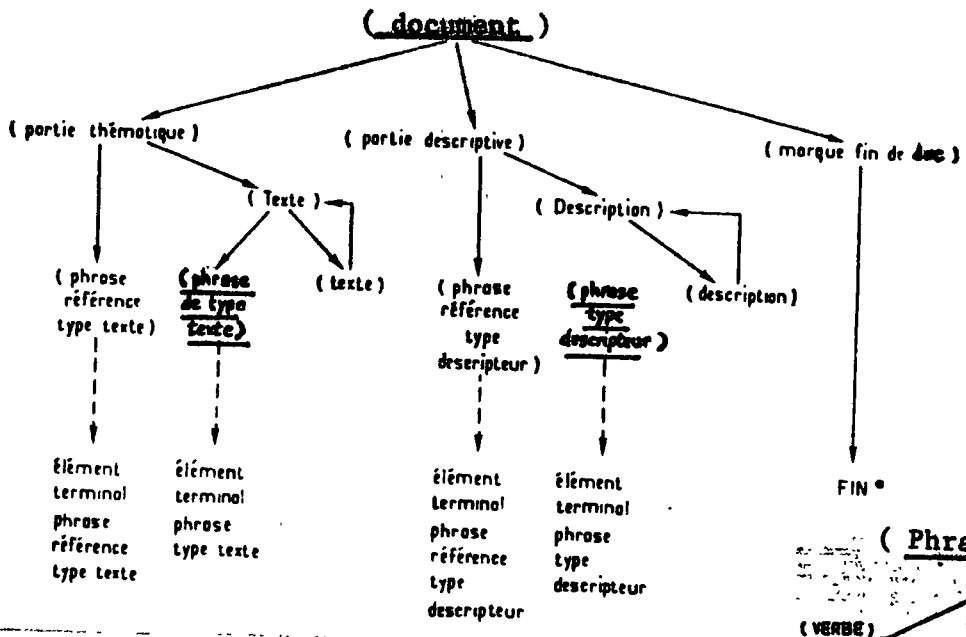
Organisation du document

Exemple de codification de la partie descriptive :

1	+	Groupe-pression	
2		Groupe volontaire voisinage	
3	+	Aéroport national	= 5
4		Bruit	= 3
5		Aéroport de Toussus le Noble	= 3

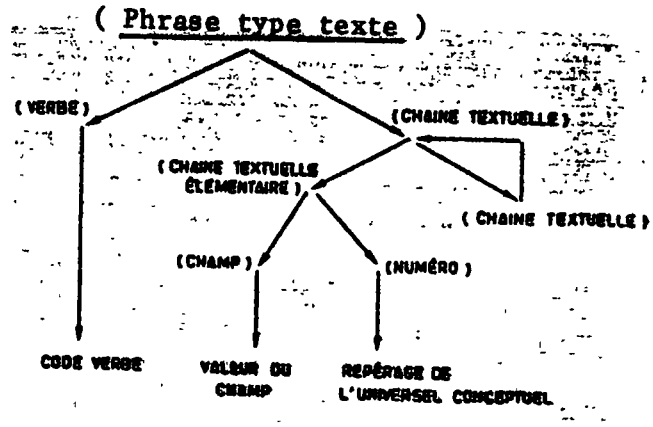
(cf. CHAUNIER, p. 79)

Règles de transformation de la génération d'un document



Articulations d'un "document"

cf. BOURELLY CHOURAQUI



Structuration des informations dans un "document":

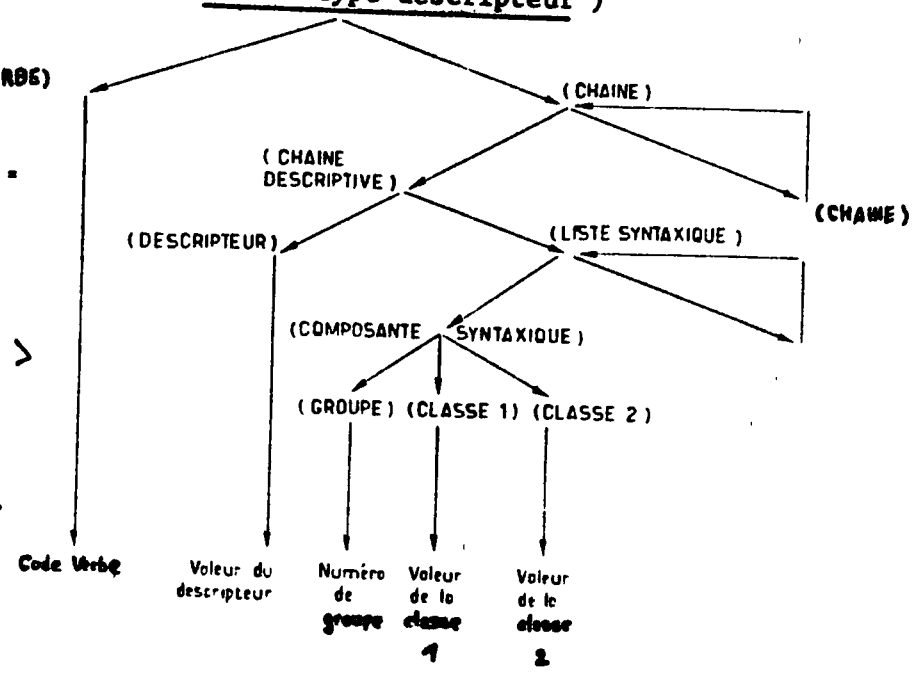
'hrase référence de type texte> : =
)de verbe 100 <chaîne textuelle>

exte> : = <phrase de type texte> |
hrase de type texte><texte>

hrase référence de type descripteur> : =
de verbe 100 <chaîne>

escription> : =
hrase de type descripteur> |
hrase de type descripteur><description >

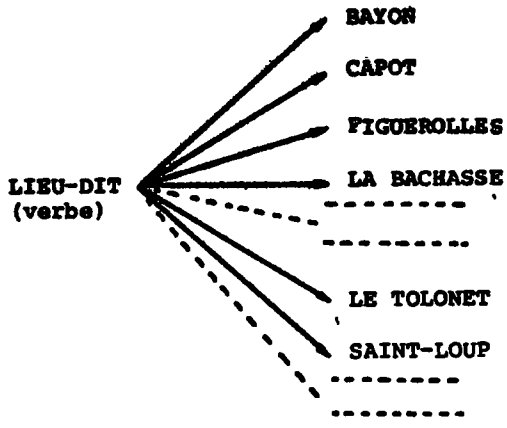
artie thématique> : =
hrase référence de type texte> |
hrase référence de type texte> <texte>



artie descriptive> : =
hrase référence de type descripteur> |
hrase référence de type descripteur><description>

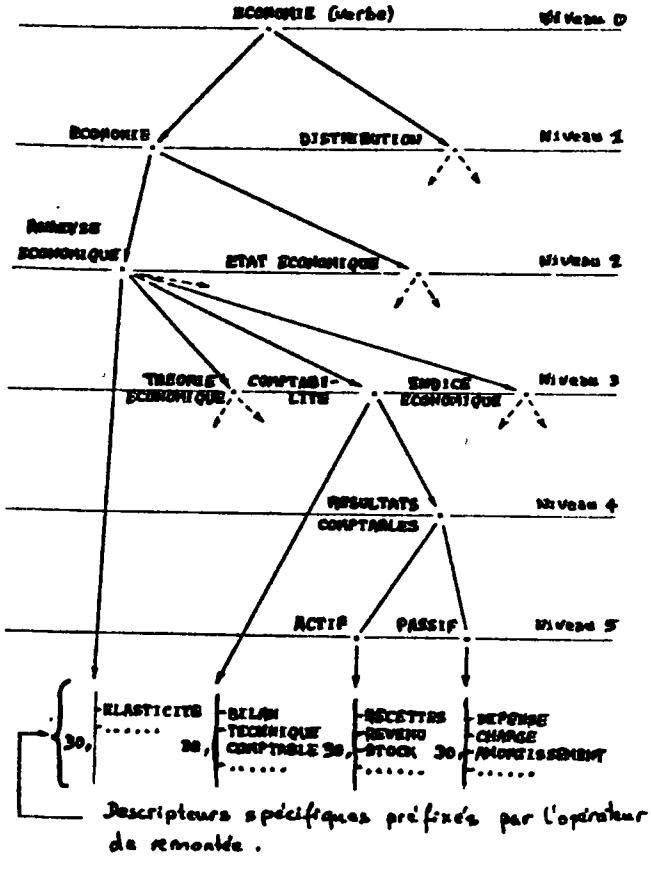
)rque de fin de document> : = FIN•

cf. l'étude du CXP sur les logiciels ...



exemple de catégorie non hiérarchisée

cf. CHOURAQUI et BOURELLY, vol. 1



exemple de catégorie hiérarchisée

Liste des produits de sortie :

— QSVP : édition de références pour les documents pertinents issus d'une recherche ;
— CPTG : comptage des documents pertinents ;
— GRIL : création d'une TABLE SYNOPTIQUE et édition des références associées à un thème de recherches ;
— CPTH : comptage des documents sélectionnés sur un thème ;
— IDEX : édition de références sous forme d'un index hiérarchisé ;
— FREF : édition de références sans recherche préalable ;
— EXTR : extraction de descripteurs ou de valeurs algébriques avec édition des références ;
— EXTG : extraction de descripteurs ou de valeurs algébriques sans édition des références ;
*** : comptage des documents pertinents et sauvegarde d'un fichier permettant de retrouver ces documents lors d'une utilisation ultérieure ;

IV. 2 LE SYSTEME TITUS II

TITUS II, modèle de deuxième génération, apporte à l'automatisation d'un système classique (description par mots-clés et recherche) des possibilités nouvelles dues à l'utilisation d'un "langage documentaire canonique" (LDC), en particulier l'indexation et la traduction automatiques. Il permet également une codification spécifique pouvant aboutir à la création d'une banque de données numériques.

Le LDC est basé sur la constatation qu'une douzaine de structures de phrases peuvent couvrir un grand nombre de résumés analytiques dans une discipline donnée, scientifique ou technique.

Ainsi, les résumés qui servent de données d'entrée au système vont être rédigés selon des structures standards, que les analystes sont amenés à "remplir" de mots-clés et de mots-outils.
Les mots-clés sont reconnus automatiquement par l'ordinateur (au moment de la fusion avec la base des données d'entrée validées), ce qui correspond à la phase d'indexation automatique, qui s'accompagne d'un processus de pondération, de paramétrage et de chaînage.

Le langage documentaire est constitué d'une série de lexiques en 4 langues, ainsi que d'une syntaxe, qui sont codifiés en un "langage-pivot" destiné à faciliter les procédures de la traduction automatique.

C'est en ce langage-pivot (ou encore langage interne) que sont condensées, en effet, les phrases d'entrée, qu'elles soient en français, anglais, allemand ou espagnol, après des opérations de consultation des dictionnaires et d'analyse syntaxique programmées.

Un traducteur automatique de sortie reconstitue par la suite les structures de phrases correspondant à la langue désirée, à l'aide également d'un analyseur syntaxique et de grammaires programmées (voir l'organigramme de TITUS II sur le schéma p. 1).

Les lexiques comprennent des "descripteurs", des "non-descripteurs" (destinés à réduire les polysémies) et des mots-outils, l'ensemble formant une suite d' "unités lexicales", avec l'enregistrement de la caractérisation de leurs formes singulier et pluriel et un "code article" correspondant aux valeurs syntaxiques que peut avoir le mot dans la phrase (1).

Aux descripteurs - mots ou groupes de mots - se trouvent également associés, sur l'axe paradigmatique, des "mots de voisinage".

Les "syntagmes standards" (au nombre d'une quinzaine) sont composés de 2 à 4 "groupes syntagmatiques" et contiennent deux "actants", qui sont des prépositions ou des locutions prépositives (au nombre de 33). Les différentes combinaisons possibles donnent un éventail de 550 structures de phrases différentes.

La syntaxe est introduite au niveau de l'Unité lexicale, qui comprend trois zones : AR pour le code article, NR pour le pluriel et LI pour la valeur grammaticale du mot suivant.

Ce système nécessite, comme le précédent, la rédaction très rigoureuse des bordereaux d'analyse. Cependant l'introduction des informations peut se faire par carte perforée ou par écran cathodique, avec contrôle et correction en temps réel.

Le traitement des informations aboutit à la constitution de fichiers inversés auxquels on peut accéder par 7 options de sélection différentes.

