DOCUMENTATION
ET INFORMATION

N. BELY / A. BORILLO / J. VIRBEL

N. SIOT-DECAUVILLE

PROCÉDURES D'ANALYSE SÉMANTIQUE APPLIQUÉES A LA DOCUMENTATION SCIENTIFIQUE

GAUTHIER-VILLARS

PROCÉDURES D'ANALYSE SÉMANTIQUE APPLIQUÉES A LA DOCUMENTATION SCIENTIFIQUE

OUVRAGES DE LA COLLECTION

« DOCUMENTATION ET INFORMATION »

Parus :

B. C. VICKERY. — **La classification à facettes.** Guide pour la construction et l'utilisation des schémas spéciaux. Traduit de l'anglais par P. SALVAN.

R. DUBUC. — **La classification décimale universelle.** Manuel pratique d'utilisation.

R.-C. CROS, J.-C. GARDIN, F. LÉVY. — **L'automatisation des recherches documentaires.** Un modèle général : LE SYNTOL.

L'organisation de la documentation scientifique. Études par J.-C. GARDIN, E. DE GROLIER, F. LEVÉRY et l'Association nationale d'études pour la documentation automatique (A. N. E. D. A.).

Z. DOBROWOLSKI. — **Etude sur la construction des systèmes de classification.**

C. LEGEARD. — **Guide de recherches documentaires en démographie.**

Économie générale d'une chaîne documentaire mécanisée. Par F. ALOUCHE, N. BÉLY, R.-C. CROS, J.-C. GARDIN et J. PERRIAULT.

R. CORMIER. — **Les sources des statistiques actuelles.** Guide de documentation.

R. DUBUC. — **Exercices programmés sur la C. D. U.**

Procédures d'analyse sémantique appliquées à la documentation scientifique. Par N. BÉLY, A. BORILLO, N. SIOT-DECAUVILLE et J. VIRBEL.

Classification médicale de la "National Library of Medicine". Traduit par Dr G. NICOLE et M. NICOLE.

DOCUMENTATION ET INFORMATION
Collection dirigée par Paul Poindron,
Directeur des études de l'Institut national
des techniques de la documentation

PROCÉDURES D'ANALYSE SÉMANTIQUE APPLIQUÉES A LA DOCUMENTATION SCIENTIFIQUE

par

N. BELY
A. BORILLO
N. SIOT-DECAUVILLE
J. VIRBEL

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

PRÉFACE DE J.-C. GARDIN

GAUTHIER-VILLARS ÉDITEUR
55, quai des Grands-Augustins, 75 - PARIS - VI
1970

Procédures d'analyse sémantique appliquées à la documentation scientifique, par N. BELY, A. BORILLO, N. SIOT-DECAUVILLE, J. VIRBEL. Préface de J.-C. GARDIN. – Paris, Gauthier-Villars, 1970. – 16 cm, XVIII-242, fig. tableaux.

(Documentation et Information).

TABLE DES MATIÈRES

Pages

Préface de J.C. Gardin XIII

1e PARTIE : CADRE MÉTHODOLOGIQUE

1. Fondements de la recherche 3

2. Paramètres expérimentaux 7

2.1. *Le langage-cible* 7
2.2. *Le domaine et le corpus* 11
2.3. *Démarche générale* 12
2.4. *Etapes de la recherche* 14

2e PARTIE : INDEXATION LEXICALE

3. Démarche de l'indexation lexicale 19

3.1. *Morphologie des termes du langage naturel* 19
3.2. *Les groupes de mots* 20
3.2.1. Définition des groupes de mots 20
3.2.2. Reconnaissance des groupes de mots 21
3.3. *Statuts des termes LN quant à leur traduction* 23
3.4. *Polysémies* 25
3.4.1. Définition 25
3.4.2. Résolution des polysémies 26
3.4.2/1 Polysémies de type "1" : "mots clef" 26
3.4.2/2 Polysémies de type "2" : "catégories grammaticales multiples" 27
3.4.2/3 Polysémies de type "3" : "analyse sémantique seule" 29

