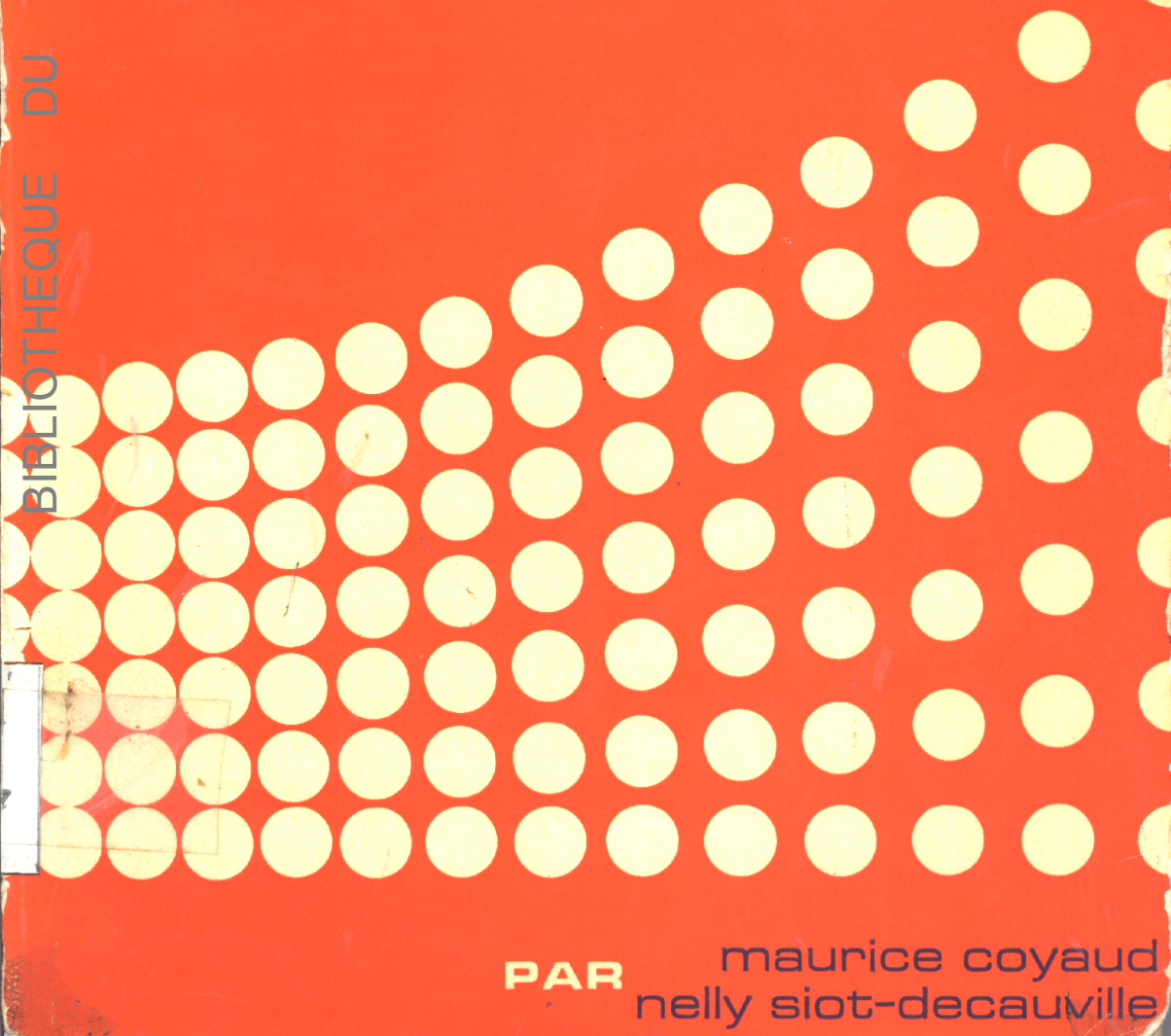
l'analyse

AUTOMATIQUE des DOCUMENTS



PAR maurice coyaud
nelly siot-decauville

INFORMATIQUE
I

Maurice COYAUD
Nelly SIOT-DECAUVILLE

L'analyse automatique des documents

mouton
paris
la haye

PLAN

ABRÉVIATIONS

I.C.C.L.	International Congress for Computational Linguitics.
S.D.C.	System Development Corporation.
A.M.T.C.L.	Association for Machine Translation and Computational Linguistics.
Amer. Doc.	American Documentation.
Congrès de Moscou	Conference on Information Processing, Machine Translation and Automatic Reading of Text, 1961.
Congrès de Cleveland	International Conference for Standards on a Common Language for Machine Searching and Translation, New York, Interscience Publishers, 1961.
J.P.R.S.	Joint Publication Research Service, Washington.
Lg.	Language.
I.F.I.P.	International Federation of Information Processing.
C.E.A.	Commissariat à l'Energie Atomique.
T.D.A.P.	Transformation and Discourse Analysis Papers. University of Pennsylvania.