Différents produits de sortie caractérisent enfin ce système documentaire, qui a l'avantage d'être opérationnel, y compris sur le plan international, grâce à sa dimension multilingue.

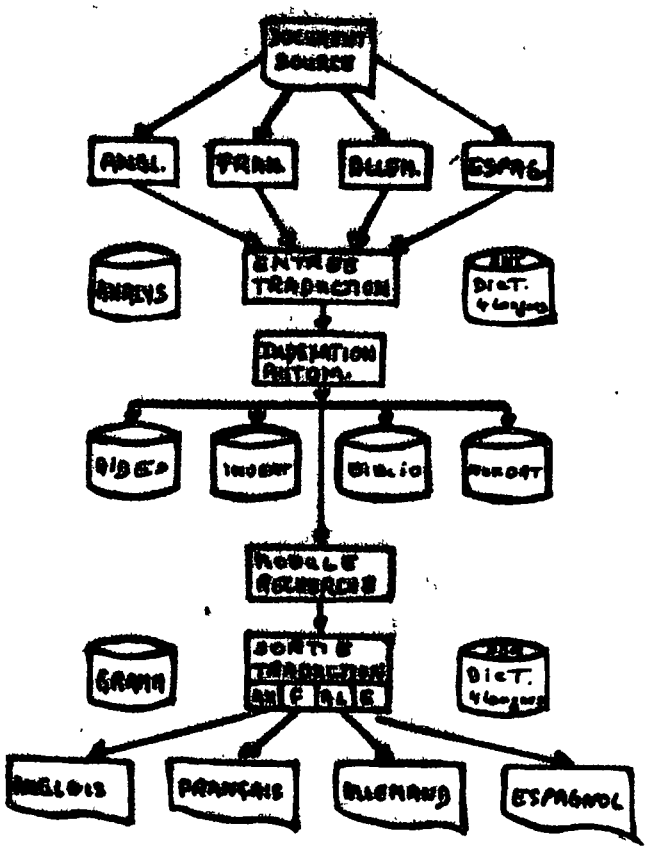
Une version TITUS IV (2) doit rendre possible la recherche en texte libre et se rapprocher, dans sa langue d'indexation et de condensation, le plus possible de la langue naturelle.

---

(1) voir p.1 du schéma

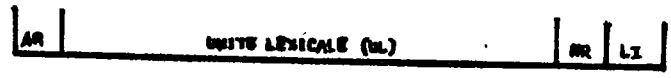
(2) cf le compte rendu d'une communication faite par J.M. Ducrot au Deutscher Dokumentarverlag 1977 paru dans Nachrichten für Dokumentation, 28, 1977, n° 6 p. 226

Organigramme de TITUS II



cf. ZINGEL

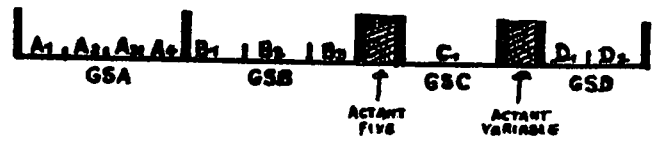
Structure du mot en LDC



AR = code unique article ou facteur immédiat
UL = Mot-clé ou Mot-outil
NR = Code nombre singulier ou pluriel
LI = Code fonction syntaxique de l'UL suivante

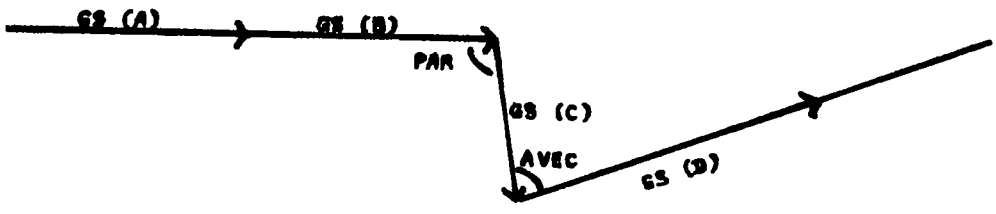
LDC : Langage documentaire canonique.

Composition des syntagmes standards



GS = Groupe syntagmatique
A_1, A_2, D_2 = Unités lexicales (UL)

Exemples d'expressions autorisées par le syntagme standard PAR avec flexion G :



cf. article DUCROT

Amélioration de la teinture du Nylon PAR traitements avec solvants
Amélioration,#teinture/#Nylon, traitement+,solvant+.
←GSA→ ←GSB→ ←GSC→ ←GSD→

Procédés de filature de la laine PAR voie pneumatique avec turbine
Procédé+,Filature/#laine, voie pneumatique,turbine.
←GSA→ ←GSB→ ←GSC→ ←GSD→

Les méthodes de confection PAR couture avec des fils invisibles
#Méthode+/confection,,couture,#fil invisible+.
←GSA→ ←GSC→ ←GSD→

Exemples de phrases TITUS

Développement aux USA d'un fil mixte polyester triacétate pour le tricotage des vêtements de nuit

BUT 7 Développement * aux U.S.A., un fil mixte * polyester * triacétate, * tricotage, * vêtement de nuit +

La période actuelle n'est pas un phénomène nouveau, en zone soudano-saharienne

NES B * Période actuelle, un phénoméne nouveau, zone soudano-saharienne
(en)

Les problèmes de la sécheresse dans l'Afrique de l'Ouest

SIM A * Problème +, * sécheresse, * Afrique de l'Oueet
(dans)

Etude du climat, de la pédologie et de la végétation en Côte-d'Ivoire.

SIM B Etude, * climat) * pédologie) * végétation, Côte-d'Ivoire
(En) énumération énumération

J. CHAUMIER, p. 78

Exemples de sorties en français et en anglais

FRANCAIS

```
DOC.NR 40002
LA TEINTURE EN MILIEU SOLVANT _ LE PROCEDE STX
IND.TEXTILE/07/1971-FASC.1003-P.527(0002P.)-005FIG.-
VAL.DOC.=2-VAL.SCIENT.=1-FRANCE-

LA NOUVEAUTE CONCERNE 1 PROCEDE DE TEINTURE AVEC SOLVANT '
STX'.LE PROCEDE PERMET LA TEINTURE SANS POLLUTION D'EAU.

LE PROCEDE A LA CONTINUE DE RECUPERATION DU SOLVANT EVITE LA
DISTILLATION DE LA TOTALITE DU BAIN DE TEINTURE.LA TECHNIQUE
DIMINUE LE PRIX DE REVIENT ET LA DUREE D'OPERATION.
POSSIBILITES D'AUTOMATISATION ET DE DIVERSIFICATION.LES
RESULTATS DES ESSAIS MONTRENT LA QUALITE DE LA
REPRODUCTIBILITE DE LA COULEUR.LA COMPOSITION DU BAIN DE
TEINTURE COMBINE 1 MELANGE HOMOGENE D'AGENT DILUANT AVEC 1
SOLVANT ET LES COLORANTS.1 SCHEMA DU PROCEDE DONNE 1 EXEMPLE
CONCERNANT LES TEINTURES DES POLYAMIDES.

MARQU.COMMERC.=STX
FABR.=GILET-THAON_PROGIL_FMC
```

ANGLAIS

```
DOC.NR 40002
LA TEINTURE EN MILIEU SOLVANT _ LE PROCEDE STX
SOLVENT DYEING - STX PROCESS
IND.TEXTILE-07/1971-ISSUE NR.1003-P.527(0002P.)-005FIG.-
DOC.VAL.=2_SCIENT.VAL.=1_FRANCE_

THE INNOVATION CONCERNS ONE SOLVENT DYEING PROCESS 'STX'.THE
PROCESS ENABLES THE DYEING WITHOUT WATER POLLUTION.

THE CONTINUOUS PROCESS OF SOLVENT RECOVERY AVOIDS THE
DISTILLATION OF THE DYE BATH TOTALITY.THE TECHNIQUE
DECREASES THE COST PRICE AND PROCESSING TIME.POSSIBILITIES
OF AUTOMATION AND DIVERSIFICATION.THE RESULTS OF THE TESTS
SHOW THE QUALITY OF THE COLOR REPRODUCTIBILITY.THE
COMPOSITION OF THE DYE BATH COMBINES ONE THINNER HOMOGENEOUS
BLEND WIHT ONE SOLVENT AND DYES.ONE PROCESS DESIGN GIVES ONE
EXAMPLE ON 8 POLYAMIDE DYEINGS.

TRADE NAME.=STX
MANUFACT.=GILET-THAON-PROGIL-FMC
```

Cf. DUCROT

V SYSTEMES TRAITANT LE LANGAGE NATUREL

V EXEMPLES DE SYSTEMES DE TRAITEMENT DES INFORMATIONS EN LANGAGE NATUREL

C'est surtout dans le domaine du traitement des langues naturelles que se développent actuellement recherches et expérimentations, ce que rend possible l'évolution rapide de la technologie des ordinateurs (meilleures capacités de stockage et de traitement).

L'objectif de départ est, dans ces systèmes:

- d'une part, de faire l'économie du travail d'analyse et d'indexation, puisque l'on fournit à l'ordinateur, en langage naturel, la totalité des textes à traiter, ce qui devrait entraîner,
- d'autre part, la disparition des distorsions et du manque d'uniformité des analyses, dus en grande partie à la subjectivité de l'intervention humaine (1).

Il faut toutefois remarquer que les textes admis en entrée sont encore relativement courts et qu'il s'agit d'ailleurs souvent d'analyses ou d'abstracts (ou de textes de dépêches de presse), ce qui signifierait que ce sont surtout le travail d'indexation et ses inconvénients que l'on évite à l'heure actuelle.

La mise au point des traitements correspondants entraîne cependant des investissements préalables très importants, puisqu'il faut, par des programmes adéquats, assortis d'une quantité non négligeables d'informations sémantiques, syntaxiques et grammaticales, rendre l'ordinateur capable d'effectuer ces tâches, qui

---

(1) "Faisant indexer le même texte par des analystes différents, on constate que le taux moyen de cohérence de l'indexation ne dépasse pas 50 %"
cf. MANIEZ J.- Le Rôle de la syntaxe dans les systèmes de recherche documentaire.- Dijon: IUT Carrières de l'information, 1976.- 2 vol.

restent éminemment complexes.

Les systèmes dont la description suit sont encore plus ou moins en phase d'expérimentation; il semble que l'on puisse les considérer comme des exemples de réalisations futures qui, lorsqu'elles deviendront opérationnelles, pourront entraîner un nouveau "bond en avant" dans le domaine des Sciences de l'Information

V. 1 SYNTAXEME

Ce système, qui se base sur les théories linguistiques des grammaires générative et transformationnelle, veut également introduire la "psycholinguistique" et les théories logiques relatives à la présupposition.

Son objectif initial est de supprimer l'étape d'indexation dite "manuelle" en effectuant un traitement direct de textes rédigés en français; il s'agit en l'occurence, pour les deux expériences décrites (1), de résumés à caractère scientifique.

Les traitements visent à réduire le texte " à la description de sa structure profonde"(2) au moyen d'un analyseur morphologique (exploration d'un lexique de formes et d'une table des désinences) et de différentes opérations syntaxiques : analyse syntaxique aboutissant à des descriptions structurales, édition des liaisons grammaticales unissant les termes pris 2 à 2, chargement de ces liaisons dans un fichier inversé.

Cependant, l'originalité du système réside dans les procédures de consultation, en mode conversationnel, qui - selon les auteurs - aboutissent à l' "exclusion totale du bruit documentaire" (3) :

---

(1) la 1ère concerne un corpus restreint de jurisprudence (relatif à la filiation - cf. SOLET
dans le second cas, il s'agit d'une recherche menée pour le compte de la DRME - cf. STEHLIN (et où la nature du corpus n'est pas précisée).