Pages

3.4.2/4 Polysémies de type "4" : "analyse syntaxique et sémantique 31
3.4.2/5 Polysémies de type "5" : "descripteurs obligatoires 34

3.5. *Inventaire des types d'outils nécessaires à l'indexation lexicale.* 35

4. Description des outils linguistiques 37

4.1. *Le lexique documentaire* 37
4.2. *Le dictionnaire automatique* 38

4.2.1. Codes affectés aux entrées 38

4.2.1/1 Codes grammaticaux 38
A. Codes de catégories grammaticales 38
B. Codes morphologiques 39
a. Codes affectés aux verbes 39
b. Codes affectés aux autres catégories variables 40

4.2.1/2 Codes affectés aux groupes de mots 42
4.2.1/3 Codes de polysémies 43

4.2.2. Conventions pour la rédaction des articles du dictionnaire 43

4.2.2/1 Transcription des signes diacritiques 44
4.2.2/2 Codes de délimitation des éléments de l'article . . 44
4.2.2/3 Règles pour la rédaction des articles 45
4.2.2/4 Exemples d'articles 45
A. Articles sans règle de résolution polysémique. 45
B. Articles avec règles de résolution polysémique 45

4.3. *Les algorithmes d'analyse* 46

4.3.1. Reconnaissance des groupes de mots 46
4.3.2. Résolution des polysémies 46

4.3.2/1 Type "2" 46
4.3.2/2 Type "3" 47
4.3.2/3 Type "4" 48

5. Le programme 49

5.1. *Les données* 49

5.1.1. Le lexique 49
5.1.2. Les désinences morphologiques 49

5.1.2/1 Désinences de noms ou d'adjectifs 49
5.1.2/2 Désinences de verbes 50

5.1.3. Le dictionnaire 50

Pages

5.2. *Le traitement* 50
5.2.1. Introduction des données 50
5.2.1/1 Introduction du lexique 50
5.2.1/2 Introduction des désinences 51
5.2.1/3 Introduction du dictionnaire 51
5.2.2. Traitement des résumés 52
5.2.2/1 Première étape 52
5.2.2/2 Seconde étape 52
5.2.2/3 Troisième étape 52
5.2.2/4 Quatrième étape 53
5.2.2/5 Cinquième étape 53
5.3. *Exploitation* 54

6. Résultats 55
6.1. *Reconnaissance des groupes de mots* 55
6.2. *Résolution des polysémies* 56
6.2.1. Résultats d'ensemble 56
6.2.2. Algorithme "1" 58
6.2.3. Algorithme "2" 58
6.2.4. Algorithme "3" 61
6.2.4/1 Contexte immédiat ("I") 62
6.2.4/2 Contexte de la phrase ("P") 63
6.2.4/3 Contexte du résumé ("R") 63
6.2.5. Algorithme "4" 65
6.2.5/1 Résolution des substantifs 66
6.2.5/2 Résolution des adjectifs 68
6.2.5/3 Les non-résolutions 69
A. Structures non-reconnues 70
B. Structures vides 70
6.2.6. Algorithme "5" 71
6.3. *Bilan d'ensemble* 71

3e PARTIE : INDEXATION SYNTAXIQUE

7. Exposé de méthode 77
7.1. *Nature des relations logiques* 77
7.2. *Format des relations logiques* 78
7.3. *Passage du langage naturel à la représentation documentaire* . . . 80
7.3.1. Méthode de reconnaissance et d'extraction des relations logiques 82
7.3.2. Utilisation du réseau notionel 84

Pages

8. Analyse de l'énoncé en langage naturel 87

8.1. *Type d'analyse* 87
8.2. *Exploitation de l'énoncé* 88
8.3. *Méthode d'analyse* 90

8.3.1. La reconnaissance des catégories grammaticales 90
8.3.2. La caractérisation de la fonction 92

8.4. *Résultats de l'analyse* 98

9. Extraction des relations logiques 101

9.1. *Extraction des relations consécutives et comparatives* 101

9.1.1. Les mots-outils 101
9.1.2. Les schémas syntaxiques 103
9.1.3. Règles de construction des syntagmes 105
9.1.4. Recherches des suppléants et des coordonnés 106
9.1.5. Vérification sémantique 108

