COMMUNAUTE EUROPEENNE
DE L'ENERGIE ATOMIQUE

EURATOM

La Commission

Direction Générale
Recherches et Enseignement

CETIS

ENSEIGNEMENT PREPARATOIRE AUX TECHNIQUES DE LA DOCUMENTATION AUTOMATIQUE

BRUXELLES, 15 - 22 février 1960

ENSEIGNEMENT PREPARATOIRE AUX TECHNIQUES DE LA DOCUMENTATION AUTOMATIQUE

EURATOM, Bruxelles, 15-22 février 1960

AVERTISSEMENT

Nous avons retardé à plusieurs reprises la parution des conférences prononcées lors de notre session de février 1960 afin de tenir compte des développements nouveaux qui avaient lieu tant dans nos propres recherches que dans celles de nos collègues des autres Centres. Mais finalement il nous est apparu qu'une publication au moins provisoire ne pouvait pas être différée davantage et c'est pourquoi nous avons adopté le compromis suivant: nous publions l'ensemble des textes sous leur forme originale et dans une présentation modeste (ronéotypée). Par contre nous préparons pour la fin de l'année 1961 ou le début de 1962, plusieurs publications plus spécialisées et qui rendront compte des derniers développements. Ces publications seront faites lorsque se manifestera un palier dans les recherches. La présentation en sera plus soignée. Les publications actuellement prévues sont les suivantes :

- Méthodologie de la traduction automatique
- Méthodologie de la documentation automatique
- Les aspects aléatoires et les aspects certains dans la notion d'information
- Les ordinateurs mixtes.

En attendant ces futures publications nous espérons que le présent texte pourra être de quelque utilité (malgré l'absence de trois textes qui n'ont pu nous parvenir à temps).

Qu'il nous soit permis à cette occasion de remercier tous ceux qui nous ont aidés à organiser cet enseignement et, en particulier, en plus de tous les professeurs et participants, les responsables de IBM Belgique (notamment MM. HIRSCHBERG et DEBROUX) et de l'Université Libre de Bruxelles (MM. GILLIS et MORLET) qui nous ont permis de réaliser l'expérience d'analyse automatique, et MM. GUERON, Directeur Général des Recherches et de l'Enseignement, MEDI, Vice-président et HIRSCH, Président de la Communauté Européenne de l'Energie Atomique (EURATOM), qui nous ont fait l'honneur de prononcer les allocutions d'ouverture et de clôture de l'Enseignement.

S O M M A I R E

* Cet article ne nous est malheureusement pas parvenu à temps.

JOURNEE DES SYSTEMES DOCUMENTAIRES

JOURNEE D'ANALYSE

JOURNEE D'AUTOMATISATION

"Mais pour rendre cette méthode ou art d'inventer " aisée à connaître et à employer aux esprits les plus gros- " siers, j'ai trouvé un moyen de la revêtir pour ainsi dire " d'un corps palpable et agréable tout à la fois. Et ce moyen " est le projet que j'ai d'une langue ou écriture nouvelle qui " se pourrait apprendre en une semaine ou deux, qu'on ne sau- " rait quasi oublier et qu'on pourrait même retrouver l'ayant " oubliée, qui aurait bientôt cours dans le grand monde lors- " qu'elle serait connue et qu'elle aurait eu l'approbation de " quelques grands personnages; mais qui outre l'usage du com- " merce et la communication des peuples divers (ce qui la pour- " rait même rendre plausible au vulgaire) aurait des avantages " incomparablement plus grands: car elle donnerait moyen de " raisonner sur les matières capables de raisonnement par une " espèce de calcul infaillible pourvu qu'on y apportât la mê- " me exactitude qu'à chiffrer, et les erreurs ne seraient que " des erreurs de calcul. Il y aurait même des preuves sembla- " bles à celles du novenaire dont on se sert dans l'arithmé- " tique; il n'y aurait point de contestation entre ceux qui " voudraient y compromettre; et non seulement on trouverait " là dedans des voies infaillibles pour arriver à la solution " des problèmes qui se peuvent résoudre par la seule force du " raisonnement, mais lors même qu'il s'agit d'une question de " fait, et ce qu'il reste encore des expériences à faire qui " ne sont pas toujours dans le pouvoir des hommes, ce calcul " serait suffisant pour nous conduire, en attendant, le mieux " qu'il est possible de faire suivant la raison sur les con- " naissances déjà données. Car par là nous pourrions estimer " les degrés de probabilité, ce qui est une chose également " importante et négligée dans la morale et dans les affaires; " nous pourrions même trouver quelles recherches ou expérien- " ces restent encore à faire afin de nous éclaircir entière- " ment autant que cela se peut par la seule force de la rai- " son; et non seulement nous pourrions projeter experimenta " crucis, comme le chancelier Bacon les appelle, pour mettre " la nature à la question, mais nous pourrions encore par là " dresser des articles ou interrogatoires pour examiner les " hommes et pour tirer d'eux la vérité sans qu'ils s'en aper- " çoivent. En un mot, le dictionnaire de cette langue serait " comme un inventaire dans ce grand magasin confus d'une infi- " nité de belles sciences qui sont déjà acquises, mais dont " les hommes ne savent pas se servir, ni en tirer toutes les " conséquences qui sont déjà en leur pouvoir. Car ils m'ont " avoué en Angleterre que le grand nombre d'expériences qu'ils " ont amassées ne leur donne pas moins de peine aujourd'hui " que le défaut d'expérience en donnait aux anciens. Ce qui " arrive faute de cette méthode que je viens d'expliquer. Et " je tiens pour assuré que les hommes ont déjà en main des " moyens de se garantir de quantité de maux qui leur arrivent, " s'ils savaient en profiter."

(extrait d'une lettre de LEIBNIZ à JEAN-FREDERIC - Février 1679)

PRESENTATION DES COURS

A. LEROY

Les différents exposés de l'enseignement n'ont pas la prétention de vous démontrer que les problèmes de la documentation sont résolus, que "la" méthode a été trouvée, que "la" machine a été construite. Au contraire, vous n'entendrez que suggérer des méthodes et des moyens pour traiter la masse des informations disponibles et des procédés qui permettront de se rapprocher d'une solution optimale.

Si le chemin qu'il faut utiliser pour arriver au but n'est pas encore complètement déterminé, le but lui-même par contre, doit l'être. Le nôtre est de concevoir une machine qui nous renseigne aussi exactement que possible sur n'importe quelle question scientifique, et ce dans la langue du demandeur.

Nous devons évidemment lui fournir tous les renseignements scientifiques que nous pouvons nous procurer et les éléments de la question que nous désirons poser. En fait l'aliment de la machine sera essentiellement l'ensemble des publications scientifiques.

Elle devra lire ces publications, les analyser, mettre les résultats de cette analyse en mémoire, trouver, lors d'une demande, les éléments intéressant celle-ci, et fournir ces éléments dans la langue voulue. Nous sommes donc conduits à étudier les possibilités d'automatisation de chacune de ces phases, considérées comme parties d'un ensemble.

Pour cela nous devons bien entendu tenir compte des problèmes qui se sont déjà posés toutes les fois qu'on a voulu réaliser une documentation efficace, même lorsque l'on n'envisageait pas son automatisation, puisque les buts s'y retrouvent en partie. M. PIETSCH traitera ces problèmes.

Sans doute pouvons-nous déjà dire que nos ennuis proviennent de la limitation des possibilités humaines. En effet le documentaliste n'est pas assez rapide : ni au moment de la lecture des textes, ni au moment de l'exploration des "schémas" qu'il peut avoir en mémoire et qui pourraient convenir pour exprimer rationnellement ce qu'il lit. D'autre part le nombre de ces schémas est forcément limité ce qui signifie qu'un documentaliste ne peut être spécialiste que d'un petit nombre de domaines scientifiques. Les mêmes limitations l'empêchent d'opérer une sélection suffisamment rapide et précise.

Nous devons donc tenir compte des moyens qui ont déjà été mis en oeuvre pour combattre ces limitations. Moyens modestes d'abord puis de plus en plus puissants, si puissants qu'on en arrive à la conclusion, étant donné une bonne partie des résultats obtenus jusqu'à ce jour, qu'ils sont mal adaptés aux problèmes qu'on leur demande de résoudre ou que les problèmes ont été rarement bien formulés.

M. de GROLIER parlera des moyens en question.

M. BRAFFORT traitera plus spécifiquement des calculateurs. On a pu dire que les calculateurs n'étaient pas bien adaptés aux problèmes de la documentation. C'est sans doute vrai par le fait qu'ils n'ont pas été conçus dans ce but. Mais on ne peut nier que certains de leur éléments de base se retrouveront dans la future machine. On peut aussi

prédire que la technique des calculateurs évoluera dans un sens qui sera à la fois favorable au traitement des informations numériques et de celles qui nous intéressent. Enfin, ces calculateurs permettent d'effectuer des expériences, qualité extrêmement importante, mais seulement si on sait tirer profit de ces expériences. Pour cela, il existe des outils qui feront l'objet des exposés de MM. SCHUTZENBERGER et GAZZANO. On conçoit facilement en effet que la théorie de l'information puisse apporter une grande aide, par exemple pour assurer le meilleur "codage", et que la recherche opérationnelle puisse permettre elle aussi d'effectuer les meilleurs choix, par exemple en ce qui concerne les "stratégies" documentaires.

