

Centre National de la Recherche Scientifique

SECTION D'AUTOMATIQUE DOCUMENTAIRE

Com./29/1967/ J.V. A.B./Grenoble

PROBLEMES SYNTAXIQUES
DE L'INDEXATION AUTOMATIQUE
DE DOCUMENTS

Andrée Borillo

Jacques Virbel

- Mai 1967 -

P L A N

0 - Introduction

1 - Définition de la représentation documentaire visée

- 1.1 - L'indexation automatique
- 1.2 - La représentation syntaxique

2 - Principes de la méthode utilisée pour l'analyse syntaxique

3 - L'analyse des textes en langage naturel

- 3.1 - Dimension de l'unité de texte analysé
- 3.2 - Mécanisme des règles
 - 3.21 - Reconnaissance de la catégorie grammaticale
 - 3.22 - Attribution du rôle syntaxique

4 - Règles de construction des syntagmes de type consécutif et comparatif

- 4.1 - Description des mots-outils
- 4.2 - Les schémas syntaxiques
- 4.3 - Exemples de classes d'outils
- 4.4 - Repérage des correlanda

5 - Construction des syntagmes de type associatif

6 - Utilisation des informations sémantiques dans les règles de construction des syntagmes

- 6.1 - Pour la construction des syntagmes consécutifs et comparatifs
- 6.2 - Pour la construction des syntagmes associatifs

7 - Résultats obtenus et leur synthèse

- 7.1 - Types de résultats
- 7.2 - Opérations de raccordements

8 - Conclusion

0 - Introduction .

On se propose d'aborder dans cette communication l'aspect syntaxique de l'indexation automatique de résumés scientifiques (1). On définit tout d'abord un modèle syntaxique propre à représenter les relations qui unissent les descripteurs extraits d'un texte en langage naturel (§ 1). Puis on expose la méthode générale que l'on compte mettre en oeuvre pour opérer le passage de l'expression en langage naturel vers cette représentation artificielle (§ 2). On définit ensuite les divers outils linguistiques utilisés à cette fin : une analyse syntaxique (§ 3) ; un ensemble de règles permettant la mise en corrélation des descripteurs ; le "réseau notionnel", organisation syntagmatique des descripteurs du langage documentaire (§§ 4-5). On montre comment ces outils utilisent des informations tout à tour syntaxiques et sémantiques (§ 6). Des opérations de synthèse sont enfin prévues pour ordonner et compléter les résultats terminaux (§ 7).

(1) Ces résumés, rédigés en français, ont été extraits du "Bulletin signalétique du C.N.R.S." et se rapportent au domaine de la Psycho-Physiologie.

1 - Définition de la représentation documentaire visée

Avant de décrire la représentation syntaxique que l'on vise, donnons une définition de l'indexation automatique, telle qu'elle est ici conçue.

1.1 - L'indexation automatique

L'indexation automatique est envisagée ici comme un cas particulier de traduction de textes écrits dans une langue naturelle vers une représentation de ces textes formée dans un langage documentaire artificiel. Ce langage documentaire, défini a priori, est destiné à exprimer le contenu des documents sous une forme de nature à faciliter les opérations de recherche des informations.

Les caractéristiques du langage documentaire ont déjà été décrites par ailleurs (1) ; rappelons simplement qu'il est constitué par un lexique, ensemble de "descripteurs" organisés sémantiquement - chaque descripteur correspondant à un concept défini dans un champ scientifique donné - et par une syntaxe qui fait l'objet du paragraphe suivant.

La phase de traduction syntaxique dont il sera question est précédée d'une phase dite lexicale, où l'on cherche à repérer les concepts attestés dans le texte naturel de départ et à les traduire par les descripteurs correspondants, ceci à l'aide d'un dictionnaire automatique. Nous supposons achevée cette phase lexicale de l'indexation (2) ; la phase syntaxique qui suit s'applique donc à une liste inorganisée de descripteurs obtenus à partir du texte.

1.2 - La représentation syntaxique

Les relations syntaxiques choisies pour exprimer les rapports logiques qui unissent les descripteurs sont empruntées à la grammaire SYNTOL (3). Elles sont ici au nombre de trois :

- la relation "consécutives" qui marque une interdépendance d'ordre dynamique entre deux concepts.

Ex. : A exerce une action sur B

(1) Voir bibliographie, in fine, n° 1

(2) Sur cette étape, voir bibliographie n° 2

(3) Voir bibliographie n° 1, pp 44 - 88

- la relation "comparative" qui indique un rapprochement formel entre deux termes.

Ex. : A est différent de B

- la relation "associative", marquant un rapport d'ordre statique, quelle que soit sa nature : appartenance, inclusion, qualification, etc.

Ex. : A est dans B
A est accompagné de B
A appartient à B etc...

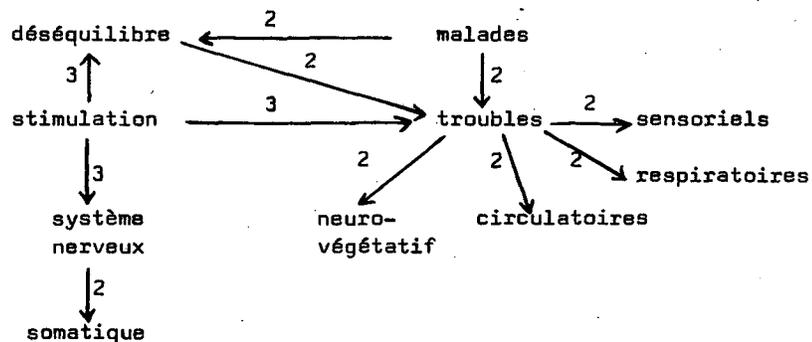
On appelle syntagme un couple de descripteurs unis par une relation. Chaque relation est orientée d'un pôle à l'autre du syntagme selon ces règles précises (1).