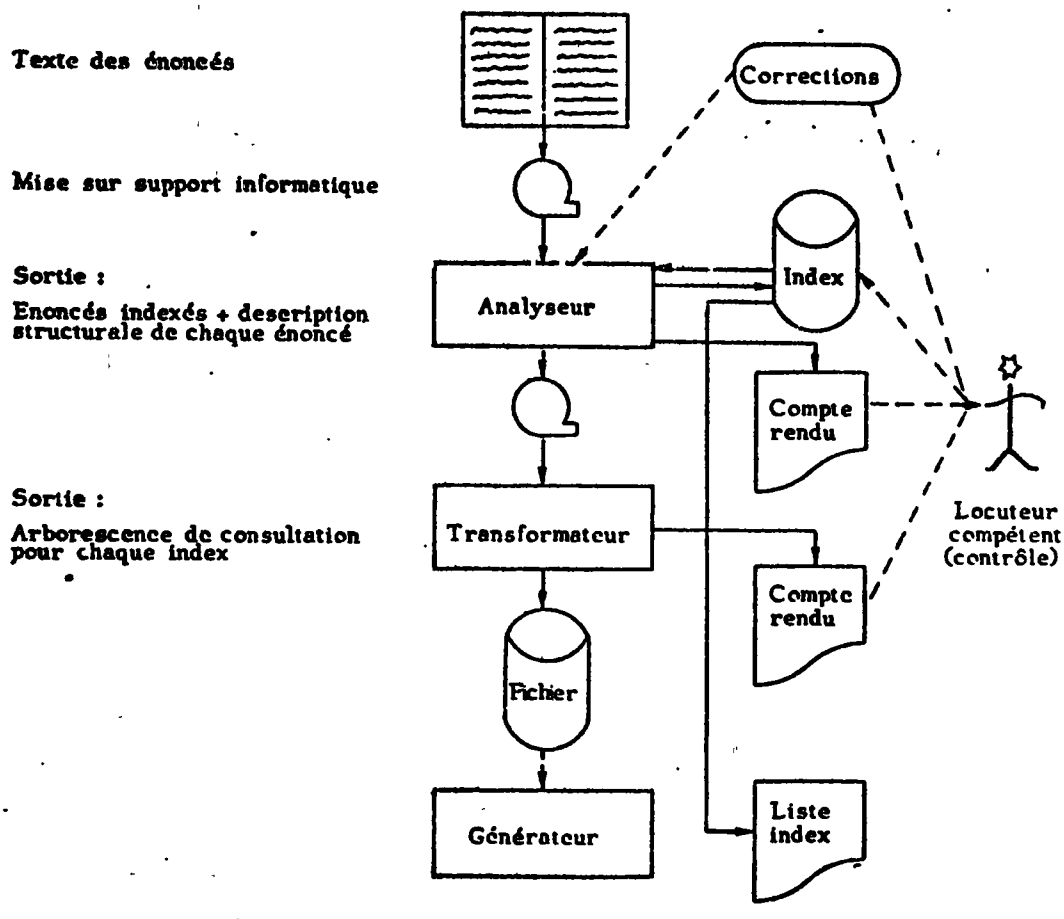
(2) On peut se demander s'il y a analogie avec les "cas profonds" mis en évidence dans le système PRECIS ...

(3) cf. SOLET

en fonction d'un "thème" en entrée, le système génère des questions qui représentent les différentes directions possibles de recherche (1 et qui sont directement dépendantes des informations intégrées dans le système (cf. l'exemple du schéma).

Les auteurs envisagent une évolution de ce système dans le domaine multilingue : possibilité multilingue de consultation et possibilité de multitraduction de la réponse (cf. STEHLIN).

(1) Ce type de cheminement à l'interrogation apparaît également dans le système ARIANE, mis en oeuvre sur la banque de données du CATED (Centre d'assistance technique et de documentation) dans le domaine du bâtiment.



Dessin de chaine du "noyau" de SYNTAXEME

Cl. SOLET

Exemple d'interrogation

```
  VOULEZ-VOUS UNE RECHERCHE AVEC THESAURUS - (O-N)
>N
  **VEUILLEZ SOUMETTRE VOTRE REQUETE**
>PROPRIETES.
  1 S'AGIT-IL D'UNE PROPRIETE THERMIQUE
  2 S'AGIT-IL DE LA PROPRIETE D'UNE FONDERIE
  3 S'AGIT-IL DE LA PROPRIETE D'UNE SUSPENSION
  4 S'AGIT-IL DE LA PROPRIETE D'UN HYDRURE
  5 S'AGIT-IL D'UNE PROPRIETE ANEOLOGIQUE
  6 S'AGIT-IL DE LA PROPRIETE DE SOLIDES
  7 S'AGIT-IL D'UNE PROPRIETE IMMUNOLOGIQUE
  8 S'AGIT-IL D'UNE PROPRIETE MECANIQUE
  9 S'AGIT-IL DE LA PROPRIETE DE DECHARGES
 10 S'AGIT-IL D'UNE PROPRISTE DONNE
 11 S'AGIT-IL D'INFLUENCER DES PROPRIETES
 12 S'AGIT-IL DE L'ETUDE DE PROPRIETES
 13 S'AGIT-IL DE LA CARACTERISATION DE PROPRIETES
 14 S'AGIT-IL D'UNE DONNEE SUR DES PROPRIETES
 15 S'AGIT-IL D'UNE RELATION ENTRE DES PROPRIETES
>5.
  *** FICHE NO =  74020200
  VOULEZ-VOUS L IMPRESSION DES FICHES RETENUES - (O,N)
>O
  ** FICHE NO:   74020200
  ETUDIER LES DIFFERENTS PARAMETRES INFLUENÇANT LES PROPRIETES
  RHEOLOGIQUES DE LA SUSPENSION SANGUINE.

  VOULEZ-VOUS REVENIR A L ETAPE ANTERIEURE - (R),
       OU EFFECTUER UNE AUTRE RECHERCHE - (O,N) ?
>O
  VOULEZ-VOOS UNE RECHERCHE AVEC THESAURUS - (O-N)
```

Cl. STEHLIN

V. 2 LE SYSTEME A L'ETUDE AU CENTRE D'INFORMATIQUE JURIDIQUE (CEDIJ) ET A L'INSTITUT NATIONAL DES SCIENCES ET TECHNIQUES NUCLEAIRES (INSTN/ CEA)

C'est le groupe de recherche en linguistique automatique, dont les travaux dans le domaine du traitement automatique du langage sont variés et nombreux (1), qui a mis au point, en collaboration avec le CEDIJ, un syetème d'indexation automatique à partir du texte intégral, dont une version est actuellement en démonstration à Saclay.

Il s'agit de textes juridiques, sur lesquels sont effectués des traitement linguistiques et statistiques.

La méthode proposée par le groupe (2) comprend la construction automatique d'un thesaurus et d'une liste de mots vides; le thesaurus est constitué :

- d'une liste de descripteurs (avec règles de reconnaissance de l'homographie, listes des termes génériques, spécifiques, synonymes et parents, fonctions de poids sémantique)

- d'un système de reconnaissance des mots composés.

Une série de programmes d'analyse linguistique et statistique ont été élaborés, comprenant :

/ une analyse morphologique (signalant notamment les erreurs typographiques)

/ une analyse syntaxique (syntaxe construite automatiquement à partir d'une méthode d'apprentissage) permettant de lever les ambiguïtés grammaticales, suivie de pro-cessus de filtrage de mots grammaticalement vides.

---

(1) Je remercie vivement M. Andreewsky pour l'ensemble des documents qu'il a très aimablement accepté de me communiquer et dont ceux que j'ai principalement utilisés sont cités dans la partie bibliographique.

(2) cf. la communication IRIA- atelier SESORI

/ une analyse statistique, avec reconnaissance des homographies, des synonymes, des termes génériques et spécifiques et de la parenté sémantique (traitement de champs sémantiques partiels)

/ des programmes de calcul de la fonction de poids sémantique.

Les expériences menées sur les textes juridiques portent particulièrement sur l'indexation automatique.

Les questions sont posées en langage naturel, sur le mode conversationnel, et font apparaître les documents " par ordre de proximité décroissante par rapport à la question ".

D'autres expériences, menées sur un " corpus considérable " (1), amènent les auteurs à la conclusion que l'indexation automatique est plus "souple" et plus "objective" que l'indexation manuelle et qu'elle donne des réponses " dans la plupart des cas meilleures que les réponses aux systèmes indexés manuellement ".

(1) cf. note CEA-N-1795.

CENTRE D'INFORMATIQUE JURIDIQUE

UNE ÉQUIPE DE RECHERCHE :

Parallèlement à l'exploitation du système DOCILIS qu'il a créé, le CEDIJ poursuit ses travaux de recherche, théorique et appliquée, en collaboration avec des organismes extérieurs.

Dans le cadre du SICOB, sont présentés deux thèmes de recherche qui ont abouti à des résultats exploitables.

Ces deux recherches sont menées conjointement par le CEDIJ et l'INSTN (CEA), dans le cadre de l'ERA n° 430 du CNRS. Leur objectif est de contribuer à l'automatisation maximum du processus de traitement des textes, notamment en ce qui concerne la saisie, l'édition, l'indexation et l'interrogation.

Les travaux entrepris ont été expérimentés sur les textes juridiques, une telle automatisation s'y révélant particulièrement nécessaire.

La détection et la correction automatique d'erreurs typographiques et le système documentaire statistique à indexation automatique et interrogation en langue naturelle sont les deux produits qui font l'objet d'une démonstration.

I - DETECTION ET CORRECTION AUTOMATIQUE D'ERREURS TYPOGRAPHIQUES.

Si les erreurs typographiques n'ont de conséquences néfastes que sur l'esthétique des textes imprimés, elles peuvent en avoir de dangereuses dans les systèmes documentaires utilisant les textes comme source d'information. Elles sont en effet de nature à provoquer des "silences" (documents pertinents non fournis en réponse par la machine) au moment de l'interrogation.

La détection-correction d'erreurs, proposée au cours de la démonstration, a pour but de traiter les erreurs qui demandent la plus grande attention aux correcteurs humains. Un tel système est donc de nature à les soulager d'une tâche astreignante en leur permettant de concentrer leurs efforts sur la correction de fautes plus subtiles dont la détection par le système serait trop coûteuse.

Pour chaque mot erroné qu'il rencontre, et à condition que l'erreur n'engendre pas un mot existant, le système propose une solution. L'action du correcteur humain se réduit donc, pour ce type de fautes, à la correction des erreurs que peut commettre parfois le système automatique.

2 - LE SYSTEME DOCUMENTAIRE STATISTIQUE A INDEXATION AUTOMATIQUE ET INTERROGATION EN LANGUE NATURELLE.

Le système présenté en démonstration est l'élément de base d'un système documentaire entièrement automatisé. Il intégrera progressivement tous les traitements linguistiques qui ont été mis au point par ailleurs.

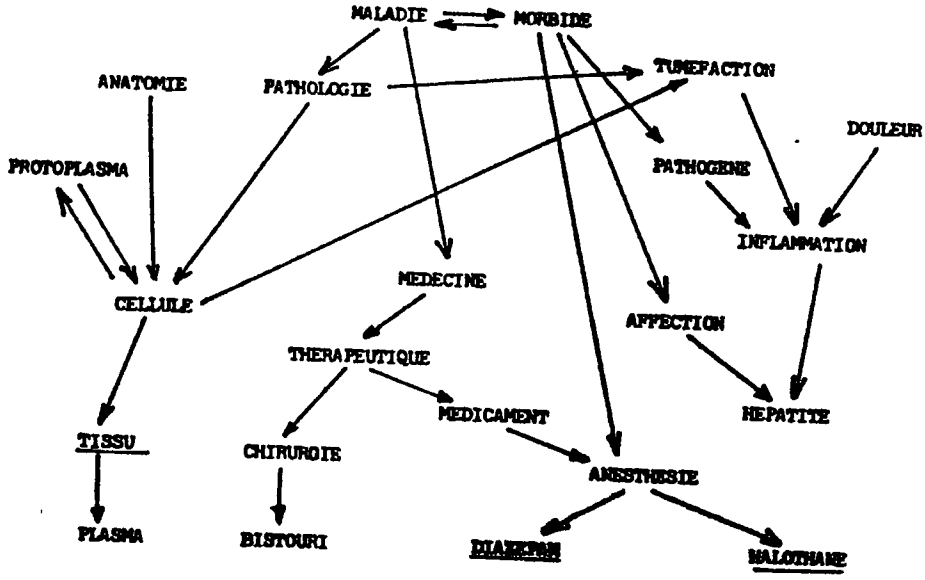
Le système est construit sur un modèle mathématique Bayesien qui a été modifié pour établir la relation entre fréquence et sémantique. Le fonctionnement en est le suivant :

- Le système d'indexation établit pour chaque document de la base de données la liste des unités linguistiques discriminantes, associées à des données statistiques, qui serviront à l'interrogation.

- Le système d'interrogation considère la question comme un document ; il l'indexe et établit une distance entre elle et les différents documents de la base. La réponse est constituée par la liste des documents classés par ordre de proximité avec la question. Il est ensuite possible de visualiser les documents sur écran, afin de vérifier la pertinence des réponses.