Ces questions sont indissociables de la statistique, qui fera l'objet des premiers travaux pratiques.

Autre domaine d'une importance évidente : la linguistique et les sciences connexes - Comment l'homme exprime-t-il sa pensée ? Qu'est-ce qu'un langage ? M. CECCATO, qui a étudié pendant longtemps ces problèmes, a bien voulu nous faire un panorama de ses vues en la matière. MM. BRAFFORT et BELEVITCH traiteront également de linguistique, respectivement d'un point de vue "structural" et d'un point de vue statistiques. Enfin les travaux pratiques de linguistique porteront sur la détermination systématique des rapports entre les mots et se présenteront comme une introduction à l'exposé de M. LECERF, de même que les travaux pratiques précédents constituaient une introduction à l'exposé de M. BELEVITCH.

Nous connaissons maintenant les problèmes et nous avons à notre disposition un certain nombre de moyens. Essayons donc de concevoir un système documentaire efficace en gardant présente à l'esprit la perspective de l'automatisation complète - Cette dernière condition fait que nous n'hésiterons pas à utiliser des méthodes qui seraient difficilement applicables si l'on ne disposait pas des moyens évoqués par MM. de GROLIER et BRAFFORT - Mais cette difficulté d'application ne provient que du fait que les méthodes envisagées demandent une trop grande vitesse d'opération lorsque les collections deviennent très importantes. M. GARDIN montrera que les principes de base appliqués au domaine des sciences humaines donnent entière satisfaction.

Au cours des travaux pratiques qui suivront on s'efforcera d'appliquer les principes énoncés dans ces exposés pour constituer le début d'un grand système documentaire efficace.

Avec les séances suivantes nous entrons réellement dans le domaine de l'automatisation - Sur le plan des réalisations pratiques d'abord : M. LECERF nous montrera comment une machine peut analyser grammaticalement des textes, de chimie pour la circonstance, et l'expérience qui suivra nous permettra de voir la réalisation pratique sur machine IBM.

Dans le même ordre d'idée, les exposés de Mme POYEN et de M. TE NUYL porteront sur les calculateurs utilisés comme outils de sélection des informations non-numériques.

La dernière après-midi sera consacrée aux perspectives. Seront ainsi évoqués les problèmes de traduction automatique par M. IUNG,

d'analyse et de sélection automatiques par moi-même. Enfin M. BRAFFORT s'efforcera de tirer la leçon de ces journées en soulignant les perspectives offertes par la convergence des méthodes développées pour le traitement des informations non-numériques et pour le calcul numérique.

HISTORIQUE DES SYSTEMES DOCUMENTAIRES

E. DE GROLIER +

I - Qu'est-ce qu'un "système documentaire" ?

Je dois vous parler, selon le programme de ce séminaire, sur l'"historique des systèmes documentaires". A vrai dire, il faudrait d'abord se poser la question : qu'est-ce que l'on doit entendre par "systèmes documentaires" ? Au sens étroit, je pense qu'il s'agit des systèmes utilisant des machines, mais sans doute faut-il replacer ceux-ci dans un contexte beaucoup plus large; vous voudrez donc bien m'excuser si je prends les choses de très loin. Car, au sens large, il semble que l'on puisse parler de système documentaire dès qu'il y a transmission d'informations sous une forme plus ou moins systématique, c'est d'ailleurs une simple paraphrase. Or ces transmissions d'informations telles que nous les connaissons aujourd'hui sont le produit d'une très longue évolution, dont les différentes étapes ont laissé des traces dans l'organisation actuelle.

II - Première période : la parole

On pourrait définir la première période comme celle de la parole; d'ailleurs faut-il préciser qu'elle part d'une époque où l'homme n'existait pas encore, si l'on convient de couvrir par le terme parole le langage des abeilles - lequel, comme l'ont montré les études de von Frisch, est sans doute plus satisfaisant du point de vue documentaire que le nôtre sur un point, car ses signaux sont non équivoques, ce qui n'est pas toujours le cas des nôtres. Néanmoins on doit reconnaître qu'il s'agit de systèmes documentaires très réduits, puisque la seule information qu'ils puissent transmettre se rapporte à la présence d'un certain pollen en un certain lieu.... Si nous voulons trouver des systèmes documentaires plus différenciés, il nous faut tout de même passer à l'homme et à ce que Pavlov a appelé le "deuxième système de signalisation", c'est-à-dire le langage proprement dit. La communication créée par la parole est directe, elle est constituée par des éléments isolés d'information qui sont transmis directement entre les locuteurs, sans délai ni intermédiaire. L'invention du langage articulé reste à la base de

+ Centre Francais d'Echanges et de Documentation Techniques (Milan)

tout système documentaire actuel, quel qu'il soit : tous les systèmes que nous connaissons sont fondés sur le langage, et c'est ce qui fait que les recherches linguistiques présentent une telle importance pour la documentation. Mais quelle machine le langage utilise-t-il ? Eh bien, le langage est une "technique du corps" qui n'a donc besoin de nul "artifact", de nulle chose qui soit extérieure à l'homme lui-même. Si maintenant nous passons de la machine au code, nous voyons qu'il s'agit de codes qui furent sans doute très simples à l'origine, mais qui ont évolué vers une grande complication et un extrême raffinement. La pensée se réalise à travers le langage non sans difficultés - celui-ci évolue avec elle vers l'expression d'idées abstraites; il s'opère une certaine normalisation des éléments du langage, qui tendent à être non-équivoques dans le langage scientifique. Mais ce langage scientifique n'est jamais complètement séparé du langage ordinaire, et il y entre nécessairement de ce fait des éléments non normalisés (et sans doute même non normalisables) : éléments affectifs ou esthétiques dont nous trouvons des exemples dans n'importe quel article scientifique, chaque fois que l'auteur essaie d'éviter des répétitions, d'obtenir une certaine harmonie dans son langage, toutes choses qui seraient parfaitement inutiles du simple point de vue logique, mais qui facilitent sans doute cependant la communication en la rendant plus "agréable" - et qui causent par ailleurs bien des difficultés quand il s'agit de traduire en langage-machine. Bien entendu, dans le langage il se trouve aussi des restes de catégories périmées, ayant perdu depuis longtemps leur valeur originelle : ainsi les genres grammaticaux. Je voudrais ici vous faire remarquer que l'usage de la parole reste extrêmement important en documentation : les résultats des enquêtes qui ont été faites sur des chercheurs ont montré que les communications orales représentaient souvent plus de 50 % des actes d'information. On ne peut donc dire que la parole, l'information orale, ait perdu son importance dans l'information scientifique actuelle, et cela d'autant moins d'ailleurs que l'on a développé des moyens techniques plus puissants pour la transmission à distance de cette parole : téléphone, radio, enregistrement phonographique, magnétophone, etc. Si vous le permettez, je vous citerai ici une anecdote. A la Conférence de San José, en mai 1958, l'un des chercheurs d'I.B.M. avait fait une recherche documentaire très poussée pour connaître les caractéristiques que l'on pouvait escompter dans un proche avenir pour les mémoires. Il avait fait appel entre autres à un "système documentaire" qui était à l'époque l'un des plus perfectionnés quant à l'automatisme, et qui utilisait la RAMAC 650; il avait ainsi obtenu un certain nombre de références, mais il nous dit : "je dois quand même vous avouer que j'ai employé aussi la vieille méthode bien connue qui consiste à consulter directement celui "qui sait"; j'ai donc interrogé individuellement ceux de mes collègues qui étaient spécialistes de ces questions et, finalement, ce sont leurs renseignements qui m'ont été les plus précieux". Eh bien, ce genre de situation est assez fréquent, vous et moi pouvons aussi citer de nombreuses expériences analogues que nous avons pu faire. Les renseignements que l'on recueille de la bouche de "celui qui sait", c'est encore aujourd'hui quelque chose de très important, souvent le plus important!

III - Deuxième période : L'écrit

Dans ce survol rapide de l'histoire de la documentation, passons à la deuxième période qui est celle de l'écrit, et qui voit naître des communications indirectes. Celles-ci qui viennent s'ajouter aux communications directes, d'homme à homme, sont rendues possibles d'abord par la naissance de l'image (déjà chez les magdaléniens) puis par l'écriture. Avec elle commence la civilisation proprement dite, qui est une "civilisation écrite", caractérisée par la présence du livre (au sens large, depuis les inscriptions jusqu'aux formes actuelles du périodique ou des rapports techniques). A ce moment, nous voyons apparaître dans la communication un "artifact" (terme anglais bien commode, qui n'a guère d'équivalent français, à part le "bidule" devant lequel certains pourraient s'offusquer peut-être). Cet "artifact", c'est le livre et il convient sans doute de marquer ici que ce n'est pas encore une chose périmée. Nous sommes tous, c'est entendu, pour la documentation automatique, mais enfin nous utilisons par ailleurs encore des livres, périodiques, etc... Nous les utilisons parce qu'ils représentent un système d'information qui a fait ses preuves, qui a permis une certaine sécurité pour la transmission des idées, malgré la fuite du temps, qui a rendu possible ce fait que, selon les paroles de Pascal (dans le fragment d'un Traité du vide) "Toute la suite des hommes, pendant le cours de tant de siècles, doit être considérée comme un même homme qui subsiste toujours et qui apprend continuellement".