Si les textes analysés sont homogènes, c'est-à-dire s'ils se réfèrent à un seul thème, comme c'est généralement le cas pour des résumés d'ouvrages de psycho-physiologie, on peut poser pour principe que la représentation documentaire de ce texte doit avoir la forme d'un graphe connexe, d'où tout isolat est donc exclu.

Voici un exemple de texte que l'on se propose d'analyser, et la représentation documentaire que l'on vise à obtenir.

Texte : "Étude de trois malades présentant un déséquilibre unilatéral avec troubles sensoriels après diverses stimulations. Ces dernières provoquent aussi des troubles neuro-végétatifs, circulatoires et respiratoires, montrant que le système nerveux somatique n'est pas seul intéressé par les stimuli".

Représentation documentaire visée



(1) Voir bibliographie n° 1, pp 51 - 66

Les termes soulignés dans le texte original sont ceux qui ont un équivalent lexical dans le langage documentaire. Ce sont ces descripteurs que l'on trouve dans la représentation documentaire (1). Les relations syntaxiques attestées entre les descripteurs sont notées par une flèche ($\xrightarrow{3}$: relation consécutive ; $\xrightarrow{2}$: relation associative).

Comme on le voit, la représentation obtenue ne compte pas d'isolat, mais constitue au contraire un graphe unique.

C'est ce type de produit que l'on s'efforce d'obtenir par la procédure décrite aux paragraphes suivants.

2 - Principes de la méthode utilisée pour l'analyse syntaxique

L'analyse syntaxique consiste donc à passer d'un texte en langage naturel vers la représentation documentaire telle qu'elle vient d'être définie. Un certain nombre d'hypothèses de travail ont été posées pour élaborer les règles nécessaires à ce passage.

a) Tout d'abord, on part du postulat que les relations consécutives et comparatives sont marquées dans le langage naturel, pris dans un domaine spécialisé, par des constructions dont on peut donner des listes finies. A la base de ces constructions, on constate qu'apparaissent des connecteurs syntaxiques ou mots-outils que l'on peut isoler et caractériser. A partir de ces repères que constituent les mots-outils, on estime pouvoir établir des règles de corrélation pour la production de syntagmes corrects et non ambigus. Par ce procédé toutes les relations consécutives et comparatives établies entre descripteurs du langage documentaire peuvent être extraites d'un texte et fournir une partie de la représentation documentaire visée.

b) De ce fait, on déduit que toutes les relations qui restent à extraire sont des relations associatives. Cette démarche qui fait de la relation associative une relation résiduelle peut se justifier dans la mesure où cette relation étant la plus chargée d'un point de vue sémantique et syntaxique, il est préférable de ne pas avoir à l'interpréter mais de l'obtenir par déduction.

Cependant même si l'interprétation de la relation associative est implicite, il faut au moins la détecter dans le texte, et il semble qu'on ne puisse opérer comme précédemment : reconnaître tout d'abord un mot-outil susceptible d'explicitement la relation, puis d'après l'outil, rechercher les correlanda. Etant donné la variété de signification que la relation associative recouvre, les types d'outils à recenser seraient très nombreux et difficiles à caractériser. D'autre part, même si l'on réussissait à dresser une liste quasi exhaustive de ces mots-outils et à les classer selon certains critères opératoires, il n'est pas

(1) Malgré l'identité orthographique que l'on n'a maintenue que pour plus de clarté.

certain que le repérage des correlanda en serait facilité. En effet, les mots-outils considérés étant de type très différent, les contextes syntaxiques dans lesquels ils apparaissent n'ont rien de commun - à la limite on aboutit presque à des cas d'espèces qu'il faut relever un à un. De plus le contexte syntaxique n'étant pas suffisant, dans beaucoup de cas il faut avoir recours au contexte sémantique des correlanda eux-mêmes, qui de ce fait jouent un rôle aussi important que les mots-outils recensés. C'est pourquoi pour la construction des syntagmes de type associatif une démarche différente est prévue. On n'utilise pas de mots-outils comme précédemment ; on relève les descripteurs qui n'ont pas été mis en relation consécutive ou comparative et on établit entre eux des corrélations de fonction à fonction du type : "nom sujet - verbe" ou encore "nom sujet - nom complément d'objet".

c) Une autre hypothèse a été posée concernant la place et l'importance relatives des informations de type sémantique et syntaxique dans la méthode d'analyse. Lors d'une expérience antérieure, une première méthode d'analyse avait été expérimentée (1); elle se caractérisait en particulier par l'usage exclusif d'outils sémantiques pour la mise en corrélation : le lexique et le réseau notionnel.

Le réseau notionnel constitue une organisation syntagmatique des descripteurs du lexique, et atteste ainsi les affinités combinatoires des concepts ; par exemple, la classe des agents (chimiques, physiques, psychologiques, etc.) est reliée par une relation consécutive aux entités pouvant être affectées par leur action (cerveau, foie, digestion, sécrétion, etc...). Dans cette méthode, la construction des syntagmes se déclenche au vu de mots-outils, termes ayant un rôle syntaxique dans la langue naturelle. Elle s'opère par consultation du réseau notionnel qui atteste quels couplages des descripteurs sont acceptables dans le contexte du mot-outil.

Les résultats ont montré (2) d'une part que cette méthode permet de résoudre nombre de difficultés lors de l'établissement des relations entre descripteurs, et d'autre part que les erreurs constatées peuvent souvent être réduites par l'apport d'informations syntaxiques.

En utilisant comme base cette première méthode, on se propose donc de ne pas limiter les moyens de l'analyse à des informations strictement sémantiques, et d'exploiter des informations grammaticales et fonctionnelles du contexte en langage naturel. Ainsi, en diversifiant la nature des informations, et en les combinant, on vise à mettre au point une méthode mixte, plus efficace que chacune des méthodes syntaxique ou sémantique prise séparément.

(1) Voir bibliographie nos 3 et 4.

(2) Voir l'évaluation des résultats dans bibliographie n° 4.

d) Cette exploitation des informations autres que sémantiques suppose une procédure d'analyse grammaticale du texte en langage naturel. En effet pour pouvoir utiliser le contexte syntaxique, il faut qu'un traitement autonome ait fait attribuer à chacun des mots de l'énoncé où se trouve l'outil syntaxique ou les descripteurs, le type de renseignement grammatical et fonctionnel que l'on attend pour procéder à la recherche des correlanda.