A.G.	Analyse grammaticale.
T.	Transformations.
M.A.G.	Métalangage d'analyse grammaticale.
L.N.	Langue naturelle.
L.D.	Langage documentaire.
L.I.	Langage intermédiaire.

AVANT-PROPOS

1. *Définition*

L'analyse documentaire automatique est un ensemble d'opérations, relevant des mathématiques, de la linguistique, de la programmation, destinées soit à sélectionner certains éléments d'un document sans les changer (notre chapitre 1), soit à transformer la forme et le contenu du document (notre chapitre 2), soit à combiner les deux types de méthodes précédents (notre chapitre 3). Les liens entre la synthèse et l'analyse automatique de données linguistiques sont étroits. Ce n'est pas un hasard si un des chercheurs en analyse syntaxique des textes a pu nommer sa méthode « analyse par synthèse ». Les travaux ne manquent pas sur le thème de la synthèse automatique : « génération » de phrases (Yngve), synthèse de la parole (vocoder, etc.). Nous n'aborderons néanmoins pas le problème des rapports entre analyse et synthèse, qui pourrait faire l'objet d'une monographie séparée.

2. *Intérêt théorique de l'étude*

Il est inutile de souligner l'importance pratique des travaux sur l'analyse automatique des documents. On sait qu'il s'agit là d'un des goulots d'étranglement de la documentation automatique, et que c'est un domaine où il y a le plus de progrès à faire. Par contre, son intérêt théorique n'apparaît pas toujours. Bornons-nous à noter que la définition d'unités minimales de traitement de l'information documentaire est une pierre de touche pour le formalisme linguistique (Ecole de Harris, etc.). C'est un problème théorique important que la définition des moyens de traiter automatiquement les textes linguistiques : pour la segmentation de la chaîne écrite en unités minimales, pour le calcul de la structure de la chaîne (fonctions des morphèmes, rapports implicites entre lexèmes, « analyse du discours »), etc.

3. *Limites de la présente étude*

Nous circonscrirons notre tâche à l'examen des recherches en analyse automatique d'un point de vue surtout linguistique ; en outre, nous nous limiterons à l'analyse de la chaîne écrite.

Un point de vue linguistique : Les questions de programmation et les questions logiques de théorie des langages formels ne seront pas abordées directement. Naturellement, dans la deuxième partie de cet ouvrage, consacrée à une étude expérimentale, on exposera en

détail le programme utilisé ; mais au cours de la première partie, où nous ferons un examen général des méthodes d'analyse automatique, il ne sera pas question de programmes, même d'un point de vue général. Notons néanmoins que c'est là un thème fort souvent abordé, notamment dans la bibliographie américaine. Citons simplement, à titre d'exemple, la communication de P. Garvin à la Conférence Internationale de 1965 sur la linguistique appliquée aux automates[1]. Le problème qu'il aborde a son importance ; il s'agit de savoir s'il faut créer un algorithme général, capable de traiter (d'absorber) toutes les grammaires, ou s'il faut au contraire — par nécessité de fait — adapter l'algorithme à la grammaire. A l'encontre des chercheurs qui semblent pencher pour la première solution, Garvin choisit la seconde, avec des arguments fort probants[2]. L'essentiel de ces arguments est qu'il n'est pas possible d'imaginer un algorithme assez général pour traiter une grammaire sensible au contexte.

Une autre implication du point de vue prédominant choisi ici est que notre étude se bornera aux problèmes d'analyse interne[3] des documents. Nous ne ferons donc que mentionner les nombreuses recherches sur l'analyse externe, notamment la constitution automatique d'index de références bibliographiques citées par les différents auteurs dans leurs articles (« citation indexes »). Le lecteur pourra se reporter à ce sujet aux travaux cités dans la bibliographie sous les noms de Lesk, Thompson et B.-A. Lipetz.

D'autre part, nous n'aborderons pas les problèmes de transcription ou transformation des notations symboliques artificielles (notation chimique, etc.) ; une bibliographie très abondante leur est consacrée : ces études constituent un tout en elles-mêmes, et méritent une monographie séparée, du fait qu'elles sont indépendantes dans une large mesure, des études concernées par le présent ouvrage.

L'analyse de la chaîne écrite : Nous nous limiterons enfin à l'étude des travaux concernant l'analyse de la chaîne écrite. Malgré des liens évidents, les problèmes posés par l'analyse de la chaîne parlée sont fort différents (linguistiquement et techniquement), et nous les laisserons entièrement de côté.