Le livre, depuis l'époque reculée de l'invention de l'écriture, a évolué sans doute, mais n'a pas été fondamentalement transformé et nos modernes périodiques ou rapports techniques sont le produit d'une évolution linéaire continue, depuis les tablettes mésopotamiennes ou les hiéroglyphes égyptiens. Certes, les deux derniers siècles ont vu apparaître de nouvelles méthodes de reproduction : photographie, puis microphotographie, aujourd'hui xérographie, celle-ci considérée par certains comme devant provoquer une sorte de "révolution" dans la transmission des informations. Je ne suis pas tout à fait d'accord sur ce point et je pense que tous ces nouveaux procédés ne changent encore rien au système de base lui-même; ce dernier n'est pas non plus transformé par l'intervention de moyens de reproduction à distance : télégraphe, téléscripteur, télévision, etc., quels que soient leurs effets, si considérables qu'ils soient. Quant au code, nous trouvons là un deuxième système de code superposé à celui du langage parlé et destiné à en être l'image : l'écriture; malheureusement, il s'agit d'une image très infidèle et d'autant plus infidèle d'ailleurs, semble-t-il, qu'il s'agit de langues de plus haute civilisation à ce point que le statisticien Herdan a écrit que la complication orthographique était "le prix qu'il fallait payer" pour développer une littérature importante à partir d'une langue donnée. Je crois qu'il exagérait un peu, mais cette divergence entre les deux systèmes de signes, oral et écrit, cause des difficultés que l'on retrouvera, très gênantes, si l'on veut passer à une documentation entièrement automatique donnant des écrits à la sortie à partir de la parole à l'entrée (ou l'inverse).

Comme l'a bien montré à Cleveland le rapport de Booth, les difficultés principales qui s'opposent à une telle réalisation sont les divergences entre l'expression orale et sa traduction écrite dans des langues comme l'anglais (et même le français).

A peu près en même temps qu'on a créé l'artifact n° 1 - le livre - se sont fondés des dépôts de ces documents, archives et bibliothèques, plus récemment évolués en centres de documentation ou services d'information. Il sont tous caractérisés par le fait qu'entre ce qui était le locuteur d'une part et l'auditeur de l'autre - devenus maintenant auteur et lecteur, entre le producteur d'un document et le destinataire de ce document, intervient désormais un troisième personnage : cet intermédiaire est le bibliothécaire, l'archiviste, le documentaliste. Ce personnage, si vous me permettez cette image, comme le Janus bifrons, a deux faces, l'une tournée vers les productions des auteurs, les livres, les périodiques, etc., et l'autre tournée vers leurs destinataires, c'est-à-dire les lecteurs. Historiquement, sa première face, celle tournée vers les livres, a eu plus d'importance au début que sa deuxième face, celle tournée vers le lecteur. Le bibliothécaire, ou l'archiviste, était autrefois un homme dont la tâche essentielle était de réunir des documents et qui devait ensuite, subsidiairement, les mettre à la disposition des lecteurs. Le bibliothécaire moderne, mais plus encore le documentaliste ou "l'information officer" (terme à nouveau difficile à traduire en français), est davantage tourné vers les lecteurs et doit se consacrer à leur service. A côté de ce personnage intermédiaire, et servant d'instruments à sa disposition, sont venus se placer des "artifacts n° 2" : tout ce que l'on appelle les "publications secondaires" : catalogues, bibliographies, recueils analytiques, etc... Je vous ferai observer que ces artifacts n° 2, destinés à faciliter au lecteur, au récepteur des documents, la connaissance qu'il doit prendre de ceux-ci, sont restés fondamentalement les mêmes en définitive depuis quatre ou cinq cents ans, et même peut-être davantage. Ils n'ont fait qu'évoluer linéairement, devenant de plus en plus gros, de plus en plus complexes, mais leur principe n'a pas changé. Il en est de même d'ailleurs des dépôts de documents, dont le principe est en somme le même qu'aux temps des bibliothèques de Ninive ou d'Alexandrie, et qui sont seulement devenus quantitativement plus importants. Je vous ferai remarquer en outre qu'il y a eu une évolution cyclique en ce sens qu'un premier sommet, atteint avec la bibliothèque d'Alexandrie et ses quelque cent mille volumes, a été suivi d'une chute verticale puisqu'au moyen âge nous trouvons des collections dont le nombre de documents s'exprime par dizaines ou centaines; la reprise a d'abord été lente : au XVIIè siècle la Bibliothèque Mazarine en est encore aux 40.000 volumes et peut cependant être considérée comme une merveille du monde; le cap des cent mille volumes n'a été franchi en France à la Bibliothèque Royale qu'à la fin du XVIIIè siècle. Puis il y a eu croissance exponentielle : le principe de l'accélération de l'histoire dont a parlé Daniel Halévy se vérifie ici aussi, et maintenant les chiffres qu'on nous donne font état (comme ceux du Dr Pietsch pour le nombre d'ingénieurs) d'un doublement en dix ans pour les pays qui évoluent le plus rapidement. Dans des pays de vieille civilisation, comme la France, le rythme de croissance est plus lent (L.J. Van der Wolk

indique dans son article du Bulletin des Bibliothèques de France, nov. 1959, P. 481, un doublement en cent ans pour les bibliothèques universitaires françaises). Par ailleurs, un autre phénomène est intervenu : la spécialisation, commencée dès le XVIè siècle pour les bibliographies spécialisées, continuée vers la fin du XVIIIè siècle avec les bibliothèques spécialisées et à la fin du XIXè avec les centres de documentation. Dans l'ensemble, tout cela n'a que peu modifié jusqu'à présent l'aspect général du travail de bibliographie ou du bibliothécaire, et les méthodes que l'on utilise aujourd'hui encore en général dans les centres de documentation sont très semblables après tout, sauf modifications de détail, à celles qui étaient utilisées par disons la Bibliothèque Royale au XVIIè siècle. L'invention de la fiche par exemple, par l'Abbé Rozier, n'avait pas modifié la structure générale du deuxième genre d'artifacts : les bibliographies ou catalogues, mais seulement sa forme extérieure. Il faut bien constater que, pour le moment, les techniques bibliographiques et bibliothéconomiques traditionnelles demeurent encore valables en ce sens que, comme le montrent les enquêtes sur l'information scientifique, ce sont des méthodes qui sont toujours largement utilisées par les usagers.

La documentation automatique va sans doute changer tout cela, mais nous sommes pour l'heure devant une situation qui est que l'on utilise pour 50% à peu près l'information orale et pour les autres 50% les formules traditionnelles développées durant la période de la "civilisation écrite" : bibliothèques, bibliographies etc... Quant aux codes, durant cette période examinée ici, ils avaient été l'objet de deux développements divergents : d'une part, avec des classifications systématiques du type à "hiérarchie forte", selon la terminologie de Mooers, c'est-à-dire avec une structure arborescente faisant usage de la seule relation d'inclusion (telle que vous la connaissez bien dans la classification des sciences naturelles, par exemple quand il s'agit de retrouver le nom d'une plante par dichotomies successives); d'autre part, avec des index alphabétiques, dont je n'ai pas besoin de vous parler longuement, qui s'étaient perfectionnés lentement au cours du temps, ceux des Chemical Abstracts représentant probablement l'état le plus évolué de ce développement à l'heure actuelle. Pourtant tout ceci ne constitue encore qu'un accroissement quantitatif de choses qui existaient déjà, disons au temps à peu près des premières bibliographies nationales au XVIIè siècle.

IV - Troisième période : les machines, la documentation automatique

Mais tout laisse penser que nous voici maintenant au seuil d'une troisième période, et c'est là sans doute que mon exposé peut commencer à vous intéresser un peu plus.

En effet, tout fait penser que nous approchons d'une sorte de révolution dans les méthodes d'information, et que nous vivons à cet

égard une époque où les changements quantitatifs graduels ont amené finalement un brusque changement qualitatif. A vrai dire, cela avait été prévu il y a déjà bien des années; j'étais à Bruxelles en 1931 et à cette époque déjà Otlet prévoyait la "documentation automatique", mais on avait alors aucun moyen pratique de la réaliser. En 1937, à l'occasion du Congrès mondial de la documentation à Paris, Pierre Bourgeois avait fait une communication dans laquelle il exposait d'une mani re plus détaillée ce que pourrait être la mécanisation du travail documentaire. Les réalisations effectives datent seulement en fait d'une quinzaine d'années. La cause fondamentale doit en être rapportée à l'accroissement exponentiel de la quantité de documentation que l'usager moyen doit "digérer" et qui, même pour le spécialiste, est devenue tellement considérable que cette "digestion" dépasse ses forces - et le temps dont il dispose. Le destinataire des documents ne peut plus faire lui-même le travail d'extraction des informations contenues dans les documents - informations qui cependant, je dois attirer votre attention sur ce point, sont la seule chose qui l'intéresse, car le document en lui-même au fond ne lui importe guère : ce qui compte pour lui, ce sont les données qu'apporte ce document et même seulement les données nouvelles. Or, jusqu'à présent, l'usager doit effectuer finalement lui-même le travail de criblage des documents pour dépister ceux qui peuvent lui apporter des informations inédites. Même s'il trouve pour cela une aide dans la systématisation des documents par des classifications, même avec le secours des publications secondaires bibliographiques signalétiques ou analytiques, "annual reviews, "Handbücher" genre Gmelin ou Beilstein, ou tout ce que vous voudrez, ce travail est devenu quelque chose qui dépasse ses forces et qui dépasse même les forces de l'intermédiaire (le bibliothécaire, le documentaliste) que nous avons vu apparaître tout à l'heure. Alors que se passe-t-il ? Eh bien, on voit que les techniques de la documentation écrite perdent finalement de leur intérêt, car elles n'arrivent plus à maîtriser le flot toujours croissant des documents. Certes, on utilise toujours les grands index, du type de celui des Chemical Abstracts - mais ceux-ci deviennent si lourds, si volumineux, qu'on ne sait pas très bien si l'on pourra encore éditer un autre index décennal des Chemical Abstracts. Il est naturel dans une telle situation que l'on cherche un secours dans la mécanisation. Sans doute n'est-il pas inutile de vous donner ici un aperçu "à vol d'oiseau" de l'évolution des machines utilisables en documentation. Les premières machines que l'on a pensé pouvoir utiliser pour la documentation, ce sont les machines à statistiques Hollerith (ou Bull ou Gammas) nées dans les années 1880, mais dont les premières applications à la documentation ne remontent guère au-delà de 1936 : donc une cinquantaine d'années au plus tard.