  Enfin, on peut poser un "document-réponse" en question, ce qui simule d'une certaine manière la notion d'association d'idées, et est particulièrement utile dans certaines recherches exhaustives, par exemple, en jurisprudence.

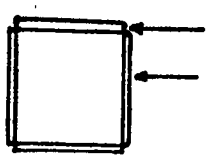
Modèle de graphe servant au calcul des fonctions de poids, à partir d'un dictionnaire de spécialité.



cf. note C -N-1694(1)

Méthode de reconnaissance des synonymes, des relations spécifique - générique et de parenté sémantique.

- synonymie



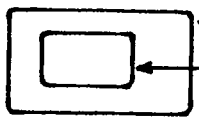
champ sémantique S_i du concept C_i

champ sémantique S_j du concept C_j

Si $S_i \cap S_j \sim S_i \sim S_j$ alors $C_i \sim C_j$

et on a une relation de synonymie.

- spécificité-généricité



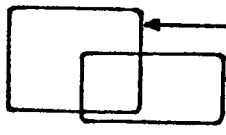
Champ sémantique S_i du concept C_i

Champ sémantique S_j du concept C_j

Si $S_i \cap S_j = S_j$ et donc $S_j \subset S_i$,

alors C_j est terme spécifique de C_i

- parenté sémantique



Champ sémantique S_i du concept C_i

Champ sémantique S_j du concept C_j

C_i et C_j sont dans une relation de parenté

car $S_i \cap S_j \neq \emptyset$.

cf. note CEA-N-1795

V. 3 PIAFDOC

Le syetème PIAF (Programmes Interactifs d'Analyee du Françaie) a été élaboré à Grenobla, par l'équipe "Intelligence artificielle" de l'IMAG, qui travaille actuellement à la miee au point de son application - sous le nom de "PIAFDOC" - sur la base des "dépêchas de presee" de la Documentation françeise, en liaieon avee le logiciel d'interrogation MISTRAL(1).

Le système, qui fonctionne sur le mode convereationnel, traite les informations en langue naturelle (il s'agit à la foie dee documents et des questions) en effectuant :

- un contrôle de saisie, avec détection de fautes d'orthographe et possibilité de correction immédiate
- une indexation automatique assietée, pour l'extraction de mots-clés.

Il associe (2) d'une part des "modules de programmes" (éditeur, composante linguietique - ou "ànalyeeur morphologiqne" -, composante documentairs - ou "Générateur de mots-clés") et, de l'autre, des paramétres linguietiques (dictionnaire de racines, suffixee et terminaieons, grammaire oompoeée de règles d'étata finid et de modèles).

Le "fichier dictionnaire" se compose d'informationa sous forms de chaînes de caractères associéee à des pointeurs chaque chaîne est suivie d'un modèle morphologique et d'un ou plueieurs pointeurs

---

(1) voir p. 1 dee schémas
(2) cf. GRANDJEAN (Projet PIAF)
(3) voir p. 4 dee schémae : extrait d'un état du dictionnaire

- ou indicateurs - suivante :

(0)	mot vide	exemples : de (0) le (0)
(*)	chaîne indécoupable	POMME DE TERRE (*)
(-)	mot à relier (élément de mot composé)	PROJET (*)(-)
(S)	existence de synonymes	SANS EMPLOI (*)(S)
(D)	détail	PROJET DE REFORME (*)(D)
(?)	mot multisens	CONCORDE (*) (?)

Un indicateur spécial, l'"indicateur de rémanence", qui n'existe qu'en code interne, est destiné à alléger le dialogue lors des demandes de choix dans la mesure où il mémorise les choix qui ont été faits une fois pour un cas d'homographie ou de polysémie (1).

Les homographes, ainsi que les différentes racines d'un verbe irrégulier, sont chaînés entre eux.

Les "MOTS MULTISENS" sont suivis de parenthèses explicatives qui sont numérotées lors d'une interruption réclamant un choix de l'analyste; par exemple :

CONCORDE (?)
1 CONCORDE (LA PLACE)
2 CONCORDE (AVION)
3 CONCORDE (HOTEL)
4 CONCORDE (PAIX)

A chaque entrée de phrase, le système met en route l'analyseur morphologique qui provoque des interruptions :

/ soit lorsque les chaînes ne sont pas reconnues, ce qui se produit en cas de faute d'orthographe, ou à la rencontre d'un mot nouveau

/ soit en cas de polysémies ou homographies; on obtient alors la liste des différentes solutions possibles.

---

(1) cf. GRANDJEAN (Projet PIAF) : on tient compte du fait que le contexte est généralement constant dans un document - et on suppose également "une certaine continuité des événements relatés dans la base" (p. 23)

Dans les deux cas, l'analyste prend une décision; il dispose pour cela de différentes commandes lui permettant :

- des corrections (ORTHO)
- l'introduction de mots nouveaux (DICT)
- le choix d'une des solutions proposées (MCL suivi du numéro de la solution choisie)
- l'ajout d'un complément à un mot-clé (ou à une chaîne) ou bien dans le texte même d'entrée (MCL + suivi du complément, ou TEXT +)

Lorsque la phrase est entièrement analysée, il reste à valider (ou à refuser de valider si l'on souhaite apporter des modifications) les mots-clés proposés par le système.

L'interrogation se fait en langue naturelle par le logiciel MISTRAL V 3, dont les performances sont cependant améliorées, puisqu'il n'est plus nécessaire :

- d'opérer sur des troncatures (on a une réduction automatique des mots à leur "modèle" par le programme)
- d'utiliser les opérateurs booléens pour reconstituer les mots composés
- d'utiliser les indicateurs de distance, en cas de recherche sur critères secondaires.

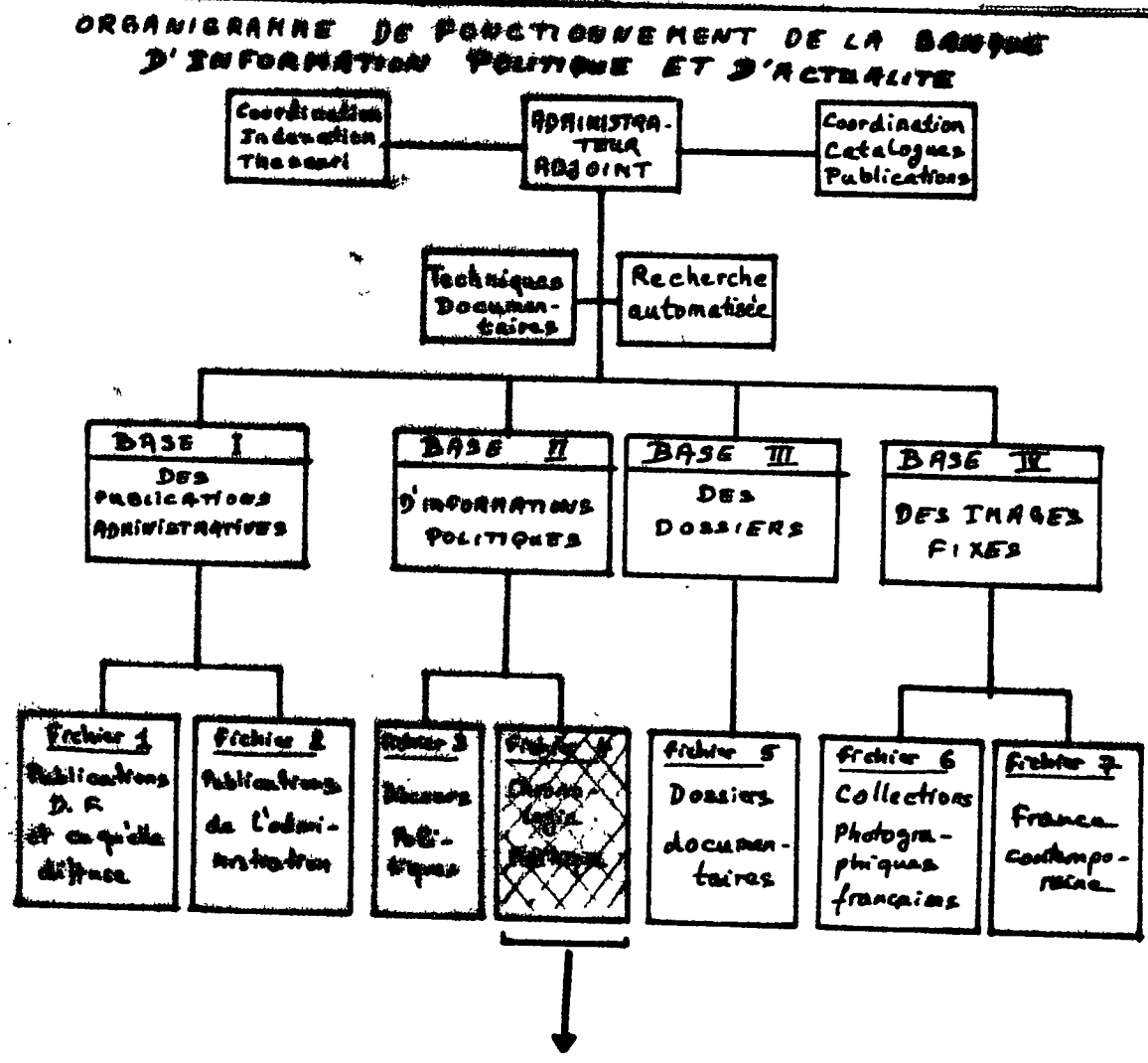
Les expérimentations se heurtent actuellement au problème de la reconnaissance des mots composés non connexes (exemple : "projet très controversé de réforme"), ainsi qu'à celui, plus général, de la tendance à la multiplication des mots composés, ce qui aboutit à élargir considérablement le dictionnaire et alloage les temps d'indexation et de recherche.

Les recherches actuelles (1) s'orientant en particulier:

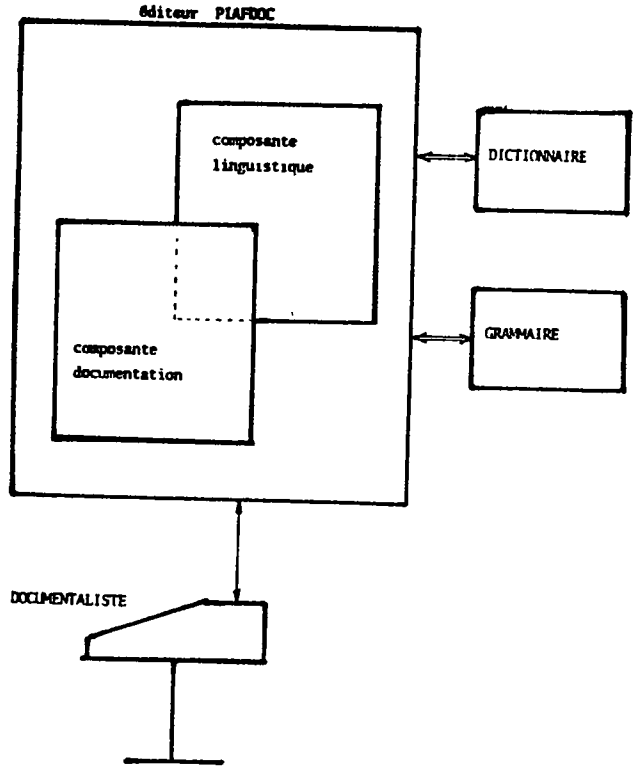
- vers la réalisation d'algorithmes permettant de lever automatiquement les homographies
- vers la définition de classes - et donc d'une grammaire - de mots composés
- vers la réalisation d'une analyse syntaxique des questions - ce qui devrait permettre une comparaison avec les structures syntaxiques des documents sélectionnés, pour une meilleure adéquation des réponses.

PIAFDOC, enfin, est présenté comme un système ayant "un taux de rappel élevé" (2).

(1) cf. GRANDJEAN (Projet PIAF)
(2) cf. GRANDJEAN (communication IRIA - atelier SESORI)



PIAFDOC :



- **un programme interactif de saisie avec :**

. détection de fautes d'orthographe et possihilité de les corriger immédiatement
. demande d'assistance lorsqu'un mot à sens multiple a besoin d'une indication de contexte
. contrôle de l'indexation dès sa création

- **un programme d'indexation automatique permettant :**

. le repérage des mots-clés (y compris les mots composés)

. l'écriture du mot-clé sous une forme normalisée

. la substitution d'un mot-clé non retenu par un mot-clé appartenant au thésaurus

. la surindexation d'un mot-clé non introduit dans le thésaurus par son générique y figurant.

ultérieurement :

- **un programme d'analyse automatique des variantes** syntaxiques (mots disjoints par exemple) et **des** contextes

- **un programme d'analyse automatique des questions qui seront** posées par l'utilisateur.