9.2. *Extraction des relations associatives* 109

9.2.1. Construction des syntagmes 110
9.2.2. Restrictions 111
9.2.3. Vérification sémantique 111

10. Synthèse des résultats 115

10.1. *Types de résultats* 115

10.1.1. Des syntagmes complets 115
10.1.2. Des syntagmes incomplets 115
10.1.3. Des descripteurs isolés ou isolats 116

10.2. *Traitement des interprétations multiples* 116
10.3. *Opération de raccordement* 117
10.4. *Les opérations de développement* 118

11. Programme 119

11.1. *Organisation générale* 119
11.2. *Introduction des données permanentes* 120

11.2.1. Les codes grammaticaux et fonctionnels 121
a) Forme externe 121
b) Forme interne 122

11.2.2. Régles d'analyse grammaticale 122
a) Forme externe 122
b) Forme interne 122

Pages

11.2.3. Schémas 122
a) Forme externe 122
b) Forme interne 123
11.2.4. Coordonnés et suppléants 124
a) Forme externe 124
b) Forme interne 124
11.2.5. Règles associatives 124
a) Forme externe 124
b) Forme interne 125
11.3. *Traitement des résumés* 125
11.3.1. Lecture des résumés 125
Traitement des parenthèses 125
11.3.2. Interprétation syntaxique d'une phrase (organigramme 2) 126
Format des résultats partiels (Voir Annexe 21B) 126
11.3.3. Recherche des syntagmes à l'aide des schémas syntaxiques (organigramme 3) 128
11.3.3.1. Orientation de l'exploration 128
11.3.3.2. Recherche des suppléants et des coordonnés 128
11.3.3.3. Médiation 130
11.3.3.4. Format des résultats partiels (Voir Annexe 21B) 130
11.3.3.5. Interprétations multiples 131
11.4. *Recherche des syntagmes associatifs* (organigramme 4) 131
11.4.1. Exploration de la phrase 131
11.4.2. Format des résultats partiels (Voir Annexe 21B) . . . 133
11.4.3. Les isolats 134
11.5. *Les raccordements* 134
11.6. *Impression des résultats finals* 134
11.6.1. Condensation à l'intérieur d'un résumé (programme 6) 134
11.6.2. Format des résultats finals (Voir Annexe 21C) (programme 7) 134
11.7. *Exploitation* 135

Pages

12. Résultats de la phase syntaxique d'indexation 137
12.1. *Evaluation de l'analyse syntaxique* 137
12.1.1. Déroulement de l'analyse 138
a) Arrêts dûs à des altérations dans les données 138
b) Arrêts dûs aux règles 138
12.1.2. Résultats de l'analyse 140
12.2. *Résultats de la recherche des relations consécutives et comparatives* 142
12.2.1. Evaluation quantitative des résultats 142
a) Les mots-outils 142
b) Les mises en relations 143
12.2.2. Evaluation de la procédure de recherche des syntagmes 144
a) Les syntagmes complets 144
b) Les syntagmes incomplets 146
c) Absence de mise en relation 147
12.2.3. Examen des résultats en fonction de leur valeur informative 150
12.3. *Résultats de la recherche des relations associatives* 153
12.3.1. Evaluation qualitative des résultats 153
12.3.2. Evaluation de la procédure de recherche des syntagmes associatifs 154
12.3.2/1 Les mises en relation incorrectes 154
12.3.2/2 Les isolats 155
12.3.3. Examen des résultats en fonction de leur valeur informative 158
12.4. *Valeur des opérations de condensation et de raccordement* . 159
Conclusion 161
Annexes 165