Le deuxième système mécanographique que l'on a cherché à utiliser en documentation a été celui connu actuellement en anglais sous le nom de "Peek-a-boo" (en allemand : Sichtlochkarten) découvert en 1915 par Taylor, perfectionné en 1920 par Soper, mais dont les premières applications documentaires quelque peu étendues remontent à 1939-1940 environ, c'est-à-dire quelque vingt-cinq ans après leur invention.

Les cartes à préperforations marginales ont été inventées par l'anglais Alfred Perkins vers 1919 et leurs applications documentaires se sont développées quelques décennies plus tard; elles ne sont guère applicables qu'à de petites collections de documents.

Le quatrième système est dénommé en anglais "film scanning", que je ne sais trop comment traduire: disons sélection à l'aide de films. L'idée en a été lancée pour un brevet Goldberg déposé en 1928 et accordé en 1931, et par des recherches de Watson-Davis et de Draeger en Amérique vers 1935, la première réalisation ayant été le Rapid Selector de Bush vers 1938-1939, dont un modèle transformé est actuellement à l'étude au National Bureau of Standards. Le même système de sélection utilisant un film continu se retrouve avec le Flip de la Benson-Lehner Corp. présenté en 1958, et Marcel Locquin a présenté très brièvement à l'Académie des Sciences, il y a quelques jours, une note sur un système qui semble - pour autant qu'on puisse en juger - en être un proche parent.

Le Dr Pietsch vous a parlé du Filmorex Samain et de la Minicard Kodak, que l'on peut considérer comme issus d'une sorte de "croisement" entre le système des machines à cartes perforées Hollerith et celui du "film scanning". Les défauts de ces procédés, c'est leur vitesse de défilement assez réduite (respectivement 600 à 700 microfiches/minute et 2.000 minicards/minute) : pour des collections de documents de quelque étendue, l'opération de sélection devient trop longue.

Nous trouvons un système qui est lui aussi issu d'une sorte de croisement, cette fois entre le "peek-a-boo" et le "film scanning", avec le projet Cordonnier de "cosélectionneuse" actuellement en cours de réalisation en France. Comme autre exemple de système à mémoire optique (il y en a, à ma connaissance, un seul exemplaire existant), nous avons la mémoire à disques créée par Gilbert King à l'International Telemeter Corporation, vers 1955, développée plus récemment par I.B.M. et qui va servir à la traduction mécanique avec le dictionnaire automatique enregistré sur cette mémoire pour le Professeur Reifler à l'Université de Washington. Pour ce qui est des calculatrices, je laisserai Braffort vous parler de leur évolution et de leur emploi en documentation. Finalement, je rejoindrai ce que vous disait le Professeur Pietsch : jusqu'à présent toutes ces machines ont été conçues pour d'autres besoins que la documentation (calcul, traitement d'informations dans les administrations ou les affaires); peut-être la première formule de mémoire vraiment adaptée aux besoins des documentalistes est-elle la Magnacard (dont le Professeur Pietsch vous a aussi parlé), grâce à sa grande capacité et à sa grande rapidité de défilement. Une loi (très empirique) a été dégagée par Calvin Mooers pour estimer le rendement des systèmes documentaires : c'est le critère du nombre de documents sur lesquels on peut opérer une sélection en une demi-heure. Si l'on applique cette règle (tout à fait empirique, encore une fois) au Filmorex du docteur Samain, par exemple, on trouve une limite d'environ 18 à 20.000 documents. Naturellement, on peut "s'en tirer" pour des collections plus importantes en pratiquant une préselection des microfiches, mais celle-ci n'est pas exempte d'inconvénients du point de vue économique.

Parlons maintenant un peu des codes. Historiquement, on a d'abord utilisé pour la documentation mécanique des codes qui n'étaient pas faits pour elle. Vers 1935-36, par exemple, quand on a commencé à travailler avec des machines Hollerith, I.B.M. ou autres, on a employé parfois la CDU (classification décimale universelle). Or, la CDU n'est pas adaptée aux méthodes mécanisées de recherche d'informations, parce que c'est à la base une classification "à hiérarchie forte" fondée sur la relation d'inclusion, et qu'une telle classification est très peu efficiente pour les machines. On s'est donc aperçu, comme le dit Mooers (dans une conférence fort intéressante qu'il a faite l'année dernière et que j'ai largement utilisée ici : "The next twenty years in information retrieval") "new mechanisms deserve new methods", les nouvelles machines demandent des méthodes nouvelles. Ces nouvelles méthodes, on peut les décrire, comme le fait Mooers, en disant qu'il s'agit de "putting together independent ideas expressing terms and selecting upon their correlative occurences", autrement dit, en français, de réaliser des combinaisons de termes indépendants simples, afin d'exprimer des concepts complexes, et d'opérer ensuite la sélection en fonction de la présence simultanée de plusieurs de ces termes simples. Ceci implique que l'on passe des classifications à "hiérarchie forte" basées sur la relation d'inclusion et sur le schéma arborescent, à des classifications à "hiérarchie faible", c'est-à-dire en forme de treillis et faisant usage d'autres relations que la relation d'inclusion. Cette évolution des classifications, à vrai dire on peut la rattacher à des origines assez anciennes. Probablement faut-il en voir les premières manifestations dans les projets de langues universelles de Leibniz et de Wilkins au XVIIè siècle, mais il faut attendre 1895 pour voir ces idées pénétrer dans le domaine de la documentation proprement dite avec le belge Paul Otlet : c'est lui qui introduisit le premier dans une classification documentaire une autre relation que la relation d'inclusion avec ce signe que les usagers de la CDU connaissent bien : le deux points, le signe de "relation générale", qui a été le premier exemple dans une classification documentaire d'une méthode pour relier des concepts simples en un concept complexe.Redécouverte par Ranganathan, en 1933, redécouverte une deuxième fois par Gérard Cordonnier en 1943, cette méthode a mené aux classifications dites "à facettes" (faceted classifications) que préconisent en Angleterre nos amis du Classification Research Group et, en France, par exemple, à l'analyse codée de Robert Pagès. Reste à savoir si ces classifications sont vraiment adaptées à la documentation automatique; personnellement, je ne le crois pas : elles ont été faites pour des documentations "manuelles", utilisant des fichiers normaux, des fichiers du type de l'abbé Rozier, et non pas des machines documentaires (à l'exception de celle de Pagès, qui travaille avec des fiches peek-a-boo).

Autre méthode de codification, celle dite des "descriptors" - encore un terme anglo-saxon difficile à traduire, peut-être "descripteur", ou disons "mot-clé" comme Braffort. Lancée par Mooers pour être utilisée avec un système particulier de machines d'ailleurs fort simple, elle a été conjuguee par lui avec l'utilisation de codes superposés, faisant usage de nombres tirés au hasard. Il n'y a plus classification proprement

dite, celle-ci est remplacée par des mots normalisés correspondant à des concepts simples répartis entre un certain nombre de catégories qui peuvent d'ailleurs se recouvrir partiellement, qui ne sont pas mutuellement exclusives, cette catégorisation servant seulement pour faciliter les contacts avec l'usager. Dans le système primitif de Mooers, il n'y a pas d'expression des relations. Un développement de cette méthode fait usage des mêmes descripteurs, des mêmes mots-clés, mais tend à exprimer les rapports entre ceux-ci et à passer, comme l'écrivent Leroy et Braffort, des mots-clés aux phrases-clés. On peut trouver une tendance de ce genre dans les travaux de Hans Selye au Canada vers 1956, puis chez Jean-Claude Gardin, en France, et l'on aboutit aux recherches de l'équipe qui nous accueille aujourd'hui. On cherche à exprimer les relations par différents moyens : chez Selye, à l'aide de signes spéciaux (qui ressemblent à des prépositions, si l'on veut une comparaison avec le langage), chez Gardin, avec des désinences qui s'apparentent aux flexions greco-latines; chez Braffort enfin, avec des diagrammes, procédé nouveau.

Une troisième méthode de codification utilise les mots du langage courant. Historiquement, elle est un peu postérieure à la méthode des "descriptors", elle est née avec les "Uniterms" de Mortimer Taube qui, à l'origine, n'étaient guère autre chose que des mots du langage, destinés à représenter une idée simple, mais non standardisés. Depuis, les "uniterms" tendent à se rapprocher des descripteurs.