Diagramme du Système PIAFDoc

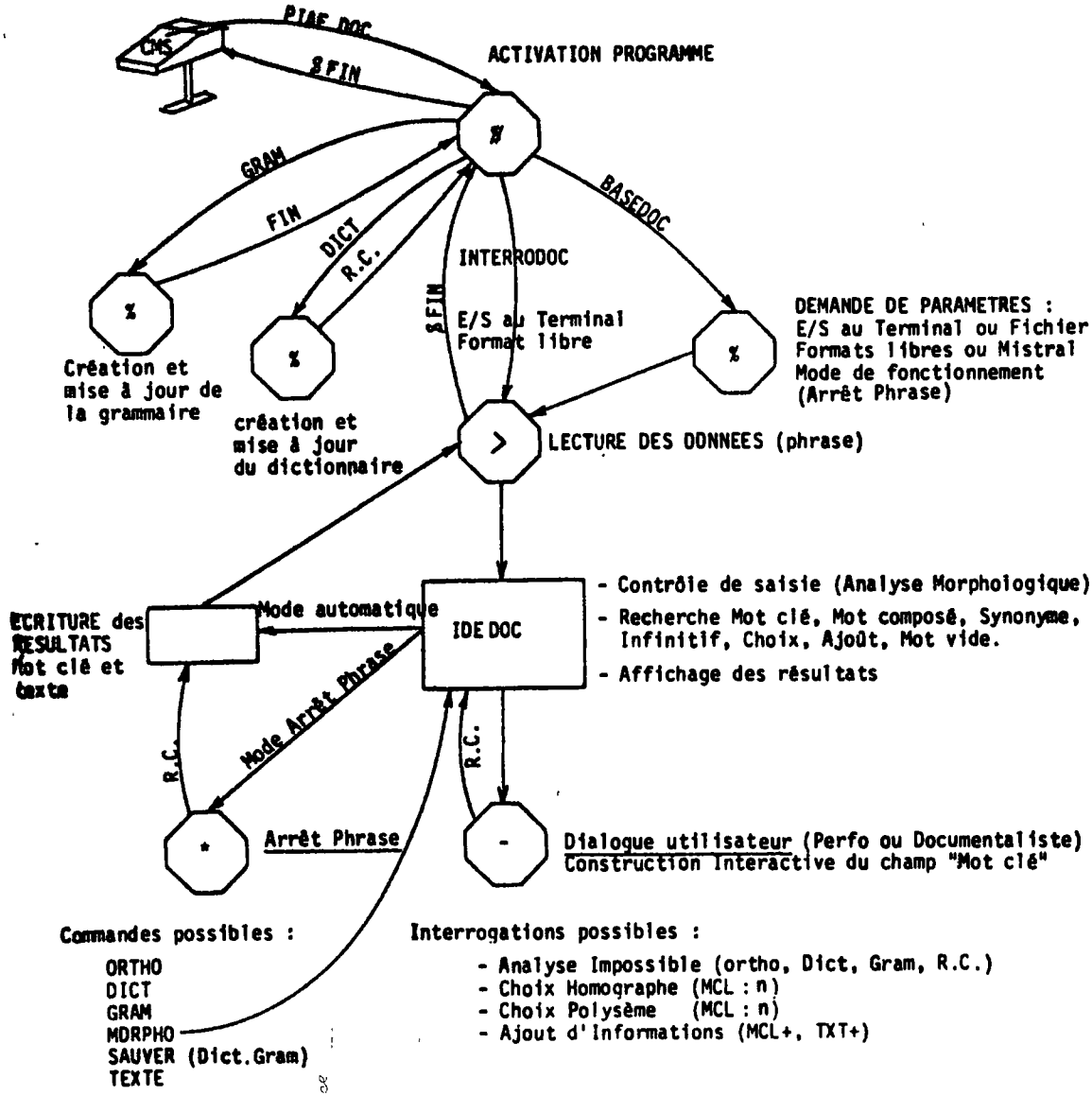
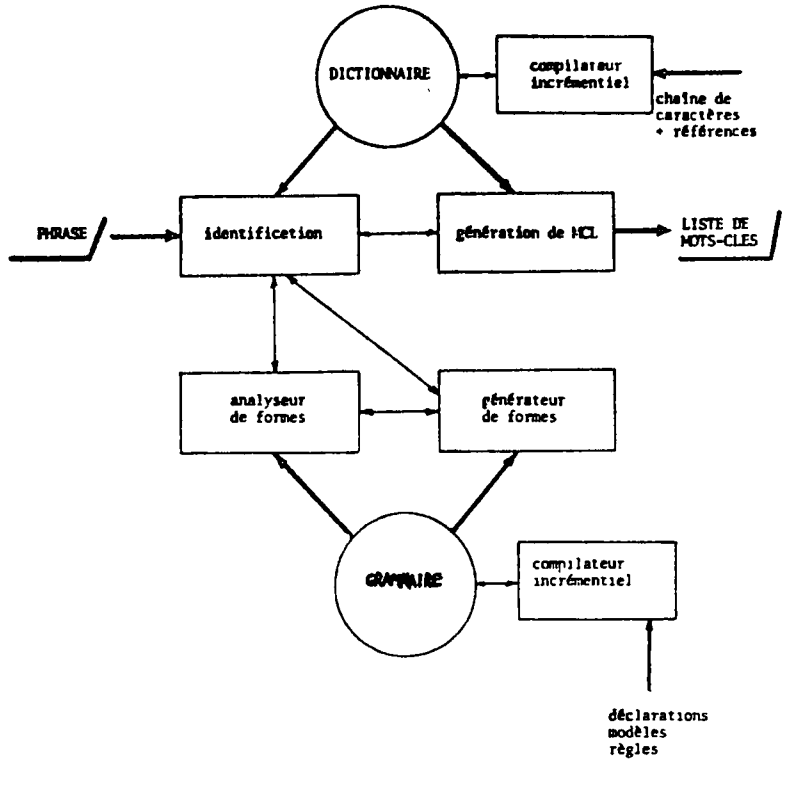
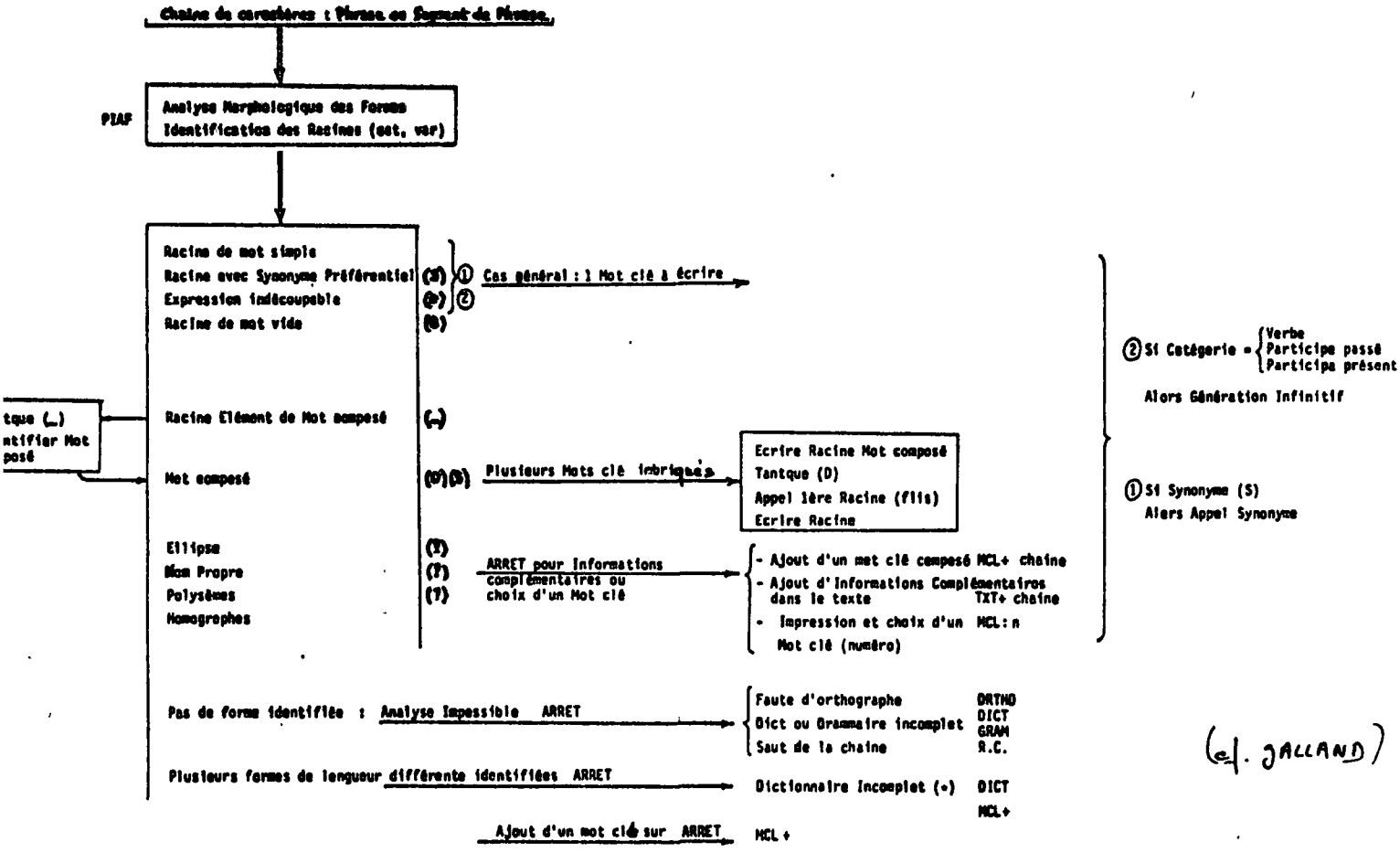


Schéma fonctionnel des modules d'exploitation et de mise au point



Construction interactive du champ des mots clés



(cf. JALLAND)

Exemples de saisie interactive :

- phrases en entrée
- → mots clés extraits par le système

```
• LES PROJETS DE LOI ADOPTES PAR L'ASSEMBLEE NATIONALE |
  1 ADOPTER
  2 ADOPTION (VOTE)
  3 ADOPTION (FAMILLE)
  CHOIX POLYSEME ? (MCL: N)
  -
  mcl: 2
▶ PROJET DE LOI$ADOPTION (VOTE)$ASSEMBLEE NATIONALE$ |$
      • LES ORGANISATIONS SYNDICALES TIENNENT COMPTE DU RECENT PROJET DE LOI SUR LA FOR
        ATION |
        1 FORMER
        2 FORMATION
        AJOUT ? (MCL+ OU TXT+ ...))
        -
        mcl+ permanente
  ⟶    ORGANISATION SYNDICALE$SYNDICAT$TENIR COMPTE$RECENT$PROJET DE LOI$FORMATION PER
        ANENTE$ |$
        *

• IL A FAIT APPROUVER DIFFERENTS PROJETS DE DECRETS |
▶ FAIRE APPROUVER$DIFFERENT$PROJET DE DECRET$ |$
  *
```

(cf. GRANDJEAN)

Extrait d'un état du dictionnaire (cf. le laboratoire de l'IMAG. Grenoble)

FILE: FILE SYSPRINT P1 CAMBRIDGE MONITOR SYSTEM

/ABANDON(*)/HOMME/ABANDONN/
/ABANDONN(*)(O)(S)/POS/ABANDON/
/ABATTAGE(*)/HOMME/
/ABATTEMENT(*)/HOMME/
/ABATTOIR/HOMME/
/ABATTOIRS DOUX (*)(S)/YEUX /ALSTHOM-CGE /
/ABDOMEN/HOMME/
/ABELIN (O)(S)/CARNOT /ABELIN PIERRE /
/ABELIN PIERRE (*)/CARNOT /ABELIN /
/ABERRATION/VOITURE/
/ABONNEMENT(*)/HOMME/
/ABORD/POS/
/ABOUTI/FINI/
/ABOUTISSEMENT(*)/HOMME/
/ABRIT(*)/POS/
/ABSENCE(*)(_)/VOITURE/ABSENT/
/ABSENCE DE VOTE (*)(S)/SOIF /ABSTENTIONNISME /
/ABSENT(*)(O)(S)/VRAI/ABSENCE/
/ABSOLU(*)/GRAND/
/ABSOLUMENT (*)(O)/HIER /
/ABSOLV/ABSOLV/ABSOUS/
/ABSOU/ABSOU/ABSOLV/
/ABSOUS/ABSOUS/ABSOU/
/ABSTENTION/VOITURE/
/ABSTENTIONNISME (*)/OEIL /ABSENCE DE VOTE /
/ABUS (*)/NEZ /
/ABUSI(*)/NEU/
/ACADEMIE(*)(_)/VOITURE/
/ACADEMIE DES SCIENCES (*)/SOIF /
/ACADEMIE FRANCAISE (*)/SOIF /
/ACCED(*)/POS/

les chaines soulignées représentent les modèles morphologiques de chaque entrée.