Une dernière hypothèse, d'ailleurs liée à la précédente, est posée en ce qui concerne la nature et les dimensions de cette analyse. L'exploitation que l'on veut faire des résultats de cette analyse grammaticale, et le rôle que l'on entend faire jouer aux informations sémantiques grâce au réseau notionnel, permettent d'avancer qu'il n'est pas nécessaire d'effectuer une analyse totale de la phrase, mais qu'une simple segmentation en syntagmes caractérisables du point de vue fonctionnel est suffisante.

C'est cette première opération d'analyse qui est exposée au paragraphe suivant (§ 3). A l'issue de cette analyse, les renseignements grammaticaux et fonctionnels étant obtenus, on aborde la phase de construction des syntagmes. Etant donné les procédures distinctes préconisées pour l'obtention des relations consécutives et comparatives d'une part, pour les relations associatives d'autre part, on examine tout d'abord la construction des syntagmes du premier type (§ 4), puis celle des syntagmes du second type (§ 5), en indiquant (§ 6) quelle est la part des informations sémantiques dans chacune des constructions. Dans un dernier paragraphe (§7), on donne un aperçu général des résultats obtenus par ces divers procédés et des solutions envisagées pour parfaire la représentation documentaire: complètement et raccords.

3 - L'analyse des textes en langage naturel

3.1 - Dimension de l'unité de texte analysé

Une segmentation des textes est déjà faite dans la phase de traduction lexicale, nécessitée par le balayage d'unités de contexte déterminées lors de l'interprétation sémantique des termes polysémiques (1). Ces unités contextuelles sont délimitées par le point -lorsqu'il n'est pas un signe d'abréviation - et le point-virgule. Pour la traduction syntaxique, ce premier découpage fait apparaître des unités trop complexes, car elles contiennent des propositions de nature différente (principales, subordonnées, complétives, etc.) ; on donne donc également le statut de séparateur aux conjonctions de subordination. L'analyse se déclenche et s'arrête alors au vu de l'un des séparateurs suivants : signe de ponctuation ou conjonction de subordination.

(1) Voir bibliographie n° 2

Cependant si les opérations d'analyse se déroulent de manière discontinue et traitent chaque membre de phrase comme une unité autonome, il ne faut pas oublier que le résultat final attendu doit perdre ce caractère disjoint pour former une représentation cohérente du texte. Nous verrons (§ 7.2) les opérations qui permettent de relier entre eux les groupes de syntagmes isolés en raison du caractère fragmentaire de l'analyse.

3.2 - Mécanisme des règles

L'analyse, s'appliquant à des membres de phrase compris à l'intérieur des frontières que nous venons de définir, s'effectue grâce à des règles de concaténation. Ces règles fixent l'enchaînement des catégories grammaticales des mots pris deux à deux dans l'ordre où ils apparaissent dans la phrase examinée de gauche à droite. Les règles établies sous forme d'équivalences comportent deux parties : la partie droite concerne le choix de la catégorie grammaticale, la partie gauche l'attribution du rôle syntaxique.

3.21 - Reconnaissance de la catégorie grammaticale

La possibilité de juxtaposition de deux catégories dans une phrase obéit à des règles de distribution. En faisant le recensement de toutes les juxtapositions permises, on arrive à résoudre certaines ambiguïtés dues à des homographies.

Ex. : Le fait de trouver une règle : "article + nom = ...", mais non pas : "pronom + nom = ...", résout l'ambiguïté du type : "Le chat" ... "Le", "pronom" ou "article", placé immédiatement avant un nom ne peut être qu'"article".

Le filtrage n'est pas toujours aussi facile :

Ex. :	Le	chat	entre	dans	la	cage
	"pro.pers."	"nom"	"verbe"	"préposition"	"pro.pers."	"nom"
	"article"		"préposition"		"article"	

D'après la liste de règles établie, on peut éliminer l'interprétation "pronom personnel" pour "le", mais pour "chat" et "entre" correspondent deux règles de caractérisation :

nom sujet + verbe =
 nom sujet + préposition =

(le premier constituant est accompagné d'un code fonctionnel attribué lors de la règle précédente, le deuxième ne comprend encore que le code de catégorie).

Etant donné les deux règles permises, l'interprétation de "entre" peut être aussi bien "verbe" que "préposition". Nul choix n'étant possible, les deux interprétations sont donc conservées.

A l'étape suivante, lorsque l'on doit faire l'enchaînement "entre"+"dans", une seule règle est possible :

verbe intransitif + préposition =

La double interprétation pour "entre" n'est plus possible, celle de "préposition" est rejetée puisqu'elle n'est plus vérifiée par ce qui suit. Après examen de toute la phrase, on obtient en ce qui concerne les catégories :

le	chat	entre	dans	la	cage
art.	nom	verbe	prép.	art.	nom

3.22 - Attribution du rôle syntaxique

Dans la deuxième partie de la règle est définie la fonction syntaxique du deuxième constituant.

Ex. : verbe transitif * + nom = verbe transitif + nom compl. objet (l'astérisque qui accompagne "verbe transitif" dans la première partie de la règle indique que le verbe n'a pas encore reçu de complément d'objet ; ce signe disparaît dans la deuxième partie puisque le complément d'objet est trouvé).

Le rôle syntaxique du nom découle ici de la nature et du rôle du mot qui précède le verbe transitif - qui lui-même a été défini par rapport au mot qui précédait, etc...

Ex. : (1) le médicament provoque des troubles
Ex. : (2) les troubles que provoque le médicament

Bien que placés de même façon par rapport à "provoque", "troubles" et "médicament" n'ont pas la même fonction. Cette différence est marquée explicitement par le verbe :

Dans (1) verbe transitif * + nom = verb. transit. + c. objet
Dans (2) *verbe transitif + nom = verb. transit. + sujet

Dans (2) la présence de l'astérisque avant le verbe indique que le sujet n'a pas encore été trouvé ; l'absence d'astérisque, après, signale par contre que le complément d'objet précédait le verbe ; "troubles" a donc une fonction de "complément d'objet", "médicament" une fonction de "sujet".