Une quatrième méthode est celle du "semantic code" de Perry et de ses collaborateurs à Cleveland, qui est une espèce de croisement entre les méthodes n° 2 (c'est-à-dire des descriptors) et n° 1 (c'est-à-dire des classifications systématiques).

Il y aurait un certain nombre de remarques à faire sur cette codification, quelque peu hybride; si l'on voulait lui trouver des "parents" parmi les langues naturelles, il faudrait se diriger d'une part vers les langues polysynthétiques (du type esquimau) et d'autre part vers les langues chamito-sémitiques (utilisation de racines consonantiques et de variations vocaliques internes pour exprimer certaines relations). Quoi qu'il en soit, ce code va servir ces jours-ci à la première expérience sur une large échelle de documentation automatique, puisque l'American Society of Metals vient d'annoncer la mise en route d'un service de documentation automatique faisant usage du code Perry avec la machine General Electric 250 dont on vous a parlé ce matin. Si nous voulions maintenant caractériser très brièvement toutes ces méthodes, nous pourrions sans doute dire que, pour l'instant, il s'agit encore de recherches. J'aurais pu d'ailleurs vous parler encore de bien d'autres tentatives, comme celle de Simon Newman au U.S. Patent Office pour constituer ce qu'il dénomme un "Ruly English", sorte de langage normalisé utilisant des relations elles-mêmes normalisées, qui aboutit assez curieusement à quelque chose d'assez analogue au langage imaginé par Orwell dans son roman "1984". J'ai essayé de décrire avec quelque détail toutes ces recherches dans un assez volumineux rapport pour l'UNESCO, sous le titre plutôt rébarbatif de "Etude sur les catégories générales applicables aux classifications et codifications documentaires".

V - Jusqu'où peut aller l'automatisation ?

Peut-être pourrions-nous enfin tenter très rapidement un essai de "propective", selon le terme récemment lancé par Gaston Berger : quelles sont les perspectives, compte tenu de ce long passé (fort long, en effet, puisqu'il remonte à l'époque - que je ne saurais préciser - à laquelle les abeilles ont commencé à danser pour faire savoir qu'il y avait des fleurs mellifères dans telle direction et à telle distance, cela fait peut-être quelques centaines de millions d'années) et de ce mouvement impétueux, le renouvellement que nous voyons se développer à une allure explosive pendant les dernières années ? Où allons-nous ? et jusqu'où pouvons-nous aller ? Je crois que les procédés classiques (j'entends par là les procédés nés, développés et mis au point au cours des périodes n° 1 et n° 2 de notre historique, la période de la parole et la période de l'écrit) ne sont pas destinés à mourir. Ils continueront d'être utilisés, de servir, probablement sous des formes quelque peu différentes, mais ils seront complétés (je dirais biencomplémentés si j'osais ce barbarisme) par des méthodes nouvelles, automatiques. Cette automatisation peut d'ailleurs comporter plusieurs stades. Le premier stade concerne l'automatisation de la production de choses classiques, par exemple de la production d'index, tels que nous les connaissons. Le Professeur Pietsch nous a fait ce matin une démonstration de l'économie que peut apporter l'usage des machines pour la production d'index classiques. Il y en a un autre exemple très remarquable et fort récent, puisqu'il date seulement du 1er janvier 1960, dû à la National Library of Medecine américaine qui a mécanisé la production de sa bibliographie de la médecine (Index Medicus) par des méthodes un peu différentes de celles du Professeur Pietsch mais qui s'en rapprochent cependant par plusieurs côtés (utilisation du flexowriter, de machines à statistiques type Hollerith). On peut aussi noter que les Chemical Abstracts étudient en ce moment très sérieusement les possibilités d'application de diverses machines à la production de leur bibliographie et de ses index. Je crois que c'est là un champ d'application très intéressant de la documentation "automatique" (il faudrait plutôt dire documentation semi-automatique, ou mieux production automatisée d'instruments bibliographiques traditionnels, conventionnels).

A un stade un peu plus "avancé" se trouvent les méthodes qui permettent de réserver à l'homme la seule tâche de faire le choix des descripteurs, des mots-clés ou des phrases-clés, et ensuite de faire faire par des machines les opérations qui suivent ce choix : dans cet ordre d'idées nous avons par exemple le travail de Frome qui est le dernier numéro de la série des U.S. Patent Office Research and Development Reports, dont le professeur Pietsch vous parlait ce matin, et qui est un exemple.

Une étape plus loin, nous pouvons imaginer une sélection purement mécanique à partir des textes originaux eux-mêmes. Il y a encore, à l'heure actuelle, très peu de réalisations effectives dans ce domaine. On peut citer par exemple les travaux poursuivis chez I.B.M. pour la confection automatique d'index à partir des titres des rapports. (Voir le rapport de H.P. Luhn, Keyword - in - context index for technical literature, Yorktown Heights, 1959, RC 127). Malheureusement l'on se heurte ici à l'absence de normalisation, l'absence de règles, à l'étape de la production des documents. Les titres (je ne vous l'apprends pas) sont très souvent imprécis, insuffisants, et la machine ne peut travailler que sur la matière première qu'on lui fournit : si cette matière première est médiocre, le résultat sera médiocre. Néanmoins, ces index obtenus selon la méthode dite "KWIC" ne sont pas du tout sans intérêt. Il y a une tentative beaucoup plus ambitieuse, qui est celle de Luhn à I.B.M., pour la production automatique non plus d'index mais bien d'analyses. J'ai fait allusion à cette expérience à la Conférence Internationale de Washington de novembre 1958, en disant qu'elle ouvrait une ère nouvelle, mais elle ouvre cette ère nouvelle plus par les perspectives qu'elle trace pour l'avenir que par ses résultats concrets actuels. Pourquoi ? Parce que, comme le dit Mooers quelque part, on utilise des machines qui sont "idiotes", qui ne sont pas dotées de la faculté d'"apprendre". Pour certains textes, l'analyse statistique assez brutale "inintelligente", selon la méthode de Luhn, peut sans doute aboutir à des résultats à peu près satisfaisants, mais on ne va pas très loin dans cette voie. Pour aller plus loin, il faudrait des machines "intelligentes"; ces "inductive inference machines" comme les appelle Solomonoff, ne sont pas nées, mais il y a de bons espoirs de les voir naître. Et, le jour où elles seront nées, il sera possible de leur confier des tâches beaucoup plus compliquées. Il y a eu deux rapports présentés à la conférence de l'UNESCO en juin dernier qui montrent une certaine avance dans cette voie. Les machines de ce type pourront probablement se charger de rédiger, à partir de données qui leur seront fournies, des textes d'information pour l'usager. C'est à partir de ce moment, pas avant, que l'on pourra parler vraiment d'une nouvelle époque et que l'on sera passé, après les étapes de la parole et de l'écrit, au langage-machine proprement dit, à l'étape où les archives scientifiques ne seront plus seulement en forme de livres ou de n'importe quel dérivé des livres, mais d'une mémoire électronique (ou autre...). Nous n'en sommes pas encore là. Mooers pense que cette étape viendra dans quelque vingt ans. Peut-être....

De toute façon, une question qui dominera sans doute la série de leçons de cette semaine sera : que doit-on entendre exactement par "langage des machines" ? Est-ce que l'on veut parler de la "langue" utilisée à la sortie ? Le Professeur Pietsch y a fait allusion ce matin. Il nous a dit qu'il n'existait pas de difficultés au niveau du langage interne, déjà bien connu des spécialistes; ce qu'il est important de rechercher à l'heure actuelle, c'est le caractère des "langues" des codifications qui seront employées pour l'alimentation des machines. Je pense, personnellement, que nous nourrirons celles-ci de plus en plus avec des matériaux bruts, et cela pour des raisons d'économie.

Quelqu'un a posé ce matin la question de savoir comment traiter, non plus les informations qui arrivent au jour le jour, mais l'arriéré.

C'est en effet un problème énorme. Pour codifier, en vue de les enregistrer sur machines, par exemple, les informations contenues dans un grand office de brevets tel que ceux de La Haye, de Washington, de Munich ou de Paris, il faudrait des dizaines de milliers d'années de travail d'ingénieur.

Des travaux de ce genre ne seront possibles que si l'on trouve des méthodes permettant que les machines puissent ingérer (si l'on veut nous permettre l'emploi de ce terme emprunté aux sciences de la digestion) des textes bruts, des textes tels qu'on les écrit, tels qu'on les publie. Il faut avouer que nous ne sommes pas encore très avancés à ce point de vue, en partie à cause des difficultés que soulève la lecture automatique des caractères. Nous avons bien déjà des machines capables de lire les chèques, mais seulement si ceux-ci sont écrits avec une écriture standardisée spécialement adaptée à la reconnaissance par la machine. Tout autre est le cas des textes "normaux" que vous et moi avons à lire et que les machines ne savent pas lire pour le moment.