/REFONTE(*)/VOITURE/
/REFORM(*)/POS/
/REFORMATEUR(*)(S)/HOMME/MOUVEMENT REFORMATEUR /
/REFORME(*)(_)/VOITURE/
/REFORME CONSTITUTIONNELLE(*)/VOITURE/
/REFORME DE L'ENSEIGNEMENT SUPERIEUR (*)/FOIS /
*REFORME DE L'ENTREPRISE (*)/SOIF /
/REFORME DE L'OFFICE (*)(O)(S)/SOIF /REFORME DE L'ORTF /
/REFORME DE L'ORTF (*)/SOIF /REFORME DE L'OFFICE /
/REFORME DE LA SOCIETE (*)/SOIF /
/REFORME DES STATUTS (*)/SOIF /
/REFORME DU DIVORCE (*)/FOIS /
/REFORME DU SERVICE NATIONAL (*)/SOIF /
/REFORME FONCIER(D)/GRAND/
/REFORME FONTANET(D)/SOIF /
/REFORME PENITENTIAIRE(*)/VOITURE/
/REFUGIE(_)/HOMME/
/REFUGIE POLITIQUE(*)/HOMME/
/REFUS(*)(O)(S)/POS/REFUS /
/REFUS /NEZ /REFUS/
/REGARD/POS/

V. 4 CONDOR (+)

CONDOR est également un système dont l'objectif est de réduire au maximum toute manipulation - manuelle ou intellectuelle - des informations, structurées ou non structurées.

Son ambition est d'aboutir à l'automatisation de la chaîne documentaire complète (collecte, description, recherche et traitement), de réaliser l' "effacement des barrières entre l'homme et le système" en étant le plus "confortable" possible pour l'utilisateur et, enfin, d'être un système "hautement adaptatif".

Le projet, subventionné depuis 1973 par le Ministère allemand de la Recherche, est mené dans le cadre de la Société Siemens où il est actuellement en "phase pilote" (1). Il développe deux systèmes de traitement :

- l'un pour les informations structurées
- l'autre, dont il sera surtout question par la suite, pour les informations textuelles en langage naturel, avec les étapes suivantes :
  - saisie directe de données et de textes par des procédés de lecture optique
  - analyse, description et classification automatiques des textes
  - processus de recherche guidée à travers un dialogue, en partant de questions formulées en langage naturel.

---

(1) cf. TAEUBER

(+) "Communication en langue naturelle avec des systèmes de recherche documentaire orientés vers le mode conversationnel"

Le second système, qui est décrit plus précisément sous le nom de "STEINADLER" (1), réalise l'analyse du langage naturel à partir d'un lexique réduit des mots-outils (environ 800 entrées réunissant prépositions, pronoms, articles, verbes auxiliaires...) et d'un algorithme d'analyse des homographes (2).

Puis il opère l'extraction des racines des mots et l'analyse des dérivés, pour aboutir à une liste de descripteurs potentiels. Il crée, en parallèle, un thesaurus de ces racines.

Les racines sont ensuite pondérées, suivant leur forme morphologique et leur fonction dans la phrase, à partir de la structure trouvée pour la phrase (arbre de structure obtenu après analyses des groupes verbaux, des groupes nominaux, des prépositions et des conjonctions).

Les "descripteurs" sont alors répartis en "classes de priorités" par des opérations statistiques sur l'ensemble des documents, puis traités à l'intérieur de chacune des classes (module IMPRIOR) où ils sont regroupés par échelons - mise en évidence des relations linguistiques entre les descripteurs.

Les différents regroupements sont enfin "mis en réseau" (module ZWPRIOR).

Le nombre des classes, qui représentent en fait des hiérarchies dépend du nombre des documents et de la distribution des concepts dans le fonds documentaire.

On stocke ainsi à la fois les concepts et les relations entre les concepts, ce qui devrait augmenter nettement le taux de précision.

---

(1) cf. PANYR - dont le développement signifie : "indexation statistique de textes et classification documentaire automatique, au vu de résultats linguistiques" - voir aussi le schéma.
(2) cf. HALLER et WIELAND

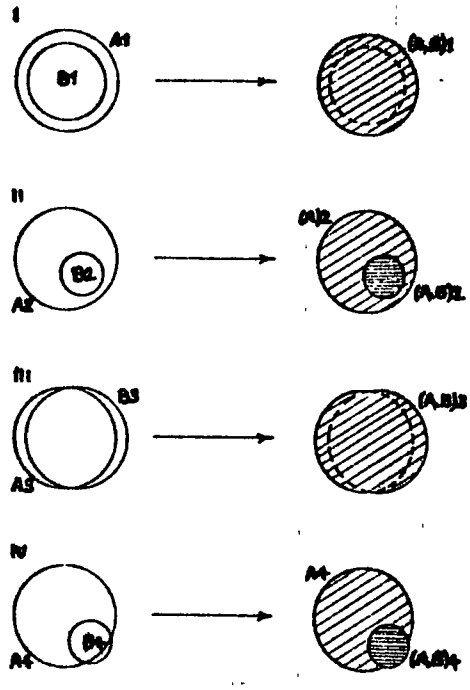
Les traitements linguistiques, et notamment l'analyse syntaxique, qui permettent d'établir les relations entre les concepts, sont encore dans une phase de teste; il convient de les développer et de les élargir avant de pouvoir espérer aboutir à la création automatique de réseaux sémantiques pour une grande quantité de textes (1).

Deux autres directions d'étude sont le développement d'un système pour la saisie et la structuration de grandes quantités de textes et l'étude d'extensions possibles à d'autres langues.

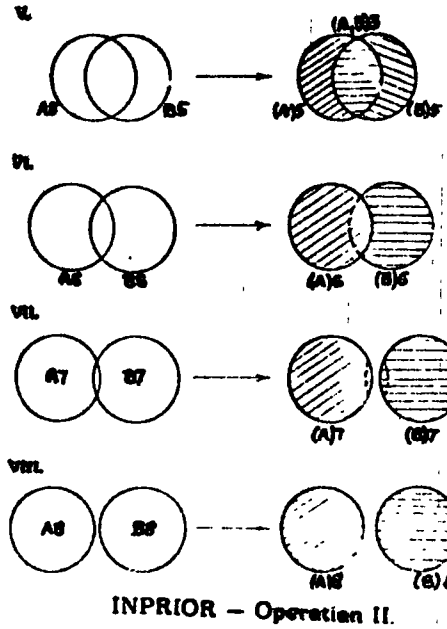
Ce dernier point en particulier devrait être facilité par le fait que la programmation se veut adaptative (c'est à dire qu'elle se compose de méthodes d'apprentissage à partir de résultats corrigés), ce qui entraîne - il faut le remarquer - des performances faibles au début, mais très avantageuses à long terme.

(1) cf. HALLER et WIELAND

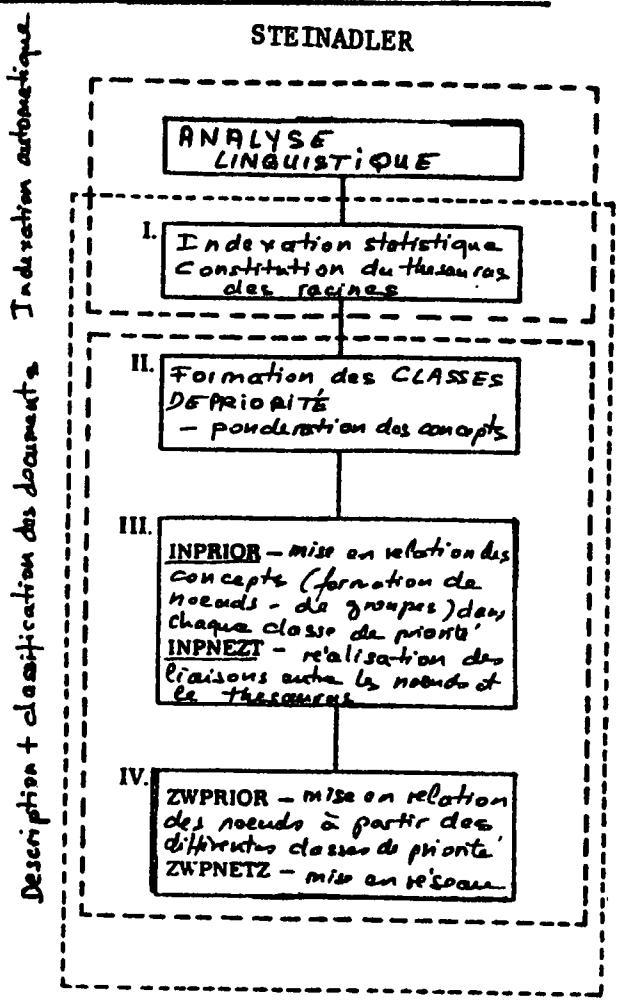
Fonctionnement et détail du logiciel steinadler -

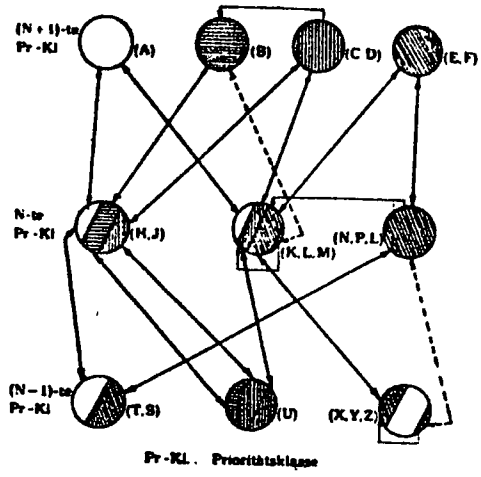


INPRIOR – Operationen I.



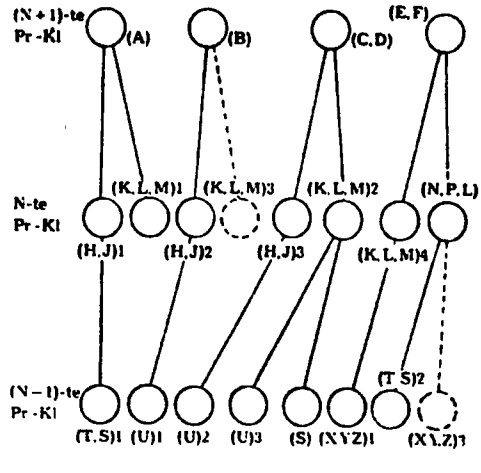
INPRIOR – Operation II.