Parfois la catégorie et la fonction d'un mot ne suffisent pas à déterminer la fonction du mot qui suit. C'est le cas par exemple des compléments prépositionnels :

- (3) Etude sur les souris de l'antagonisme...
- (4) Etude sur le foie de la souris...

Ici aucune règle ne peut montrer que "de l'antagonisme" dépend de "étude" alors que "de la souris" dépend de "foie". Au lieu de formuler une règle arbitraire qui ferait par exemple dépendre un complément prépositionnel du premier mot qui le précède, on a choisi de faire du complément prépositionnel, lorsqu'il ne suit pas un nom sujet, une structure autonome dont la nature est indiquée par un code spécifique.

Ex. : Etude sur les souris / de l'antagonisme /:"compl. prépositionnel" (1)
Ex. : Etude sur la foie / de la souris /:"compl. prépositionnel"

De la même façon, des indices spécifiques ont été choisis pour les membres de phrase coordonnés.

Ex. : Etude de l'activité bioélectrique du cerveau et des troubles du comportement.

On n'est pas autorisé à établir systématiquement une liaison entre les éléments les plus rapprochés de part et d'autre de la coordination. Ici les deux noms "cerveau" et "troubles", même introduits par les mêmes prépositions "de", ne sont pas mis en liaison par "et". Cette liaison est en fait entre "activité" et "troubles".

C'est pourquoi on préfère en général ne pas essayer de rattacher le membre de phrase coordonné. On le considère comme disjoint et on lui donne une spécification qui le fait reconnaître des autres structures autonomes (les compléments prépositionnels par exemple). Le seul cas particulier est celui des adjectifs qui, lorsqu'ils sont de part et d'autre d'une conjonction de coordination, sont reliés et reçoivent le même code de fonction.

Ex. : des lobes frontaux et latéraux

Mais : Etude de l'activité bioélectrique / du cerveau /
/ et des troubles / du comportement
"membre de phrase coordonnée"

On agit de même pour les prépositions relatives, certains membres de phrase isolés par des virgules, etc.

- La dépression du rythme alpha / qui est
"groupe nominal sujet" prop.

caractéristique / de la réaction / d'orientation /
"relative" "compl.préposit." "compl.préposit."

reflète un processus / d'excitation /
"groupe verbal" "compl. préposit."

(1) Dans les exemples qui suivent, le "/" indique les points de segmentation de la phrase.

Comme on vient de le montrer, si l'on considère l'arborescence que l'analyse structurale de la phrase devrait engendrer, on s'arrête à des résultats intermédiaires, loin du sommet.

On va voir comment les informations fournies par l'analyse grammaticale sont reprises et utilisées dans les constructions des syntagmes SYNTOL.

4 - Règles de construction des syntagmes de type consécutif et comparatif

La construction s'effectue à partir des mots-outils. Nous verrons en premier lieu ce que sont les mots-outils, et comment ils sont spécifiés (§ 4.11-4.13) ensuite comment, suivant les règles attachées aux mots-outils, se déroule la recherche des descripteurs à corréliser (§ 4.14).

4.1 - Description des mots-outils

Les mots-outils sont des mots ou groupes de mots du langage naturel qui marquent des liaisons de type consécutif ou comparatif.

On peut les distinguer du point de vue :

a) grammatical : ils peuvent être des verbes, des noms, des adjectifs, des prépositions.... Voici quelques mots-outils de type consécutif :

verbes : inhiber, provoquer, augmenter...

noms : suppression, effet de, réaction à....

adjectifs : épileptogène, destructeur, consécutif à....

prépositions : par suite de, en fonction de...

b) formel : certains comportent un seul terme (ex. : provoquer, épileptogène...) d'autres peuvent être fragmentés et admettre des insertions entre leurs différents constituants (réagit le plus souvent à..., a un effet passager sur...); d'autres enfin se présentent comme nécessairement discontinus (action de l'atropine sur le foie, réaction du muscle à la stimulation...). Dans une première expérience (1), une distinction très importante du point de vue opératoire avait été faite entre les premiers, les outils autonomes, et les derniers, les grilles, cas particulier d'outils discontinus dont les insertions pouvaient être des descripteurs. Cette distinction n'est pas retenue dans la nouvelle procédure.

(1) Voir bibliographie n° 3

c) sémantique : certains mots-outils reçoivent une traduction en SYNTOL : ils sont à la fois descripteurs et connecteurs, on dit qu'ils sont mixtes. D'autres ne sont pas retenus pour leur signification mais seulement pour leur fonction dans la phrase, ce sont les outils purs.

Deux critères sont tout d'abord retenus pour la classification des outils : la catégorie grammaticale et l'interprétation lexicale. D'après ces deux critères, on peut dresser une liste des mots-outils de type consécutif.

	Catégorie grammaticale				Interprétation lexicale	
	verbe	nom	adjectif	préposit.	mixte	pur
inhiber	X				X	
provoquer	X					X
modification		X			X	
effet		X				X
epileptogène			X		X	
par suite de				X		X

Dans cette classification, on remarque que seuls les verbes et les noms sont soit mixtes, soit purs, tandis qu'on n'a pas trouvé de mots-outils adjectifs qui ne soient pas mixtes, d'outils prépositions qui ne soient pas purs.