TABLE DES ANNEXES

		Pages
ANNEXE 1 :	Catégories grammaticales	167
ANNEXE 2 :	Codes morphologiques des catégories variables (autres que le verbe)	169
ANNEXE 3 :	Codes morphologiques des verbes	172
ANNEXE 4 :	Reconnaissance des groupes de mots (GpM)	175
ANNEXE 5 :	Résolution des polysémies type "2" : homographies substantif-verbe (N.V) et substantif-participe présent (N.T.)	176
ANNEXE 6 :	Résolution des polysémies type "2" : homographies adjectif-verbe (A.V.)	177
ANNEXE 7 :	Résolution des polysémies type "2" : homographies substantif-adjectif (N.A.) et substantif-verbe-adjectif, (N.V.A.)	178
ANNEXE 8 :	Résolution des polysémies type "2" : homographies substantif-participe passé (N.B.)	179
ANNEXE 9 :	Résolution des polysémies type "4" : analyse syntaxique	180
ANNEXE 10 :	Echantillon du dictionnaire automatique	181
ANNEXE 11 :	Echantillon du lexique	184
ANNEXE 12 :	Echantillon des règles d'analyse syntaxique	187
ANNEXE 13 :	Liste des mots outils	193
ANNEXE 14 :	Schémas consécutifs (exemples)	198
ANNEXE 15 :	Schémas comparatifs (exemples)	201
ANNEXE 16 :	Table des coordonnés et des suppléants	204

Pages

ANNEXE 17 : Recherche des coordonnés et des suppléants 207
ANNEXE 18 : Règles associatives 208
ANNEXE 19 : Echantillon du réseau notionnel 210
ANNEXE 20 : Echantillon du listing de la sortie lexicale 213
ANNEXE 21 : Echantillon d'analyse syntaxique 225
21-A : Texte soumis à l'analyse syntaxique 226
21-B : Résultat intermédiaire 234
21-C : Résultat final de l'analyse syntaxique 240

PRÉFACE

On voit paraître de temps en temps, sous couvert de recherches sur les techniques documentaires, des ouvrages d'une allure un peu ésotérique, dont le sens le plus riche n'est pas nécessairement celui qui se donne le plus immédiatement. Tel est selon nous le cas du livre rédigé par Andrée Borillo et Jacques Virbel, au terme d'une étude de trois ans menée conjointement avec Nathalie Bely et Nelly Siot-Decauville, au Laboratoire d'Automatique Documentaire et Linguistique du Centre National de la Recherche Scientifique. L'objet de cette étude semblait pourtant simple, et bien défini : il s'agissait de trouver une démarche relativement générale pour mécaniser l'analyse des textes scientifiques, telle que la pratiquent les documentalistes sous des noms divers (indexation, classification, catégorisation, etc.). La mécanisation de l'indexation – pour nous en tenir à cette appellation désormais bien acquise – a un intérêt d'abord pratique : des dizaines de milliers de spécialistes consacrent une part notable de leur temps à exprimer le contenu de documents scientifiques toujours plus nombreux, en vue de faciliter les recherches rétrospectives ultérieures ; le recrutement et la formation d'analystes compétents, pour cette tâche, sont de plus en plus difficiles, et il est naturel que l'on cherche à contourner l'obstacle, ici comme ailleurs, par la mécanisation de celle-ci. Le mérite premier de l'étude qui suit est d'apporter des indications enfin précises sur le coût de cette solution, si c'en est une. Entendons-nous, en effet : nos auteurs n'entendent pas vanter les avantages de l'indexation automatique par rapport aux méthodes d'analyse traditionnelles, ou comme on dit "manuelles" ; leur propos est seulement de montrer le genre d'outils qu'il faut fournir à la machine pour qu'elle puisse analyser elle-même le contenu de textes scientifiques donnés, dans des termes comparables à ceux qu'auraient pu préconiser des interprètes humains.

Par outils, il faut entendre évidemment des outils de calcul, au sens large du mot, et plus précisément, dans le cas qui nous occupe, un ensemble ordonné de règles assurant le passage automatique d'un texte écrit dans une