Pr-Kl. Prioritätsklasse

ZWPRIOR mises en relations



Réseau de classification

(cf. PANYR)

Liste des catégories grammaticales

Nr	Kurzform/Erklärung		Beispiel(e)
01	NOM	Nomen	das Erlebnis, der Baum
		nominales Adjektiv	das Schöne
		nominaler Infinitiv Eigenname	das Erleben Fritz
02	ATT	attributiv verwendetes – Adjektiv – Partizip Präsens – Partizip Perfekt	 das schöne Erlebnis die schweigende Mehrheit das geglückte Experiment
03	VRB	finites Vollverb	sie gingen nach Hause
04	INF	Infinitiv Vollverb	sie versprachen zu gehen
05	PAR	Partizip Perfekt als Teil einer Verbalphrase	sie sind gegangen
06	ADV	unflektiertes Adjektiv oder Partizip in prädikativer oder adverbialer Verw.	das Erlebnis ist schön nach Hause kommend/gekommen, sagte er...
07	ZAN	Zahl in nominaler Verwendung	17, siebzehn
08	ZAT	Zahl in attributiver Verwendung	17 Männer
09	DET	Determinator (nominalgruppeneinleitender Artikel, Quantor etc.)	die/einige/alle Manner
10	REL	Relativpronomen	das Erlebnis, das schön ist
11	QAT	Quasiattribut (Quantor etc. in attributiver Verwendung)	die vielen Blumen
12	NML	Nominal (Quantor etc. in nominaler Verwendung)	Einige der Besten gewannen
13	PRN	Pronomen (durch „Subkategorisierung" unterhalb der Wortklassenebene unterschieden in:) – – Pers Pron – Demonstr Pron	 Er kam Dieser kam.
14	VBH	finites Hilfsverb (sein/haben)	Er ist gekommen
15	VBM	finites Modalverb (mussen etc)	Er wird/will kommen
16	INH	Infinit Hilfsverb	Er muß gekommen sein
17	INM	Infinit Modalverb	Er wird geschlagen werden
18	PRP	Präposition	Er kam nach Munchen
19	POP	Postposition	meiner Meinung nach
20	PA1	Prapositionaladjunkt 1	Er kam bis nach Munchen
21	PA2	Prapositionaladjunkt 2	Er kam von Munchen her
22	APT	Adverbialpartikel	Er war zu/sehr schon
23	VAD	Verbadjunkt	Er machte die Tur zu
24	KON	Konjunktion (nebenordnend)	Er kam und ging
25	NBS	Subjunktion	Da es regnet werde ich naß
26	FAV	Funktionsadverb	Heute/moglicherweise komme ich
27	IPA	Infinitivpartikel	Es ist deutlich zu sehen
28	NEG	Negation	nicht
31	SAZ	Satzbegrenzer (soweit von Segmentierung als solche erkannt)	.!/?/;,
32	SBZ	Subsatzbegrenzer	,/(/)/=
33	HHZ	Hervorhebungszeichen	"/ etc

```
PINGUIN-AUSGABE
01  AUSS**.
02    PROP
03      VMOD
04      | TPKZ
05      VR11
06      | K16 **SEIN
07      SUBJ
08      | NG
09      |   DETD
10      |   | K9  **DER
11      |   NK
12      |     NK
13      |     | K1  **EINSATZ
14      |     ATTR
15      |       NG
16      |         DETD
17      |         | K9  **EINER
18      |         NK
19      |           NK
20      |           | NK
21      |           | | K1  **ANALYSE
22      |           | ATTR
23      |           |   K2  **AUTOMATISCH
24      |           ATTR
25      |             K2  **LINGUISTISCH
26      SUBJ
27        NG
28          DETD
29          | K9  **EIN
30          NK
31            NK
32            | K1  **WEG
33            FIOA
34              PG
35                PRPF
36                | K16 **ZUR
37                NG
38                  DETD
39                  | BIND
40                  NK
41                    NG
42                    | K1  **BEWAELTIGUNG
43                    ATTR
44                      NG
45                        DETD
46                        | K9  **DER
47                        NK
48                          NK
49                          | K1  **PROBLEM
50                          ATTR
51                            PROP
52                              VMOD
53                              | TPKZ
54                              VOEL
55                              | K3  **AUFTAUCHEN
56                              SUBJ
57                              | NG
58                              |   DETD
59                              |   | BIND
60                              |   NG
61                              |     K1  **PROBLEM
62                              LKTP
63                                PG
64                                  PRPF
65                                  | K16 **BEI
66                                  NG
67                                    DETD
68                                    | K9  **DER
69                                    NK
70                                      NK
71                                      | K1  **VERARBEITUNG
72                                      ATTR
73                                        NG
74                                          DETD
75                                          | BIND
76                                          NK
77                                            NK
78                                            | K1  **TEXTMENG
79                                            ATTR
80                                              K2  **GROSS
```

Exemple d'arbre de structure

(cf. HALLER et WIELAND)

VI CONCLUSION - TABLEAUX SYNOPTIQUES

VI ESSAI DE CONCLUSION ET TABLEAUX SYNOPTIQUES

Dans le cas de systèmes documentaires, les problèmes de traitement des langues qui se posent sont ceux :

- de l'expression du contenu des documents
- de l'expression du contenu des questions
- des différents traitements possibles entre ces deux types d'expressions, pour arriver à la meilleure adéquation des réponses.

Dans la plupart des grand systèmes automatisés opérationnels à l'heure actuelle, les deux premiers aspects sont traités "en-dehors du système", et seul le troisième est en oeuvre des traitements, plus ou moins sophistiqués, entre les différents contenus des documents entrés et des questions, exprimés généralement sous forme de "mots-clés" ou de "descripteurs".

Cette procédure est basée sur l'existence de langages combinatoires - ou thesaurus; elle est orientés surtout vers l'automatisation de la recherche.

On pourrait considérer que l'étape suivante est celle où sont introduites des relations entre les termes. C'est le cas des deux premiers aspects traités dans cette note, avec toutefois entre eux une différence d'objectif, puisque les systèmes OLPI, NEPHIS, RELATIONAL INDEXING et PRECIS aboutissent simplement à des éditions d'index permutés, alors que SATIN 1 et TITUS II peuvent être envisagés comme des systèmes documentaires "complets".

Dans ces deux derniers cas, les contraintes augmentent, puisqu'il faut non seulement respecter un thesaurus, mais encore utiliser un métalangage très rigoureusement défini.

Cependant, les critiques principales faites aux systèmes dits "classiques", qu'ils soient d'ailleurs automatisés ou non - à savoir : manque d'uniformité, absence d'objectivité, fort taux de condensation, restent également valables pour ces systèmes, où le "traitement intellectuel" se fait <u>avant introduction des informations.</u>

Le traitement automatique du texte intégral en langage naturel représente une manière tout à fait nouvelle d'aborder le problème du traitement de l'information : en effet, l'utilisation de procédures automatiques et programmées devient un garant d' "objectivité" et d'uniformité, et chacun des systèmes présentés annonce des prévisions de performances accrues : selon le cas, suppression totale du bruit documentaire, meilleures réponses que dans un système indexé manuellement, taux de rappel élevé, augmentation du taux de précision...

On peut penser cependant que le problème de fond n'a fait que se déplacer et qu'il s'est même considérablement compliqué, dans la mesure où il s'agit d'introduire dans un ordinateur, avec toutes les formalisations que cela suppose en langage binaire, tout un ensemble d'informations - en particulier linguistiques - susceptibles de "traduire" une très grande complexité.

On trouve alors, comme préliminaires à la mise en oeuvre de ces systèmes, la référence à des théories linguistiques récentes (1), à la psychologie, à la psycholinguistique et aux mécanismes de la pensée, ce qui donne une idée de l'importance et de la variété du champ à couvrir...

---

(1) pour n'en citer que quelques-uns, on trouve surtout les noms de CHOMSKY, SALTON, F. de SAUSSURE, TESNIERE ...

Or on peut remarquer que les traitements actuellement réalisés sur le texte intégral en langage naturel suivent - à des variantes prés - une progression similaire dans les différents systèmes, consistant en particulier à reconnaître des formes (analyses morphologiques) puis des relations (analyses syntaxiques). Mais, ainsi que le signalait déjà Y. COURRIER dans son "état de la question" sur l'indexation automatique (1), le véritable travail d' "extraction du sens des documents" semble être "encore loin des réalisations".

Certes, certains problèmes sémantiques peuvent être résolus par le recours à des "lexiques" plus ou moins élaborés (et évolutifs) - ce qui revient d'ailleurs à réintroduire une forme plus performante de thesaurus, où les concepts (ou chaînes de caractéres) retenus comme descripteurs (2) sont réduits à leurs racines et où l'on tient compte de formes grammaticales variées (substantifs, mais aussi adjectifs, verbes conjugués, etc.), mais ce n'est qu'un aspect de la question, alors que le champ des recherches qui restent à mener dans le domaine sémantique est très vaste.

L'on essaie également de "traduire" les relations entre les termes, ce qui peut conduire à divers types de "réseaux" (réseau de questions, comme dans SYNTAXEME, ou réseau de concepts et de "classes de priorité", comme dans CONDOR).

Mais on peut considérer que, dans le domaine du traitement automatique des langues et en particulier du langage naturel en documentation, les recherches en sont encore plutôt à un stade exploratoire; il suffit de remarquer, à cet égard, que les essais qui se font concernent le plus souvent des textes courts qui sont déjà des résultats d'une élaboration (résumés, abstracts, dépêches de presse) et non, en fait, un texte véritablement intégral ...

---

(1) voir bibliographie p.3
(2) il s'agit souvent de toutes les "unités linguistiques" obtenues aprés filtrage à travers de listes de mots vides ou de mots-outils

On peut cependant s'attendre à des développements rapides, dus en particulier aux améliorations technologiques croissantes dans le domaine de l'informatique. Cela permet-il d'augurer, pour un avenir proche, des réalisations vraiment innovatrices ? La réponse dépend encore beaucoup des investissements mis dans la recherche...

TABLEAUX SYNOPTIQUES

On trouvera dans les pages suivantes trois tableaux qui tentent de rassembler les informations collectées en situant les systèmes décrits :

- d'une manière générale, les uns par rapport aux autres
- puis, plus précisément dans le cas de la mise en oeuvre d'opérations linguistiques, selon que les informations d'entrée sont préparées (utilisation d'un métalangage) ou non
- enfin, en ce qui concerne le traitement du langage naturel, selon les différents processus utilisés.

1 TABLEAU GENERAL

opérations sur les données d'entrée / caractéristiques des systèmes	Nom du système	Mode de saisie	Existence d'un thesaurus	Traitements effectués en machine	Produits - Recherche documentaire
Préparation pour l'édition	OLPI	conversationnel	non	tris tenant compte des indications données (entrées interdites - associations de mots - intercalation de mots-outils)	édition d'index ou de catalogues permutés
	NEPHIS	bordereau (conversationnel prévu)	non		
	Relational indexing	bordereau	non		
	PRECIS	bordereau	constitution simultanée d'un thesaurus		
Description élaborée, à l'aide d'un métalangage	SATIN 1	bordereau	Thesaurus à facettes	différents fichiers très "fins" stockage des relations données numériques	Produits de sortie très diversifiés
	TITUS II	- bordereau - saisie possible par écran cathodique	Thesaurus en 4 langues	indexation automatique traduction automatique données numériques	sorties des notices en plusieurs langues
Pas d'opérations: traitement du texte en langage naturel (en fait, il s'agit surtout de résumés, d'abstracts ou de dépêches de presse)	SYNTAXEME	conversationnel	lexique avec "mise à jour" évènementielle	chargement des liaisons grammaticales	recherche "guidée" en langage naturel
	INSTN/CEA-CEDIJ	conversationnel	constitution d'un thesaurus	contrôle de saisie indexation automatique classification automatique (essais)	interrogation en langage naturel
	PIAFDOC	conversationnel	construction d'un dictionnaire	contrôle de saisie indexation automatique assistée	interrogation en langage naturel par MISTRAL V 3
	CONDOR	conversationnel	construction d'un dictionnaire des racines	indexation automatique classification automatique	recherche "guidée" en langage naturel

2 Préparation et données stockées pour les traitements linguistiques

Nom du système	préparation des informations d'entrée	Constitution de thesaurus	Procédures d'interruption	Détail des traitements "linguistiques" internes	données correspondantes programmées
SATIN 1	utilisation d'un métalangage vocabulaire + syntaxe	lexique avec catégories - hiérarchisées - non hiérarchisées	/	analyse syntaxique à l'entrée (provoquant des rejets en cas d'erreur)	
TITUS II	utilisation d'un "langage documentaire canonique" lexiques + syntaxe (actants)	lexiques en 4 langues + leur codification en "langage-pivot"	possibilité de saisie avec correction en temps réel	entrée : consultation des dictionnaires ; analyse syntaxique ; stockage en langage-pivot sortie : analyse syntaxique avec grammaires ; consultation de dictionnaires	- table des morphèmes (= unités non autonomes) - grammaires programmées
SYNTAXEME	aucune préparation - langage naturel	lexique avec "mise à jour environnementale"	possibilités de modification à la saisie et de demandes d'assistance	- analyse morphologique (production d'une séquence codée) - analyse syntaxique (production de descriptions structurales) - transformation (édition des liaisons grammaticales)	- lexique des formes et désinences - grammaire (règles d'analyse)
INSTN (CEA) - CEDIJ		constitution d'un thesaurus	détection-correction d'erreurs	- analyse morphologique - analyse grammaticale et procédure complémentaire (détection des homographes) - fonction de reconnaissance des relations (SYN - TS, TS et TA) - fonctions de poids sémantique	- liste de terminaisons - liste de mots vides
PIAFDOC		vocabulaire évolutif (chaînes de caractères suivies d'indicateurs)	• contrôle de saisie avec correction • demandes d'assistance (choix à faire)	- programme de segmentation - analyse morphologique (détection d'homographies syntaxiques, de polysémies et cas d'élision)	- modèles morphologiques - liste de mots vides - grammaire (règles d'états finis - modèles - déclarations)
CONDOR		Construction d'un thesaurus de racines	(non précisées)	- analyse de la classe des mots - analyse des homographes - extraction des racines - analyse des dérivés → analyse syntaxique (arbre de structure de la phrase) (puis traitements statistiques)	- liste d'environ 800 mots-outils