Par ailleurs, il sera aussi nécessaire de réaliser de grands progrès dans le domaine de la linguistique appliquée à la documentation automatique. Vous verrez dans les leçons qui vont venir combien les recherches structuralistes de Harris, de Chomsky et d'autres, ou les recherches de Ceccato sur l'expression des relations ont déjà été utiles dans ce domaine, et je crois que leur développement ne fait que commencer. La recherche des informations deviendra sans doute, comme le prédit Andreev en URSS, une partie de la linguistique appliquée. A ce propos, je pense qu'il faut attirer votre attention sur les différences qu'il y a cependant entre la recherche automatique des informations et la traduction mécanique. Il est vrai que l'on m'a fait l'honneur de me charger (et c'est là vraiment une charge, au sens fort du mot) de présider le sous-comité "recherches" d'un comité dont le nom est à peu près intraduisible en français "International Continuation Committee on Mechanical Translation and Information Retrieval" (peut-être pourrait-on le traduire approximativement par "Comité International pour la Traduction Mécanique et la Documentation Automatique"). On a donc jumelé là les deux questions, mais on n'a pas encore réussi à déterminer exactement en quoi la recherche automatique des informations et la traduction mécanique se rapprochent et en quoi elles diffèrent. Je crains pourtant que cette question dépasserait les limites de temps qui me sont imparties ce matin.

Je voudrais terminer en disant que c'est sans doute vraiment une sorte d'"expérience historique" que nous vivons ici cette semaine. Je n'ai pas l'habitude d'être emphatique, mais tout de même je crois qu'il faut remercier l'EURATOM, et tout particulièrement Braffort et Leroy, d'avoir, pour la première fois, mis sur le plan de l'enseignement pratique des méthodes d'avant-garde qui étaient restées jusqu'à présent sur le plan de la recherche ou des discussions entre quelques spécialistes.

JOURNEE DE MATHEMATIQUES

INTRODUCTION A LA JOURNEE DE MATHEMATIQUES

P. BRAFFORT

Au cours de la journée d'hier, M. Pietsch et M. de Grolier ont présenté un tableau des centres de documentation et des systèmes de documentation qui vous ont permis de voir à quel point les méthodes utilisées et les problèmes posés étaient nombreux. Comment, alors, choisir un système correct dans votre propre centre de documentation?

Il existe de nombreux systèmes tant du point de vue technique que du point de vue des méthodes de classification, et souvent dans les conférences on se trouve en présence de petits combats singuliers entre les tenants de tel ou tel système qui estiment avoir trouvé la pierre philosophale. C'est notamment le cas lorsqu'on se situe au niveau pré-automatique et au stade de la fabrication de systèmes de classification. Vous savez qu'il existe notamment la célèbre classification décimale universelle qui est encore utilisée dans un certain nombre de centres et de périodiques; vous savez aussi que d'autres systèmes, tel que le système CORDONNIER, le système PAGES etc... emploient les techniques linguistiques compliquées, et ce n'est que depuis peu de temps qu'on a senti la nécessité de justifier l'emploi de tel ou tel système.

En fait, les auteurs, dans la plupart des cas, pensaient que les qualités de leur système devaient sauter aux yeux et que, par conséquent, il n'était pas besoin de les justifier davantage. Mais avec l'automatisation des fonctions documentaires, on est préoccupé par d'autres problèmes : les problèmes financiers. Lorsqu'il s'agit de payer quelques documentalistes, les payer fort peu en général, cela va très bien. Quand il faut acheter des machines, qui représentent des investissements plus considérables, alors on nous demande des comptes. Pour ces comptes, nous voudrions essayer de présenter quelques techniques de calcul. Dans quel domaine ces techniques de calcul devront-elles être appliquées ? Pour le savoir, il faut se reporter aux quelques textes très peu nombreux dans lesquels des auteurs se sont efforcés de justifier l'utilisation de systèmes.

Mais qui dit estimation dit utilisation de techniques statistiques. Ces techniques portent sur plusieurs domaines : d'une part, les documents; il s'agit donc de statistiques sur le nombre de documents intéressants retrouvés lors d'une recherche rétrospective,

la proportion des documents corrects par rapport aux documents incorrects, la valeur d'un échantillonnage de documents pour juger la qualité d'un système documentaire etc... D'autre part, il faut aussi connaître les coûts des recherches tant en ce qui concerne les salaires des documentalistes ou des personnes qui se servent des machines que des coûts liés au temps d'utilisation des machines. C'est à l'aide de ces éléments qu'on peut aborder d'une façon quantitative l'estimation d'un système documentaire. Sans avoir élaboré complètement de telles méthodes et pour n'avoir fait que des essais partiels, on se rend cependant compte dès maintenant qu'il n'y aura pas une solution universelle et qu'il n'est pas question de décider à la fin d'une enquête quel système, quelle méthode de classification et quelle mécanique résolvent le problème de la documentation. Et ceci parce qu'il n'y a pas un problème unique de la documentation. Il suffit de travailler pendant un certain temps dans un grand organisme scientifique comme l'Euratom pour se rendre compte que, à chaque niveau d'organisation de l'institution scientifique en question, il se pose un problème documentaire en particulier. C'est ainsi que le problème de la documentation générale disons de l'Euratom n'est pas le même que celui de la documentation de la direction des recherches qui, lui-même, n'est pas le problème de la documentation du groupe d'études sur les échanges thermiques. Un laboratoire qui étudie les échanges thermiques peut avoir à manipuler un ensemble de documents, de quelques dizaines de milliers de documents, et l'Euratom peut avoir à manipuler quelques millions de documents. On se rend bien compte que le problème de coût et du nombre sont fondamentalement différents. Vous voyez, par conséquent, que c'est l'aspect numérique et statistique des populations documentaires en question qui détermine justement quel système pourra ou ne pourra pas être qualifié de système optimal.

D'un autre côté, en même temps qu'on s'approche d'une automatisation plus complète on abandonne les classifications à tendances philosophiques pour en venir à des systèmes d'expression des textes scientifiques plus proches du langage naturel, plus linguistiques. C'est un autre aspect de la statistique qui apparait : celui de l'utilisation des statistiques pour l'étude des textes, pour l'étude du langage, notamment du langage scientifique spécialisé. Ceci explique la répartition des différentes conférences qui sont présentées aujourd'hui et demain, ceci fait apparaître également une relation entre les cours d'aujourd'hui et les cours de demain. Dire qu'on va faire de la statistique ne veut pas dire qu'il faut négliger les aspects que l'on peut appeler "certains" des structures mathématiques, ces aspects sont, au contraire, ceux qui ont été les plus couramment exploités jusqu'alors, ainsi qu'on le verra au cours de la journée de linguistique.

Lorsqu'on pénètre un peu profondément dans l'étude du langage, que ce soit langages naturels ou langages artificiels (comme les systèmes formels), on éprouve l'unité profonde du syntaxique et du sémantique. Eclairer ces rapports, c'est éclairer les rapports du statistique et du certain dans les structures mathématiques. C'est pourquoi la journée de Mathématiques se devait de débuter par l'analyse d'une notion qui est à la racine de ce point de rencontre : c'est la notion d'information.

THEORIE DE L'INFORMATION

M. SCHUTZENBERGER +

Dans l'état actuel de son développement, la théorie de l'information se présente comme une théorie essentiellement mathématique assez peu élémentaire et il me sera donc difficile d'en donner autre chose qu'un aperçu sommaire.

Schéma général

Tous les problèmes qui se posent dans la théorie mathématique des communications peuvent être considérés comme des cas particuliers du schéma suivant :

un émetteur choisit un "message" dans une certaine liste; ce message est codé en un certain signal qui est transmis sur une ligne où il est éventuellement soumis à des perturbations aléatoires (bruit).

A la réception, le signal est décodé et on compare (idéalement) le message qui a été envoyé au message que le récepteur croit avoir décodé, de façon à calculer le gain (positif ou négatif) qui résulte de l'opération. Pour compléter le bilan, on fait intervenir le "coût de transmission" comportant à la fois des frais fixes et des dépenses proportionnelles à la longueur du message transmis.

En raison du caractère extrêmement général de ce modèle, il est clair qu'aucun problème vraiment intéressant ne peut être posé avant que soient délimités de façon plus stricte chacun des éléments qui le composent.

Je vais donc reprendre un par un les différents éléments et montrer quelles hypothèses il est naturel de faire - ou du moins quelles hypothèses mènent à des résultats praticables.

Les messages : Une division à priori s'impose : ou bien la liste des messages est finie, ou bien la liste des messages est infinie et, dans ce cas, nous ne pourrons en parler utilement que si nous lui avons conféré une structure.

+ Faculté des Sciences de Poitiers

Encore faudra-t-il distinguer deux cas : dans le premier, chacun des messages est, en réalité, une suite de "messages élémentaires" extraits d'une liste plus simple; c'est le cas le plus évident : la liste des messages que vous pouvez vouloir envoyer par la poste est infinie, ce sont toutes les phrases ou suites de phrases possibles; mais, de fait, chacun de ces messages n'est qu'une suite (infinie) de symboles appartenant à un ensemble fini : celui des signes typographiques (l'alphabet, les chiffres, les signes de ponctuation etc...) qui constituent l'ensemble des messages élémentaires.

A l'opposé, la structure de la liste des messages pourra - ou devra - dans certains cas, être prise dans toute sa complexité : c'est ce qui se produirait par exemple si le "message" était une mesure dont l'expression exacte n'a à priori aucune chance de n'exiger qu'un nombre fini de décimales.