langue naturelle -- en l'occurence, le français -- à une représentation de ce texte qui soit censée en exprimer le sens, du point de vue largement intuitif où se placent habituellement les documentalistes. On ne saurait être surpris de la complexité relative des règles en question, telles qu'elles sont illustrées par l'expérience d'indexation automatique dont ce livre est le compte rendu : l'analyse sémantique d'un texte scientifique, fût-il déjà résumé, est une opération éminemment intelligente, qui exige une double compétence, sur le plan de la langue tout d'abord, mais aussi sur le plan de la pensée scientifique elle-même, puisqu'enfin l'on n'attend plus aujourd'hui d'un documentaliste omniscient qu'il soit capable de dégager indifféremment le sens d'un article de physique théorique ou de sociologie. La machine doit être instruite de la même manière dans ces deux ordres de compétence ; et l'on trouvera dans les pages qui suivent une bonne mesure de ce qu'il en coûte, sous forme de ce que les auteurs appellent judicieusement l'"investissement intellectuel" nécessaire : construction de dictionnaires et réseaux sémantiques spécialisés, exprimant une certaine organisation de la connaissance scientifique dans un domaine particulier, élaboration de grammaires *ad hoc*, assurant une relative normalisation du discours sur le plan du lexique et de la grammaire, établissement de programmes permettant la mise en œuvre de ces outils logico-linguistiques par une machine, telles sont les tâches initiales que présuppose l'indexation mécanique, dès lors qu'on attend d'elle des résultats comparables à ceux d'une analyse conduite par des cerveaux humains.

Quant à la rentabilité de cet investissement, elle est à son tour illustrée par les temps de calcul observés au cours de l'expérience, sur un nombre suffisamment élevé de résumés scientifiques — un millier environ — pour fournir au moins un ordre de grandeur des coûts, en termes d'argent cette fois. Dès maintenant, il apparaît que l'automatisation de l'analyse documentaire est une entreprise justifiée, du point de vue économique, sans que l'on ait à transiger sur la qualité finale du produit ; et tout porte à croire que la balance ne cessera de pencher davantage en faveur de cette option, à mesure que le coût des machines baissera par rapport à celui du travail humain. Mais la décision, sur ce point, relève d'une politique à long terme de l'information scientifique, qui dépasse l'objet d'une étude théorique.

Théorique, cette étude l'est en effet par la nature des problèmes qu'elle pose, au moins de façon implicite, au-delà de son objectif appliqué immédiat. L'un d'eux est évoqué au début du livre : c'est celui de la généralité du modèle utilisé pour définir le langage-cible de l'indexation, à savoir ici le SYNTOL, déjà décrit dans un ouvrage de la présente collection. La thèse des auteurs, selon laquelle la plupart des "langages documentaires" existants peuvent être décrits ou traduits dans le formalisme binaire du SYNTOL, sans perte d'informations notable, mériterait à elle seule une étude particulière. Non qu'elle nous semble fragile : bien au contraire, elle se trouve

plutôt renforcée par l'apparition d'études récentes, notamment en U.R.S.S. et aux Etats-Unis, où les langages d'information proposés, sous des noms divers, présentent des caractéristiques formelles tout-à-fait comparables à celles du SYNTOL, tant sur le plan de l'organisation sémantique (ex. : les "data bases", reductibles – et parfois déjà réduites, par leurs auteurs même – à des réseaux de relations binaires), que sur le plan de l'expression syntaxique proprement dite, d'où le SYNTOL tire son nom. Nous avons déjà eu l'occasion de signaler ce fait en présentant la seconde édition du livre consacré au SYNTOL, il y a deux ans ; bornons-nous à en marquer une fois encore l'importance, dans la perspective nouvelle de l'indexation mécanique. En effet, s'il était établi que les langages-cible de cette opération font appel à des modes d'expression fondamentalement voisins, au-delà des différences de vocabulaire, et que ces modes d'expression entrent eux-mêmes sans peine dans les schémas binaires du SYNTOL, la démarche décrite dans les pages qui suivent trouverait du même coup un champ d'application extrêmement étendu, au moins pour ce qui concerne l'analyse de textes scientifiques rédigés (ou résumés) en français.