3 Essai de grille de classement selon les processus linguistiques utilisés

	Analyses morphologiques	Analyses syntaxiques	Aspects sémantiques * (approches)	Traitements statistiques
Analyses morphologiques	• SYNTAXEME • INSTN (CEA) - CEDIJ • PIAF DOC • CONDOR	• SYNTAXEME • INSTN (CEA) - CEDIJ • CONDOR	• INSTN (CEA) - CEDIJ • CONDOR	• INSTN (CEA) - CEDIJ • CONDOR
Analyses syntaxiques	• SYNTAXEME • INSTN (CEA) - CEDIJ • CONDOR	• SYNTAXEME • INSTN (CEA) - CEDIJ • CONDOR	• CONDOR	• CONDOR
Aspects sémantiques (approches) *	• INSTN (CEA) - CEDIJ • CONDOR	• CONDOR	• INSTN (CEA) - CEDIJ • CONDOR	• INSTN (CEA) - CEDIJ • CONDOR
Traitements statistiques	• INSTN (CEA) - CEDIJ • CONDOR	• CONDOR	• INSTN (CEA) - CEDIJ • CONDOR	• INSTN (CEA) - CEDIJ • CONDOR

remarque 1 :
cette grille devrait être plus diversifiée, (mais les informations dont je disposais ne me l'ont pas permis).

* remarque 2 :
la prise en compte des aspects sémantiques est évidemment celle qui pose le plus de problèmes - elle fait l'objet de recherches dans les deux systèmes cités.

remarque 3 :
il est déjà difficile de résoudre les problèmes issus des analyses syntaxiques... (prise en compte des relations, réalisation de réseaux de concepts...)

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

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

remarque préliminaire :

La présente bibliographie est à la fois plus restrictive et plus complète que celle remise au moment du Conseil de Perfectionnement du DESS :

elle est plus restrictive, dans la mesure où un certain nombre de références (ou trop générales, ou trop spécifiques ou encore correspondant à des documents que je n'ai pas pu consulter) ont été éliminés.

Elle est plus complète, car j'ai recueilli, entre temps, des références complémentaires sur les systèmes décrits.

I REFERENCES GENERALES

/ PRESENTATION DES LOGICIELS ET DES SYSTEMES DOCUMENTAIRES

- Logiciels et systèmes documentaires/ Etude réalisée par le CXP (Centre d'Expérimentation de Packages).- Paris: ADBS, 1975 - 1976.- 3 vol., 218 + 132 + 371 p.- (Les Cahiers de l'ADBS)

/ OUVRAGES GENERAUX SUR LES PROBLEMES D'ANALYSE AUTOMATIQUE DES LANGUES EN DOCUMENTATION

- ANDREEWSKY A.
Apprentissage, analyse automatique da langage, application à la documentation.- Paris : Dunod, 1973.- 275 p.

- BELY N. BORILLO A. SIOT-DECAUVILLE N. VIRBEL J.
Procédures d'analyse sémantique appliquées à la documentation scientifique.- Paris : Gauthier-Villars, 1970.- 243 p.- (Documentation et information)

- CHAUMIER J.
Le Traitement linguistique de l'information documentaire : l'analyse documentaire.- Paris : Entreprise moderne d'édition, 1977.- 126 p.

- COYAUD M. SIOT-DECAUVILLE N.
L'Analyse automatique des documents.- Paris; La Haye: Mouton, 1967.- 149 p.- (Informatique)

- TRYSTAM J.P.
La Documentation automatique.- Paris : Dunod, 1971.- 124 p.- (Dunod économie)

/ SEMINAIRES DONT SONT EXTRAITES CERTAINES DES REFERENCES CITEES PLUS LOIN

- COMMISSION DES COMMUNAUTES EUROPEENNES. Luxembourg.
Franchir les barrières linguistiques : 3e congrès européen sur les systèmes et réseaux documentaires, Luxembourg, 3-6 mai 1977.- München: Verlag Dokumentation, 1977.- 2 vol., 730 + 180 p.

- INSTITUT DE RECHERCHE D'INFORMATIQUE ET D'AUTOMATIQUE. Rocquencourt
Etat des recherches sur les systèmes d'information documentaire, Saint-Vallier, 31 mai-3 juin 1978.- Rocquencourt : IRIA, atelier SESORI.- Dossier de communications.

/ ARTICLES ET TRAVAUX GENERAUX SUR L'INDEXATION AUTOMATIQUE

- COURRIER Y.
L'Indexation automatique : état de la question et perspectives d'avenir.
Documentation et bibliothèques, 23, 1977, n° 2 p. 59-72.

- GALLAND H.
Les Problèmes d'indexation automatique à travers l'étude de quelques systèmes : Mistral, Piafdoc, Stairs, Golem-Passat.- Grenoble: 1977.- 75 p.- (Mémoire de DEA. IMSS. 1977.)

- SPARK JONES K.
Automatic indexing.
Journal of documentation, 30, 1974, n° 4 p.393-432.

II L'EDITION AUTOMATIQUE D'INDEX

/ OLPI

- BASER KH. COHEN S. DAYTON L. WATKINS P.
Online indexing experiment at Chemical Abstracts Service : algorithmic generation of articulated index entries from natural language phrases.
Journal of chemical information and computer sciencs, 18, 1978, n°1 p. 18-25.

/ NEPHIS

- CRAVEN T.
NEPHIS : a nested-phrase indexing system.
Journal of the American Society for Information Science, 28, 1977, n° 2, p. 107 - 114.

/ RELATIONAL INDEXING

- FARRADANE J. GULUTZAN P.
  A Test of relational indexing integrity by conversion to a permuted alphabetical index.
  Intern. Classificat., 4, 1977, n°1 p. 20 - 25.

/ PRECIS

- AUSTIN (D.).- PRECIS : a manual of concept analysis and subject indexing.- London: 1974.

- FERRIER A.M.
  Présentation du système d'indexation "PRECIS" d'après l'expérience faite par le Département des Arts du Spectacle de la Bibliothèque nationale.
  Bulletin des bibliothèques de France, 23, 1978, n°3 p. 161-169
  Bulletin de la DICA, 2, 1977, n° 3-4 p. 17-21.

- SØRENSEN J.
  PRECIS : un système multilingue.
  in : Franchir les barrières linguistiques... vol. 1 p. 297-325 (1)

- VERDIER J. AUSTIN D.
  Recherche sur le potentiel translinguistique de PRECIS.
  in : Franchir les barrières linguistiques... vol. 1 p. 327-344 (1)

III LES SYSTEMES NECESSITANT UNE DESCRIPTION D'ENTREE ELABOREE (UTILISATION D'UN METALANGAGE)

/ SATIN 1

- BOURELLY L. CHOURAQUI E.
  Le Système documentaire SATIN 1.- Paris : CNRS, 1974-1978.- 2 vol.
  / Description générale et manuel d'utilisation.- 1974.- 398 p.
  / Génération et aide à la mise au point.- 1978.- 397 p.

- BOURELLY L. CHOURAQUI E.
  La Représentation des données documentaires : structure du métalangage SATIN 1.
  Communication IRIA - atelier SESORI (2)

---

(1) voir référence complète en p. 2
(2) voir référence complète en p. 3

- BOURELLY L. CHOURAQUI E.
La Description des données dans le système documentaire SATIN 1.
Automatisme, 22, 1977, n° 1-2 p. 27-39.

- BRANCA-LACOMBE G. TOMASSON R.
Intérêt de la constitution de banques de données élaborées dans le domaine biologique : utilisation du logiciel SATIN 1.
Documentaliste, 15, 1978, n° 5-6 p. 3-10.

/ TITUS II

- DUCROT J.M.
Le Système TITUS II.
Information et documentation, 1973, n° 4 p. 3-40.

- ZINGEL H.G.
Experiences with TITUS II.
Intern. Classificat., 5, 1978, n°1, p. 33-37.

IV LES SYSTEMES DE TRAITEMENT DU TEXTE INTEGRAL EN LANGAGE NATUREL

/ SYNTAXEME

- SOLET M.
SYNTAXEME : un système de gestion de la connaissance du langage naturel adapté à la consultation documentaire.
Documentaliste, 11, 1974, n° spécial p. 58-64.

- STEHLIN F.
Analyse morphologique de textes en langue naturelle : application à un système documentaire réalisé pour la DRME : SYNTAXEME.- 18 p.
Communication IRIA - atelier SESORI (1)

/ INSTN (CEA) - CEDIJ

- ANDREEWSKY A. FLUHR C.
Indexation automatique, maintenance et gestion d'un système documentaire : 1ère partie, aspects théoriques.- Paris : CEA, 1973.- 27 p.- (Note CEA-N-1694(1)).

---

(1) voir référence complète en p. 3

- ANDREEWSKY A. FLUHR C. RAMBOUSEK J.
Automatisation de l'analyse discriminante, de l'indexation, de la recherche hiérarchisée des documents et de l'aide à la décision.- Paris : CEA, 1973.- 56p.- (Note CEA-N- 1650)

- ANDREEWSKY A. FLUHR C.
Indexation automatique, construction automatique des thesaurus, classification automatique.- Paris : CEA, 1975.- 36p.- (Note CEA-N- 1795)

- CENTRE D'INFORMATIQUE JURIDIQUE. Paris.
[Plaquette de présentation des travaux de recherche sur la détectio et la correction automatique d'erreurs typographiques et sur le système documentaire statistique à indexation automatique et interrogation en langue naturelle].- 1p. + 2p. d'annexes.

- Résumé des problèmes de l'indexation automatique tels qu'ils sont abordés par le groupe de recherche en linguistique automatique (A. ANDREEWSKY, F. DEBILI, C. FLUHR, Y. HLAL, L. NICAUD).- 5 p.
suivi de :Déroulement des expériences d'indexation automatique en coopération avec le CEDIJ. - 2 p.
Communication IRIA - atelier SESORI (1)

/ PIAFDOC

- CHARRON J.
PIAF : Programme d'indexation automatique du français.- 4 p.
[compris dans la plaquette de présentation de la Banque d'Information Politique et d'Actualité.- Paris : Documentation française, sd

- GRANDJEAN E.
Application d'un système de traitement de langues naturelles pour l'indexation automatique Emploi du logiciel PIAFDOC .- 16 p.
Communication IRIA - atelier SESORI (1)

- GRANDJEAN E.
Projet PIAF : application à la documentation automatique. Définition et utilisation du produit-prototype PIAFDOC.- Grenoble: IMAG Equipe intelligence artificielle, 1978.- 40 p.

---

(1) voir référence complète en p. 3

/ CONDOR

- BANERJEE N.
CONDOR - Communikation in natural language with dialog oriented retrieval systems.
in : SCHNEIDER W. SAGWALL H.- Computational linguistics in medicine.- Amsterdam; New York; Oxford: 1977 p.163-175.

- FISCHER M.
Le Logiciel documentaire CONDOR.- Lyon: 1978.- 22p.- (Note de synthèse. DESS d'Informatique documentaire. 1978.)

- HALLER J. WIELAND U.
Die Erschliessung natürlichsprachlicher Information im Informations System CONDOR.
Nachrichten für Dokumentation, 29, 1978, n° 4-5 p. 177-183.

- PANYR J.
STEINADLER - ein Verfahren zur automatischen Deskribierung und zur automatischen thematischen DokumentenKlassifikation.
Nachrichten für Dokumentation, 29, 1978, n° 4-5 p. 184-191

- PORT J.
Probleme der maechinellen Sprachverarbeitung.
Nachrichten für Dokumentation, 29, 1978, n°1 p. 8

- TAEUBER D.
CONDOR : ein integriertes Datenbank und Informationesystem.
Nachrichten für Dokumentation, 29, 1978, n° 3 p. 127-130.