Le codage : Toutes les distinctions qui viennent d'être faites valent aussi bien pour les signaux : ou bien ceux-ci seront en nombre fini ou bien ils seront des suites de signaux élémentaires - ou bien le signal sera continu. Du point de vue mathématique, ces cas commandent des techniques assez différentes et divisent la théorie en plusieurs chapitres dont le développement est fort inégal.

Le bruit : Puisque nous opérons dans l'abstrait, le bruit sera simplement la donnée pour tout signal émis de la probabilité pour que le signal soit reçu sous telle ou telle forme.

Naturellement, dans chaque problème concret il se posera une question préalable consistant à déduire du modèle de la structure physique ces probabilités, mais, au niveau où nous traitons cette question, il est possible de supposer que cette réduction a déjà été effectuée.

Le décodage : Peut-être qu'à priori ceci vous semble l'aspect le plus important de toute la question, et cependant nous allons l'escamoter entièrement en prétendant que la situation qui a été décrite jusqu'ici n'a été qu'un langage nouveau pour "coder" le problème général de la statistique mathématique, et en priant nos collègues statisticiens de résoudre la question pour nous.

Le rôle du théoricien des communications a été d'établir le schéma général, de choisir le meilleur code et, ceci fait, de nommer au poste de décodage un statisticien qui sera censé opérer au mieux en appliquant ses propres techniques.

Le bilan : Peut-être est-il plus simple de citer en exemple les cas les plus fréquents :

Si la liste des messages est finie, on prendra presque toujours comme fonction de coût d'erreur celle qui correspond à une pénalisation d'une unité si le message a été incorrectement décodé et à aucune pénalisation dans le cas opposé.

Si le message est une quantité continue, on prendra souvent une pénalisation proportionnelle au carré de la différence entre la valeur émise et la valeur reçue.

En ce qui concerne le coût de transmission, le plus fréquent sera de le supposer proportionnel au temps, c'est-à-dire à la longueur des signaux codés.

Il est clair que pour un niveau donné de bruit, en répétant par exemple un très grand nombre de fois la transmission de chaque message (c'est-à-dire en augmentant le coût de transmission) on peut réduire autant qu'on veut ces chances d'une erreur (c'est-à-dire la valeur moyenne du coût d'erreur). C'est, si vous voulez, ce qu'on fait lorsque sur un chèque on écrit en toutes lettres (c'est-à-dire beaucoup plus longuement) la somme à payer - ou quand on répète plusieurs fois le même mot pour mieux se faire entendre.

Cependant, puisque nous avons supposé que les coûts d'erreur et les coûts de transmission pouvaient être comptabilisés dans la même unité ("time is money") il doit exister dans chaque problème précis un optimum qui se situe à une certaine distance (à trouver) entre les deux extrêmes théoriques :

- répétition très longue et donc frais de transmission élevés, mais coût d'erreur faible.
- pas de transmission du tout, donc frais nuls pour cette rubrique, mais maxima pour la rubrique des risques d'erreur.

La théorie de la capacité de Shannon.

Quand j'ai parlé de répétition du message, j'ai été beaucoup trop schématique : il est possible en général de trouver des codes qui sont plus efficaces que la répétition pure et simple; si vous voulez un exemple, pensez que vous avez jugé plus sûr d'envoyer deux lettres à un ami lointain pour l'informer d'une affaire d'importance; s'il y avait une chance sur cent pour qu'une des lettres se perde, il n'y a plus qu'une chance sur dix mille (ce qui est très peu) pour que toutes les deux soient égarées.

Mais, outre cette augmentation de la sécurité, vous avez pu avoir un avantage marginal : à côté de l'essentiel, vous avez peut-être présenté des détails secondaires différents dans chacune des deux lettres et si, par chance, elles arrivent toutes les deux, votre ami sera mieux informé que par un seul message.

L'un des mérites de Shannon est d'avoir conçu qu'un phénomène analogue est la règle : il est théoriquement possible d'allonger le code (donc d'accroître la sécurité) et, en même temps, de transmettre plus d'information.

En outre, et c'est là l'essentiel, encore que je ne puisse entrer dans les détails, Shannon a montré qu'il existe une certaine quantité relativement facile à calculer, ne dépendant que du bruit, et qui donne une limite à l'efficacité de tout codage convenable pour un schéma donné. C'est ce que l'on appelle la "capacité" de la ligne.

Essayons encore de prendre un exemple simple : vous admettez que si les télégraphes étaient infaillibles (si les lignes étaient "non bruyantes") on pourrait considérer que la quantité de détails que vous pouvez transmettre à votre correspondant est proportionnelle au prix du télégramme.

Le théorème de Shannon montre que la situation n'est pas modifiée si les chances que des erreurs se produisent sont non nulles.

Grâce au phénomène que je mentionnais plus haut, vous pouvez, par un codage astucieux, faire deux choses à la fois : diminuer à volonté le risque d'erreur d'une part, d'autre part transmettre des détails en quantité proportionnelle à la longueur du texte. Tout simplement à cause du bruit, tout se passe comme si le prix unitaire de chaque mot était majoré (par rapport au cas idéal où le dispositif de transmission est parfait) par un certain facteur fonction de cette capacité.

Pour être honnête, je dois mentionner que naturellement tout ceci ne peut être démontré que sous des hypothèses précises, assez limitées et qu'en particulier quand les messages et les signaux sont continus la théorie n'est encore qu'assez peu développée.

Le cas non bruyant :

Le cas qui peut-être vous intéressera le plus est le cas non bruyant; celui où la ligne fonctionne à la perfection, car il contient l'essentiel des applications de la théorie aux problèmes qui vous préccupent.

Pour en parler plus commodément, je vais faire un changement de terminologie dans le modèle initial et introduire en même temps les hypothèses qui semblent naturelles dans ce cas.

Nous supposons donc désormais que l'ensemble des messages possibles est fini et, encore plus concrètement, que le modèle est le suivant :

L'émetteur est simplement une grande collection qui comprend les objets X_1, X_2.....etc. avec les fréquences p_1, p_2,..... etc. La ligne, elle, est simplement un alphabet comportant un nombre fini de "lettres" et tous les mots que l'on peut faire avec les lettres. Le codage consiste simplement à attribuer une fois pour toutes à chaque objet un "mot code" dans cet alphabet. Il est commode de supposer que l'alphabet n'a que deux lettres.

Enfin, le fonctionnement du modèle est réduit au schéma très simple suivant:

Le système des mots code ayant été choisi, l'émetteur tire au hasard un objet de la collection et transmet le mot code au récepteur pour que celui-ci sâche quel objet a été tiré.

On veut que la transmission soit sans erreur (deux objets distincts ont des mots code différents). Les mots code doivent être choisis de telle sorte que, compte tenu de la fréquence relative des différents objets, la longueur des mots soit la plus petite possible en moyenne.

Reformulons encore ce modèle de façon différente en attribuant un rôle plus actif au récepteur : à chacune des lettres successives du mot qu'il lit sur la bande où est enregistré le signal reçu, le récepteur reçoit la réponse à une question implicite du type suivant :

Quelle est la n-ième lettre du mot code de l'objet choisi ?

En d'autres termes, quand le récepteur voit sur la bande que la n-ième lettre du mot est une certaine lettre x , il apprend que l'objet inconnu appartient à un sous ensemble des objets qui, dans le code, choisi, ont un x à la n-ième lettre.

Finalement, vous voyez que nous pouvons laisser de côté l'aspect codage et que le fonctionnement peut être décrit de la façon suivante :

L'émetteur tire au hasard un objet X et le récepteur lui pose une série de questions de la forme suivante :

Est-ce que X appartient ou non à tel sous ensemble de la collection ?

Le problème est alors de savoir choisir les questions à poser de telle façon que le récepteur puisse le plus rapidement possible identifier l'objet.

C'est ici qu'intervient la formule de Shannon qui permet de fixer a priori des limites assez strictes au nombre moyen minimum des questions qui seront nécessaires en fonction de la répartition initiale des probabilités.

Pour rendre intuitif ce résultat, considérons d'abord les deux cas limites suivants ou, pour simplifier, nous supposons que l'objet appartient à une collection de 2^N objets.

Dans le premier cas nous faisons l'hypothèse que toutes les probabilités - sauf une - sont nulles : dans ce cas - qui est évidemment trivial - le récepteur sait à l'avance, sans avoir besoin de poser de question, quel est l'objet.

Dans le deuxième cas, nous faisons l'hypothèse que tous les types d'objets sont également probables à priori. Si A est un sous ensemble de $n' < 2^N$ objet et si le récepteur demande "est-ce que l'objet appartient à A ?", il peut :

- soit apprendre que tel est bien le cas - et alors on est ramené au problème précédent avec n' au lieu de 2^N ;
- soit apprendre que tel n'est pas le cas et alors on est encore ramené au problème précédent, mais avec $n'' = 2^N - n'$ objets au lieu de 2^N.

En particulier, si $n' = 2^{N-1}$, le récepteur a une chance sur deux qu'on lui réponde oui et, quelle que soit la réponse, il est ramené à la position initiale mais avec une collection deux fois plus petite.

Clairement, avec N questions de ce type il parviendra sûrement à identifier l'objet.

Il n'est pas absolument trivial de montrer que, quelle que soit la méthode employée - N questions au moins sont nécessaires.

Le théorème de Shannon affirme que, si les probabilités initiales sont $p_i = (i = 1, \ldots\ldots n^t)$ il est possible de trouver un système qui fournisse la réponse en moins de $\sum p_i \log_2 1/p_i = H$ questions en moyenne; ($\log_2$ = logarithme de base 2); dans l'exemple précédent on avait $p_i = 2^{-N}$ et H était bien égale à N.