Un malentendu risque cependant de se répandre à ce sujet. La convergence des langages documentaires, sur le plan des structures logiques mises en œuvre, conduit certains auteurs à voir en eux autant de modèles approchés d'un hypothétique "langage (unique) de la science", débarrassé des impropriétés que manifestent les langues naturelles au regard des impératifs du discours scientifique. Et l'on croit trouver une confirmation de cette thèse dans le fait que les énoncés du langage documentaire se prêtent parfois à des calculs qui rappellent ceux de la logique propositionnelle : détection de tautologies ou de contradictions, déductions, implications, etc. De là à suggérer que l'analyse sémantique des textes scientifiques est par conséquent une activité de recherche, et non de documentation, il n'y a malheureusement qu'un pas ; on saura gré aux auteurs du présent ouvrage de n'avoir à aucun moment pu laisser entendre qu'ils l'avaient franchi, alors même qu'ils exposaient des procédures apparentées a celles de l'"intelligence artificielle", comme il est aujourd'hui d'usage de les nommer (voir par exemple § 10). Mais la question n'en demeure pas moins fondée, du moins en théorie : au fur et à mesure que le langage-cible de l'indexation tend à prendre des formes justiciables d'un "calcul sémantique" – l'expression est de plus en plus fréquente dans la littérature de l'informatique documentaire – l'algorithme d'analyse (du langage naturel au langage d'indexation) tend lui-même à devenir un outil de raisonnement que l'on peut manipuler à des fins autres que celles de la recherche rétrospective d'informations, *stricto sensu.*

En pratique, et s'agissant de disciplines où le langage scientifique est solidement fondé, cette éventualité n'est qu'une vue de l'esprit : nous n'en sommes pas encore arrivés au moment où les outils syntaxiques et séman-

tiques de l'analyse documentaire, en physique ou en astronomie, pourraient conduire à des découvertes dont le langage propre de ces sciences, hautement formalisé, n'aurait pas été déjà porteur. . . La situation n'est pas la même, néanmoins, dans certains domaines de recherche apparentés à ce que l'on est convenu d'appeler les sciences de l'homme. Là, en effet, les exigences logiques de l'analyse automatique du langage *spécialisé* en arrivent à révéler des inconséquences, sinon des modèles cachés, dont les spécialistes eux-mêmes commencent à reconnaître l'utilité, du point de vue de la connaissance scientifique, et non plus seulement documentaire de leur champ. Les applications de l'informatique au traitement de la documentation médicale ou juridique, par exemple, ont parfois des conséquences théoriques de cet ordre ; et le lecteur curieux de ces incidences pourra s'en faire une idée assez bonne, ici même, en essayant d'imaginer la forme et le contenu qu'auraient dû revêtir les outils de l'indexation mécanique, tels qu'ils sont définis plus loin, si les textes considérés avaient été tirés de la sociologie, par exemple, et non de la physiologie.

Cet exercice ne serait d'ailleurs pas purement spéculatif : notre lecteur aurait en effet à sa disposition un nombre déjà notable de "dictionnaires" conceptuels établis dans des circonstances assez voisines, à savoir comme outils d'une analyse automatique de textes très divers, intéressant le sociologue, le psychanalyste, l'anthropologue, etc. Les plus connus, sinon les plus élaborés, sont ceux que l'on associe généralement au système de programmation construit sous la direction de Philip Stone, à l'Université de Harvard, sous le nom de "General Inquirer". Le but de ce système est, comme ici, la mécanisation de l'analyse sémantique visant des textes écrits dans une langue naturelle, en l'occurence, l'anglais ; et les moyens employés sont largement les mêmes, à savoir un langage-cible d'indexation possédant son lexique et sa grammaire propres. On définit l'analyse mécanique, à nouveau, comme l'application de règles logico-linguistiques établissant les correspondances voulues entre la langue naturelle d'entrée et ce langage artificiel de sortie. Dans l'état présent du General Inquirer, cependant, ce calcul consiste pour l'essentiel en des consultations de table visant la seule "phase lexicale" de l'indexation, selon la terminologie adoptée dans les pages qui suivent : la machine remplace les mots du texte par les descripteurs équivalents indiqués dans tel ou tel dictionnaire *ad hoc*. Les opérations plus raisonnées qu'impliquent la résolution des polysémies et, plus encore, l'attribution des fonctions logiques dans la phase syntaxique de l'analyse, sont laissées pour le moment à la diligence de "pré-éditeurs" ou de "post-éditeurs", dont le rôle demeure prépondérant.