1. *Plan général*

Le présent ouvrage est divisé en deux parties. La première doit servir à dresser un tableau général des recherches effectuées sur le problème de l'analyse automatique documentaire. Ce tableau a des lacunes (par exemple en ce qui concerne les recherches en Union Soviétique)

1. « Some comments on algorithm and grammar in the automatic parsing of natural language », International Conference on Computational Linguistics, New York, 1965.

2. La même position est adoptée par Jane Robinson ; cf. « Endocentric constructions and the Cocke parsing logic », p. 1 (*ibid*).

3. C'est-à-dire fondée sur les éléments proprement linguistiques du document : le texte de l'auteur, débarrassé d'éléments périphériques comme les références bibliographiques, les schémas, etc.

mais nous n'avons pas la prétention d'être complets. Plus qu'un inventaire des recherches effectuées[1], c'est un inventaire des problèmes fondamentaux que nous avons en vue.

La seconde partie de ce livre est consacrée à l'exposé d'une expérience d'indexation automatique de documents, effectuée à la Section d'Automatique Documentaire du C.N.R.S. en 1962-65.

L'expérience d'indexation automatique qui fait l'objet de la deuxième partie de ce livre a été menée sous la direction de M. J.-C. Gardin, à qui nous exprimons nos vifs remerciements. Nous remercions également Mlle Radmila Zygouris, qui a constitué le « réseau notionnel » expérimental, en psychophysiologie, Mlles M. P. Ferry et C. Charmot, qui ont participé à la compilation du dictionnaire expérimental[2].

La partie linguistique de la présente étude est due à M. Coyaud ; N. Siot-Decauville est l'auteur du programme expérimental d'indexation automatique vers le langage documentaire Syntol, et a rédigé la section 4.3 de ce livre.

M. C. et N. S.-D.

1. On trouvera sur ce thème des bibliographies quasi exhaustives dans *Current research and development in scientific documentation*, publié périodiquement par la National Science Foundation, et dans un rapport récent de Mary Stevens : « Automatic Indexing : A State-of-the-Art-Report », NBS, Washington, 1965.

2. La compilation de ce dictionnaire a été effectuée grâce à un contrat d'un an accordé par l'Euratom, qui s'est également chargé de la perforation du corpus expérimental de cinq cents résumés. A la suite de ce contrat (1961-1962), un rapport final a été adressé à l'Euratom ; il est maintenant publié par les Presses Académiques Européennes à Bruxelles : le *Syntol*, t. 4, « Analyse automatique » (1964).

0. TYPOLOGIE GÉNÉRALE

Les méthodes d'analyse automatique des documents peuvent être classées en deux grandes catégories, suivant la nature des opérations mises en jeu : ou bien celles-ci consistent en un pur *choix* de certains éléments compris dans les documents (suivi d'un réarrangement ou d'une extraction éventuels de ces éléments) ; ou bien il s'agit d'une *transformation* du contenu du document. Nous donnerons ci-dessous des exemples détaillés et commentés de méthodes appartenant à ces deux catégories. Indiquons d'abord de façon générale un point de vue qui permet d'apprécier pratiquement la différence entre ces deux grandes catégories : le point de vue des résultats. La première catégorie de méthodes a produit des résultats passés dans l'usage courant depuis plusieurs années : index de permutations, index KWIC (Key Word in Context), entre autres. Le second groupe de méthodes, plus ambitieuses celles-là, n'a pas encore porté à terme ses fruits. A cela, rien d'étonnant, dès qu'on pense à la complexité des opérations engagées dès qu'il ne s'agit plus de *choisir*, sans les modifier, des éléments dans les textes, en vertu de critères statistiques, ou autres, mais de *transformer* le contenu de ces textes dans un langage (artificiel ou naturel) plus condensé et plus précis.

Les produits obtenus peuvent alors revêtir des formes bien diverses : soit le document est soumis à une *analyse grammaticale* tendant à manifester la structure des phrases — et dans ce cas, les graphes obtenus constitueraient des représentations adéquates des documents ; soit le document est traduit dans un langage documentaire[1] et cette traduction est appelée *indexation*, parce qu'elle consiste essentiellement à dégager les notions les plus importantes du document (et parfois leurs relations sémantiques) à l'aide de termes empruntés à un lexique documentaire (*i.e.* un index de descripteurs ou de mots-clés, ou *indexing language*) ; dans un troisième cas le document pourrait être transformé en un *résumé* en langue naturelle, analogue aux résumés faits par des analystes humains. Ce dernier type d'analyse documentaire est évidemment le plus ambitieux, et ne se verra pas illustré par des résultats avant longtemps ; mais puisqu'il y a des chercheurs qui travaillent dans cette voie, il serait injuste de ne pas évoquer leurs travaux, ne serait-ce que pour mémoire.

1. Le langage documentaire sera défini au § 1.20.