Cependant la classification des mots-outils ne se fonde pas seulement sur la combinaison de ces deux critères. Une distinction supplémentaire est retenue qui fait appel au type de construction syntaxique du mot-outil, c'est-à-dire à la nature grammaticale et la fonction syntaxique des termes qu'il est susceptible de mettre en relation. Ce troisième critère de spécification ne va

pas donner lieu à une nouvelle dichotomie des classes déjà indiquées, mais va conduire à reconsidérer les distinctions faites suivant les catégories grammaticales. En effet, la nature des compléments d'un mot dépend en premier lieu de sa catégorie. Ex. : un nom peut être complété par un adjectif ou un autre nom, un verbe attend un sujet généralement nom ou pronom, des compléments directs ou indirects introduits par des prépositions etc... Ensuite, à l'intérieur de chacune des catégories grammaticales des différences se font jour : Ex. les verbes transitifs sont complétés par des noms sans l'intermédiaire de prépositions, ils peuvent se construire à la forme passive, certains verbes intransitifs régissent certaines prépositions etc...

C'est cette notion de différenciation syntaxique que l'on va exprimer sous forme de règles qui, attachées à une sous-classe de mots-outils, indiqueront toutes les combinaisons syntaxiques possibles pour ce type d'outil. A ces combinaisons, on a donné le nom de "schémas syntaxiques".

4.2 - Les schémas syntaxiques

Les schémas syntaxiques sont des suites continues ou discontinues dans le langage naturel de codes grammaticaux et fonctionnels énoncés dans un ordre donné avant et/ou après l'outil, dans les limites d'une zone déterminée, sans spécification de distance par rapport à l'outil.

Ex. : Au mot-outil de type consécutif : "effet" (de... sur...) est associé une série de schémas syntaxiques dont voici quelques exemples (1).

Nom	{ sujet ou compl.}	+ de + compl. prépositionnel + sur + compl. prépositionnel
Nom	{ sujet ou compl.}	+ sur + compl. prépositionnel + de + compl. prépositionnel
Nom	sujet	+ avoir + <u>nom compl. objet</u> + sur + compl. prépositionnel
Nom	{ sujet ou compl.}	+ qui + avoir + <u>nom compl. objet</u> + sur + compl. prépositionnel
Nom	{ sujet ou compl.}	+ avoir + <u>nom compl. objet</u> + adjectif

(1) Dans ces exemples, le code grammatical et fonctionnel représentant le mot-outil est souligné.

Ex. : sujet : L'effet de l'atropine sur le foie est étudié...

compl. d'objet : On étudie l'effet de l'atropine sur le foie....

compl. prépositionnel : Etude de l'effet de l'atropine sur le foie...

Pour un autre mot-outil, "provoquer" par ex., la liste des schémas est tout à fait différente :

Nom Sujet + <u>Verbe</u> + compl. d'objet		l'hexorbarbital provoque le sommeil
Nom Sujet + qui + <u>Verbe</u> + compl. objet		l'hexorbarbital qui provoque le sommeil
Sujet		
Nom ou + que + <u>Verbe</u> + sujet		le sommeil que provoque l'hexorbarbital
compl.		
Nom Sujet + être + <u>Part. passé</u> + par		le sommeil est provoqué par l'hexorbar-
+ compl. préposition		bital

Pour pouvoir caractériser de la sorte un mot-outil, il faut qu'il apparaisse dans des contextes syntaxiques suffisamment stéréotypés. C'est le cas en général des verbes de types consécutifs ou comparatifs qui sont soit transitifs directs et impliquent un complément d'objet, soit des verbes transitifs indirects tels que "agir sur", "dépendre de"....

Les noms eux aussi ont une construction stable ; de même que les verbes à gouvernant fort, ils appellent certaines prépositions :

	dans
effet de...	sur ...
	chez

Il est possible de dresser une liste finie des prépositions ou locutions prépositives pouvant servir de charnière à un mot-outil donné ; ici "effet" peut se construire avec l'une des prépositions indiquées "dans, sur, chez".

Dans les schémas syntaxiques donnés en exemple plus haut, la stricte juxtaposition des éléments requis n'est pas obligatoire. Un nom Sujet peut être séparé du verbe par des compléments de nom, des adjectifs, des adverbes.... mais il est reconnaissable ; sa qualité de sujet le rend caractérisable quelle que soit la distance dans la phrase qui le sépare du verbe. De même le complément d'objet peut être séparé du verbe.

Ex. : "L'hexorbarbital, administré à faible dose, provoque généralement le sommeil".

Certains mots-outils sont des prépositions ou des locutions prépositives : "par suite de", "à cause de", "à la différence de". Pour ce type d'outil il est difficile d'établir des schémas syntaxiques rendant compte des termes mis en relation. On peut connaître le correlandum qui suit la préposition, mais l'autre peut avoir un emplacement très variable.

"Troubles de mémoire par suite de traumatisme"

"Par suite de traumatisme le malade a perdu la mémoire"

"Le malade devient amnésique par suite de traumatisme"

C'est pourquoi, dans les cas où on ne peut trancher, on n'établit, pour les prépositions qu'un schéma syntaxique incomplet du type :

préposition + nom

préposition + infinitif

Ex. :
"à la suite de contusions"

"à force de parler"

4.3 - Exemple de classes d'outils

Les schémas syntaxiques ont été utilisés pour affecter les outils à des classes distinctes. Cependant comme critère d'affectation on n'a pas pris le schéma syntaxique mais sa condensation : le schéma syntaxique. Au lieu de contenir des codes spécifiés comme le schéma, le schéma comprend des macro-codes qui représentent une généralisation au niveau de la catégorie grammaticale des mots :

"Nom + préposition + nom + préposition + nom" est le schéma correspondant à la fois à :

Nom	{ compl. ou sujet }	+ de + nom + sur + nom
Nom	{ compl. ou sujet }	+ à + nom + par + nom

Ainsi on peut dire que "effet de... sur...", "réaction à... par" ont le même schéma syntaxique.

Voici le tableau de quelques mots-outils de type consécutif rangés suivant leur schéma syntaxique.

	catégorie de l'outil				cat. et/ ou fonction des corrélanda					nature lexicale de l'outil	
	verbe	nom	adj.	préposit.	nom avant	nom après	adj. après	compl. préposit. 1	compl. préposit. 2	mixte	pur
provoquer	X				X	X					X
inhiber	X				X	X				X	
agir	X				X			X			X
employer	X					X		X			X
régulation		X						X	X	X	
réaction		X						X	X		X
cause		X			X			X			X
origine		X			X		X				X
							X	X			
action		X			X			X	X		X
							X	X			
sensible			X		X			X		X	
consécutif			X		X			X			X
épilep- togène			X		X					X	
par suite de				X			X				X

Un même type de classification a été élaboré pour les mots-outils de type comparatif qui se répartissent suivant une dizaine de classes environ.