Ces différences, autant que l'analogie de la visée, donnent plus de relief encore à l'exercice de nos quatre auteurs : l'invention de règles capables d'assurer sans faillir – ou sans beaucoup faillir – l'interprétation automatique de termes homographiques ou la construction mécanique de

graphes syntaxiques même rudimentaires, comme le sont ceux du langage-cible dans l'expérience relatée plus loin, était déjà un projet bien ambitieux ; Andrée Borillo et Jacques Virbel ont su lui trouver un aboutissement qui fera date, même si l'on ne retient pas à l'avenir tous les tours de leur démarche. L'intégration de ces règles en un algorithme d'analyse déjà remarquablement efficace, malgré les limites avouées d'un programme tourné vers l'expérimentation plus que vers la performance, est un accomplissement non moins méritoire, dû à la science de Nathalie Bely et de Nelly Siot-Decauville en matière d'informatique non-numérique. Aussi devrait-on souhaiter qu'un tel travail ait maintenant des suites, dans l'une ou l'autre des directions qui viennent d'être suggérées : applications concrètes dans le domaine de la documentation automatique, études comparatives des langages-cible de l'analyse documentaire sur le plan sémantique et syntaxique, relations entre ce genre d'analyse et celle que pratiquent les exégètes de textes de toutes sortes dans les sciences de l'homme (articles de presse, biographies, contes populaires, mythes, etc.), lorsqu'ils cherchent de la même manière à objectiver les opérations sous-jacentes, pour les besoins de la mécanisation, si ce n'est pour d'autres raisons plus sérieuses, etc.

Les prolongements possibles ne manquent pas, on le voit ; il en est un que nous avons délibérément gardé pour la fin de cette courte présentation, cependant, parce qu'il est à nos yeux le plus évident, et sans doute aussi le plus fécond. Tout lecteur averti des progrès de l'analyse linguistique, dans les dernières décennies, devinera certaines parentés entre le projet appliqué qui fait la matière de ce livre et les recherches théoriques de la linguistique formelle, notamment dans le domaine de l'analyse syntaxique. On sait en effet qu'un des objectif de celle-ci est de définir des suites d'opérations – un calcul, par conséquent – permettant de reconnaître ou d'engendrer des propositions tenues pour équivalentes, à certaines nuances près, sur le plan des structures logiques profondes. Sans entrer ici dans les discussions toujours ouvertes concernant les frontières que l'on croit alors devoir affirmer, ou au contraire nier, entre syntaxe et sémantique, analyse de la langue et analyse de la pensée, etc., bornons-nous à souligner le parallélisme apparent entre cet objectif de l'analyse transformationnelle, en linguistique, et le but de l'analyse documentaire elle-même. Non qu'il soit raisonnable de considérer celle-ci comme une variante de celle-là : la vérité, plutôt, est que les macro-mécanismes de l'analyse documentaire, tels qu'ils sont posés dans le présent ouvrage, sont des sortes de "court-circuits" d'une analyse plus fine du discours, au sens où l'entendent les transformationalistes (Z.S. Harris et ses épigones aux Etats-Unis, Maurice Gross en France, etc.). Il n'est pas dit cependant que les premiers ne puissent aider au progrès de la seconde, en raison même de cette relative parenté des visées ; les travaux de linguistes comme I. Mel'Chuk et I. Jolkovski, en U.R.S.S., le suggèrent en tout cas fortement, tant le concept de (méta) langage-cible y joue un rôle

essentiel, dans l'expression des structures "syntactico-sémantiques" sous-jacentes aux formes d'expression plus diverses des langues naturelles. La double étiquette du Laboratoire d'Automatique *Documentaire* et *Linguistique*, où les recherches ci-dessous ont pu naître – avec l'appui décisif, en son temps, de la Délégation Générale à la Recherche Scientifique et Technique – laisse bien augurer de cette interaction nécessaire entre modèles théoriques et systèmes appliqués, dans l'analyse automatique des textes scientifiques.

J.C. Gardin