Cette classification va faciliter la recherche des descripteurs entrant dans une relation de type consécutif ou comparatif.

4.4 - Repérage des correlanda

Le repérage des correlanda s'opère grâce à des règles de corrélation qui accompagnent chacun des schémas syntaxiques attachés aux types de mots-outils. Ces règles indiquent non seulement quels sont les termes des schémas qui sont mis en relation mais aussi dans quel sens s'oriente la relation.

Exemple, pour "provoque" et "inhibe" (quelques règles seulement) :

type d'outil	Code	schéma syntaxique	règle (1)	Vérif. sé- mantique
provoque	VT	nom sujet + <u>verbe</u> + nom compl. objet	1 <u>3</u> → 3	S
	VT	nom sujet + qui + <u>verbe</u> + nom compl. objet	1 <u>3</u> → 4	
	ST	nom $\left. \begin{array}{l} \text{compl.} \\ \text{ou} \\ \text{sujet} \end{array} \right\}$ + <u>part. passé</u> + prép. PAR + nom compl.	1 ← <u>3</u> 4	
inhibe	VT	nom sujet + <u>verbe</u> + nom compl. objet	$ \begin{array}{c} 2 \quad 2 \quad 2 \\ \swarrow \quad \searrow \\ 1 \quad 3 \quad 3 \end{array} $	S
	VT	nom sujet + qui + <u>verbe</u> + nom compl. objet	$ \begin{array}{c} 2 \quad 3 \quad 2 \\ \swarrow \quad \searrow \\ 1 \quad 3 \quad 4 \end{array} $	
	ST	nom $\left. \begin{array}{l} \text{compl.} \\ \text{ou} \\ \text{sujet} \end{array} \right\}$ + <u>part. passé</u> + prép. PAR + nom compl.	$ \begin{array}{c} 2 \quad 3 \quad 2 \\ \swarrow \quad \searrow \\ 1 \quad 3 \quad 4 \end{array} $	

(1) Le chiffre à gauche de la flèche représente le numéro d'ordre dans le schéma syntaxique du 1er terme mis en relation, le chiffre à droite de la flèche celui du 2ème terme.

Ex. : 1 3 → 3 signifie que c'est le "nom sujet" (1er terme du schéma) et le "nom complément d'objet" (3ème terme du schéma) qui sont mis en relation de type 3 (relation consécutive).

La différence entre les règles de construction de "provoque" et "inhibe" vient du caractère mixte du deuxième outil : en plus de la relation consécutive entre "nom sujet" et "nom compl. objet" s'établissent deux relations associatives entre l'outil lui-même et les deux correlanda "nom sujet" et "verbe", et "verbe" avec "compl. objet".

Ce tableau peut se lire ainsi : si dans une phrase du texte soumis à l'analyse apparaît un outil de type "provoque", suivant le code attribué à cet outil par l'analyse grammaticale (ici VT : verbe conjugué transitif ou ST participe passé), on recherche le ou les schémas syntaxiques dans lequel on peut le rencontrer. Puis, parmi les schémas possibles, on choisit celui qui contient les mêmes catégories que celles accompagnant le mot-outil dans la phrase. Suivant le type de catégories qui forment le schéma on indique la relation qui les relie en spécifiant l'orientation. Ici les règles "sujet - compl. d'objet" et "sujet - compl. d'agent" ont une orientation différente.

Pour que la construction de syntagme soit possible il faut naturellement que les deux termes mis en relation aient une équivalence en SYNTOL. Sinon la règle ne peut s'appliquer telle quelle. Deux solutions peuvent être envisagées : on peut soit accepter un syntagme incomplet si l'un des termes n'a pas reçu de traduction au cours de la phrase lexicale, soit rechercher autour du terme non traduit un qualificatif ou un complément de nom qui pourrait, lui, avoir un équivalent documentaire et faire office de corrélandum à sa place, les deux solutions pouvant être d'ailleurs exploitées concurremment.

Le code "préposition" est suivi dans les exemples donnés de PAR. Ceci signifie que la mise en relation entre le premier terme du schéma et le quatrième n'est acceptée que si le troisième - la préposition - appartient à la liste PAR. Cette liste dressée à l'avance contient une série de prépositions ou locutions prépositives acceptables dans la même position d'un schéma, c'est-à-dire ici susceptible d'introduire un complément d'agent.

PAR =		par
		à l'aide de
		grâce à
		sous l'effet de
	

D'autres listes semblables existent, contenant des prépositions interchangeables dans des situations données.

SUR =		sur
		dans
		chez
		pour
	

Dans la dernière colonne "vérification sémantique", la présence d'un 5 - comme c'est le cas pour la troisième règle dans le tableau - signifie que les informations grammaticales fournies par le schéma syntaxique ne sont pas sûres et qu'il faut avoir recours au réseau notionnel pour vérifier la validité de la construction. En l'occurrence, la chaîne composée d'un nom introduit par une préposition, même si celle-ci appartient bien à la liste PAR, succédant à un

participe passé, lui-même succédant à un nom, peut très bien ne pas représenter une chaîne du type "le sommeil provoqué par l'hexorbarbital" mais par exemple :

"(le sommeil) du malade provoqué par intervalles (par l'hexorbarbital)".

Dans les cas, où une préposition apparaît dans un schéma syntaxique, et qu'elle ne dépend pas d'un nom sujet, la vérification sémantique est appliquée pour que soit attesté le bien fondé de la relation.

Nous venons d'indiquer la démarche générale pour la construction des syntagmes de type consécutif et comparatif. Nous allons voir maintenant comment s'établit la troisième relation, la relation associative.

5 - Construction des syntagmes de type associatif

Dans notre hypothèse de travail (§ 2), nous avons indiqué que toute relation qui n'a pas été interprétée comme comparative ou consécutive doit être interprétée comme associative. Encore faut-il s'assurer qu'une relation existe, autre que consécutive ou comparative, entre les descripteurs qui n'ont pas été reliés, ou entre ceux-ci et les descripteurs entrant dans les syntagmes déjà formés.

Or, dans le cas de la relation associative on ne veut pas utiliser de mots-outils comme repères : tout d'abord parce qu'ils sont trop nombreux et hétérogènes. Ils comprennent pratiquement tous les verbes transitifs directs et indirects, toutes les prépositions, les conjonctions de coordination, les adjectifs susceptibles d'avoir un complément, etc. Il faut également ranger dans ce groupe tous les outils implicites, qui ne sont pas à proprement parler des mots-outils, mais représentent les cas où la seule juxtaposition de deux descripteurs suffit à indiquer une relation de type associatif. En général ces descripteurs correspondent à des noms et adjectifs, ex. : "alvéoles pulmonaires".

De plus, il serait long et difficile de caractériser chaque type d'outil par des schémas syntaxiques, car les constructions dans lesquelles généralement ils s'inscrivent sont mal définies.

Devant cette difficulté de caractérisation syntaxique des mots-outils, on a préféré partir des descripteurs eux-mêmes.

A l'issue des opérations de construction des syntagmes de type consécutif et associatif, un certain nombre de descripteurs du texte soumis à l'analyse n'ont pas été mis en relation. En prenant chacun de ces descripteurs un à un, on essaie de les relier à un second descripteur appartenant à la même unité contextuelle, que celui-ci soit déjà mis en relation ou pas en confrontant les codes grammaticaux et fonctionnels qui leur ont été associés lors de l'analyse syntaxique. Une liste de règle indique quels types de rapports syntaxiques sont à retenir et quelle orientation ils déterminent dans la construction des syntagmes.

Par exemple :

	règle	orientation	validité relation
verbe transitif + nom objet =	1 <u>2</u> → 2		
nom sujet + verbe =	1 <u>2</u> → 2		
nom + adjectif =	1 <u>2</u> 2	S	
nom sujet + compl. prépositionnel =	1 <u>2</u> 2	S	S
nom compl. + compl. prépositionnel =	1 <u>2</u> 2	S	S
nom compl. + nom coordonné =	1 <u>2</u> ↔ 2		S

Dans ce tableau, la présence d'un S dans la colonne "Orientation" indique que les seules informations grammaticales et fonctionnelles ne suffisent pas à indiquer l'orientation de la relation. La règle n'indique pas l'orientation, qui ne sera établie qu'après consultation du réseau notionnel.

6 - Utilisation des informations sémantiques dans les règles de construction des syntagmes

On constate que dans la construction des syntagmes de type 2 aussi bien que pour celle des syntagmes de type 3 ou 4, les informations grammaticales et fonctionnelles fournies par le contexte naturel ne peuvent pas toujours à elles seules répondre de la validité d'une mise en relation. Les constructions jugées douteuses sont vérifiées par consultation du réseau notionnel et acceptées ou rejetées suivant le résultat de la vérification.

Il n'est pas dit que ce contrôle sémantique arrive dans tous les cas à produire des résultats corrects. Il se peut par exemple qu'une relation entre deux descripteurs soit tenue pour valide par le réseau notionnel et cependant qu'elle ne reflète pas l'idée qui veut être exprimée dans un texte. Cependant on peut considérer le contrôle sémantique comme une sorte de filtre supplémentaire ; il écarte un certain nombre d'erreurs que la seule vérification syntaxique est impuissante à éviter et dans le cas d'incertitude pèse sur le choix de l'orientation d'une relation ou de son interprétation.

Ce contrôle se fait à différentes étapes de la traduction syntaxique. Comme on l'a vu aux § 4 et § 5, on a recours au réseau notionnel lors de la construction des syntagmes lorsque le résultat semble incertain ou qu'un choix se révèle impossible du point de vue syntaxique. On l'utilise également lorsque la

traduction syntaxique des unités contextuelles étant terminée, on doit faire le raccordement des différents groupes de syntagmes ou d'isolats pour former un graphe unique. Dans ce paragraphe nous n'aborderons que le premier problème, la vérification sémantique en cours d'analyse, réservant pour le paragraphe 7 l'utilisation du réseau notionnel pour les opérations de complèment.

Nous allons résumer tous les cas où la vérification sémantique se révèle nécessaire en cours d'analyse.

6.1 - Pour la construction des syntagmes consécutifs et comparatifs

a) Lorsque les critères grammaticaux sont insuffisants pour retrouver les correlanda. Cela se produit lorsque l'un des descripteurs du schéma syntaxique correspond à un complément prépositionnel (§ 4.4). Un S inscrit dans la colonne "vérification sémantique" signale la nécessité d'informations sémantiques.

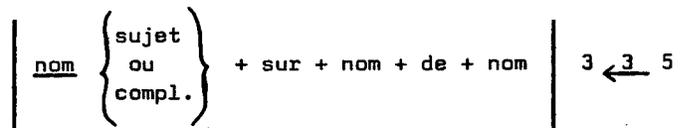
Ex. : Effet sur le sommeil / de l'animal / de l'hexorbarbital / ...

Dans cette phrase il faut choisir entre

sommeil $\leftarrow 3$ animal

et sommeil $\leftarrow 3$ hexorbarbital

puisque les deux compléments "de l'animal" et "de l'hexorbarbital" peuvent, l'un aussi bien que l'autre, faire partie du schéma syntaxique:



En consultant le réseau notionnel on s'aperçoit que la liaison animal-sommeil ne peut donner naissance qu'à une relation associative, orientée du premier vers le deuxième correlandum :

animal $\xrightarrow{2}$ sommeil

tandis que la liaison sommeil-hexorbarbital est bien de type consécutif et peut se traduire par :

hexorbarbital $\xrightarrow{3}$ sommeil

Ce sont donc les deux descripteurs "sommeil" et "hexorbarbital" qui sont retenus ici dans la phrase et mis en relation.

b) La vérification sémantique est également nécessaire lorsque le mot-outil est ambigu. La relation qu'il détermine peut alors avoir plusieurs

sens suivant la classe sémantique des termes mis en relation.

Ex. : emploi de la réserpine pour l'inhibition motrice
emploi d'un outillage particulier pour l'expérience...

L'outil "emploi de... pour..." détermine une relation consécutive dans un cas, alors que la relation est associative - instrumentale ici - dans la deuxième phrase.

Des outils tels que "employer pour", "après", "provenir de"... ont respectivement un sens instrumental, temporel, locatif et peuvent également avoir un sens consécutif. Pour eux, quelles que soient les règles de construction appliquées, la vérification sémantique est obligatoire, et la consultation du réseau fournit à la fois l'interprétation de la relation et son orientation.

6.2 - Pour la construction des syntagmes associatifs

C'est seulement la validité de la relation qui est à vérifier. La relation n'est pas à interpréter puisque si elle existe, on sait qu'elle est associative.

a) Comme dans le cas des autres syntagmes, la vérification sémantique se révèle nécessaire lorsque sont mis en relation des compléments prépositionnels ou des membres de phrase coordonnés.

Ex. : Réaction de croissance / à la lumière / des phycomyces.

La règle syntaxique "nom compl. + nom compl." fait construire pour cette phrase :

soit : lumière $\xrightarrow{2}$ phycomyces

soit : croissance $\xrightarrow{2}$ phycomyces

("croissance" et "lumière" étant déjà reliés dans un syntagme de type consécutif croissance $\xleftarrow{3}$ lumière, il reste à trouver une liaison pour "phycomyces"). La première avec "lumière" n'est pas attestée dans le réseau notionnel et de ce fait est interdite tandis que la deuxième, valide, est acceptée :

croissance $\xleftarrow{2}$ phycomyces

b) La vérification sémantique est également nécessaire dans les cas où une relation entre deux descripteurs étant sûre d'un point de vue syntaxique, son orientation reste indéterminée. Entre un nom sujet par exemple et son complément, une relation associative ne fait pas de doute ; cependant elle ne s'établit pas toujours dans le même sens.

Ex. : La réaction de l'animal....
La réaction de peur....

7 - Résultats obtenus et leur synthèse

A l'issue de l'analyse syntaxique d'un texte et de l'application des règles de construction des syntagmes, les résultats obtenus s'expriment sous des formes diverses.

7.1 - Type de résultats

7.11 - Les syntagmes complets reliés ou non entre eux sous forme de graphes. Ils représentent le type de construction visée par les procédures décrites jusqu'ici.

7.12 - Les syntagmes incomplets composés seulement d'un descripteur et d'une relation spécifiées et orientées.

a) Ces syntagmes incomplets peuvent être dus à l'absence de l'un des corrélandes dans l'unité contextuelle. C'est une absence que l'on note parfois dans les titres des textes analysés.

Ex. : "Etude des effets de la réserpine".

En général un tel titre est complété par une phrase du type :

Ex. : "Effets sur le comportement du chat..."

Ces deux phrases séparées par un signe de ponctuation forte sont analysées séparément et l'on obtient ainsi : réserpine $\xrightarrow{3}$

et $\xrightarrow{3}$ comportement.

b) Les syntagmes incomplets sont également dus à la présence dans la phrase d'un terme naturel n'ayant pas d'équivalence en SYNTOL et cependant mis en relation avec un descripteur.

Ex. : La stimulation provoque un ensemble de troubles.

La relation consécutive s'établit entre "stimulation" et "ensemble". Le second mot n'ayant pas d'équivalent documentaires, on obtient : stimulation $\xrightarrow{3}$

c) De même la présence d'un pronom engendre la construction d'un syntagme incomplet. Il reprend en général un nom contenu dans une unité contextuelle différente et par conséquent non accessible en cours d'analyse.

Ex. : "Effets de la réserpine. Elle provoque des troubles".
"Etude de la réserpine. Son effet sur le comportement".

Dans les deux cas, la deuxième phrase ne contient pas l'agent de l'action ; on obtient :

$\xrightarrow{3}$ troubles

$\xrightarrow{3}$ comportement.

Ex. : les deux syntagmes isolés $A \xrightarrow{3} B$ et $C \xrightarrow{3} D$ peuvent être associés dans un graphe $A \xrightarrow{3} B$ si le réseau notionnel atteste l'existence d'une relation associative entre A et C, ou si ces deux descripteurs appartiennent à la même classe sémantique dans le lexique.

$$\begin{array}{c}
 A \xrightarrow{3} B \\
 \downarrow 2 \\
 C \xrightarrow{3} D
 \end{array}$$

8 - CONCLUSION

Une des caractéristiques de la démarche qui vient d'être décrite est la place qu'y occupent respectivement les données syntaxiques et sémantiques : les différents outils qui ont été définis concourent en effet à n'utiliser des informations grammaticales que dans le cadre relativement restreint où elles sont sûres, et à les amalgamer avec des informations sémantiques provenant du réseau notionnel.

La préparation pratique d'une nouvelle expérience sur ordinateur est en cours, qui déterminera la validité et les limites de cette méthode.

BIBLIOGRAPHY

- Abraham S. Some questions of phrase structure
Grammars, I, CL IV, 1965
- Abraham S. : An integrated language theory (to appear at
Mouton & Co.)
- Abraham S. : A theory of structural semantics
- Kiefer F. : Mouton & Co, The Hague-Paris, 1966
- Chomsky N. : aspects of the theory of syntax, M.I.T.
Press, Cambr., Mass., 1965
- Katz, J.J. : The structure of a semantic Theory
- Fodor, J.A. : Language, 39, 1963
- Katz, J.J. : An integrated theory of linguistic
- Postal P.M. : Description, Cambr. Mass, 1964