Ceci est l'aspect en quelque sorte négatif du théorème. Il y a aussi un aspect positif qui est le suivant :

- Il existe une procédure pour poser les questions qui aboutit à l'identification de l'objet en moins de H + 1 questions en moyenne.

Très schématiquement, cette procédure consiste à choisir chacune des questions successives de telle manière qu'elle ait, a priori, une chance sur deux d'avoir une réponse positive : ceci correspond assez

bien à ce que suggère l'intuition : une question est d'autant plus "informative" que sa réponse est plus inattendue.

Prises ensembles, ces deux inégalités justifient donc que l'on utilise cette quantité H pour définir l'information (ou plutôt d'ailleurs l'ignorance) que le récepteur avait à priori sur l'objet. Naturellement, elle n'est une description de cette ignorance que relativement au système des probabilités p_i et nous reviendrons plus en détail sur ce point.

J'insiste toutefois sur le caractère opérationnel de la définition : l'"information" est définie par la quantité de travail nécessaire pour aboutir à la connaissance - et non pas sur telle ou telle propriété mystérieuse de la fonction $\sum p_i \log_2 1/p_i$ encore que, bien entendu, ce n'est qu'en vertu de ses propriétés mathématiques particulières que cette fonction est apte à caractériser ces limites opérationnelles.

Considérons à titre d'exemple la signification qui peut être attachée à la mesure de l'information qu'apporte la connaissance d'une lettre d'un texte : S'il s'agit d'un texte dans lequel les lettres sont arbitrairement choisies (un système de mots code), ce sera d'après les relations précédentes à peu près 5 unités puisqu'il y a à peu près 32 signes typographiques.

S'il s'agit d'un texte français, ou anglais, ou allemand, cette valeur tombe à environ 50% du chiffre précédent. En effet, comme on le sait, la fréquence des différents signes est inégale, des contraintes à courte et à longue distance les relient assez étroitement entre eux - bref la connaissance à priori que l'on a sur une lettre d'un texte réel est assez élevée.

A la limite, s'il s'agit d'une lettre manquante dans un contexte connu, l'information apportée est presque nulle puisque, sans poser aucune question, il est presque toujours possible au lecteur de réparer l'omission ou de corriger l'erreur.

TRAVAUX PRATIQUES DE LINGUISTIQUE STATISTIQUE

J. LARISSE

Loi de Zipf Mandelbrot

I. Théorie On étudie dans un texte d'un auteur donné dans une langue donnée la relation entre la fréquence relative f=f (r) d'utilisation d'un mot et son rang r, c'est-à-dire le nombre entier qui numérote ce mot dans la liste par ordre de fréquences décroissantes des mots utilisés. L'expérience montre qu'entre f et r existe en première approximation la loi dite d'Estoup Zipf:

$$f(r) = \frac{PT}{r} \qquad \begin{array}{l} T = \text{nombre total des mots du texte} \\ P = \text{facteur de normalisation} \end{array}$$

ce qu'on peut encore écrire:

$$\log f(r) = \log PT - \log r$$

On peut approcher les résultats expérimentaux de manière plus satisfaisante en utilisant la formule:

$$f(r) = PT\,(r + \rho)^{-B} \quad \text{ou}$$
$$\log f(r) = \log PT - B \log (r + \rho)$$
$$\log f(r) \sim \log PT - B \log r \text{ pour } r \text{ grand.}$$

La fig. 1 montre la modification apportée à la loi d'Estoup - Zipf en introduisant les deux paramètres:

B = qui caractèrise la pente de la partie rectiligne (r grand)

ρ = qui rend compte de l'aplatissement de la courbe expérimentale aux hautes fréquences.

Il existe plusieurs manières différentes d'expliquer par des hypothèses simples cette loi. Celle-ci se présente en linguistique, tout au moins sous certains aspects, comme la loi a laquelle on doit aboutir à partir d'hypothèses simples.

Signalons en particulier avec Mr. Mandelbrot que dans l'hypothèse de la génération d'un discours par un processus probabiliste de Markov des lettres et d'un signe spécial: l'intervalle, on explique la régularisation des fréquences suivant la loi d'Estoup-Zipf sous sa forme généralisée:

$$f(r) = P\,(r + \rho)^{-B}$$
avec la restriction $B > 1$

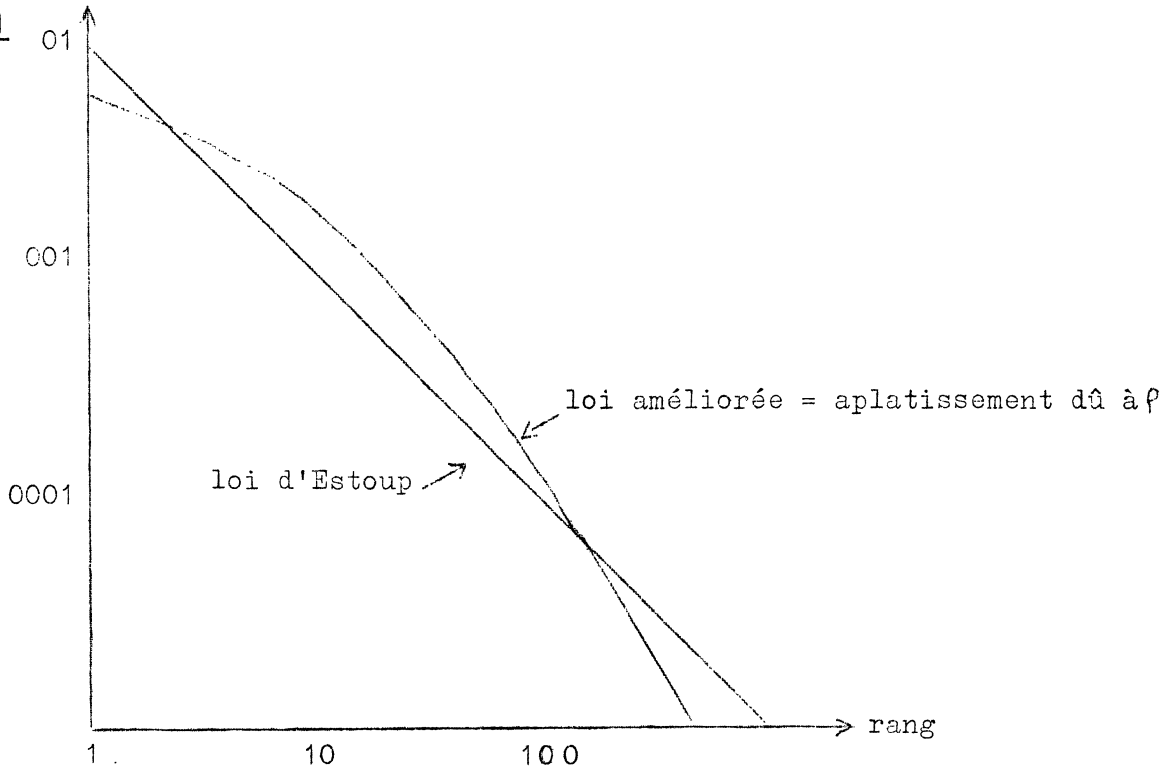
On peut également parvenir aux mêmes conclusions, et de plus B quelconque, par un critère de type thermodyna mique. On définit ce qu'on peut appeler "l'état d'un discours" et on pose que cet état doit être "le plus probable". +

+ Pour un développement non mathématique de ces considérations voir "Linguistique Statistique Macroscopique" dans l'ouvrage de Apostel, Mandelbrot, Morf: "Logique Langage et Information", Presses Universitaires de France, 1957.

Avec M[r] Belevitch on peut montrer que les lois de Zipf et de Mandelbrot se présentent comme les deux premiers termes de l'approximation de Taylor d'une loi de destribution arbitraire de la probabilité d'utilisation d'un mot dans un texte.++

II. Travail pratique: Chaque groupe de 4 élèves dispose d'un ouvrage en langue anglaise, allemande ou française, d'une photocopie de l'Index alphabétique. La manipulation consiste à étudier la fonction f (r) limitée aux mots de l'Index. En première approximation on retrouvera la loi de Zipf. On propose qu'un élève lise lentement un texte chosie en commun, un deuxième signale les mots figurant dans l'Index et les deux autres recopient les mots par ordre alphabétique en les cochant chaque fois qu'ils se présentent. De cette manière le décompte est facilité. Ensuite, classer les mots par ordre de fréquences décroissantes, le numéro d'ordre étant le rang, et tracer sur papier bilogarithmique la courbe f (r). Les résultats se précisent en moyenne sur un texte de quatre à cinq pages. Il est recommandé d' nalyser une dizaine de pages, les courbes permettront de tirer des conclusions sur la construction des Index.

fig. 1



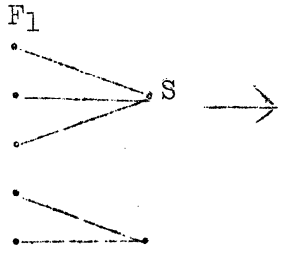
++On the Statistical laws of linguistic distribution". Annales de la Société Scientifique de Bruxelles, 1959.

JOURNEE DE LINGUISTIQUE

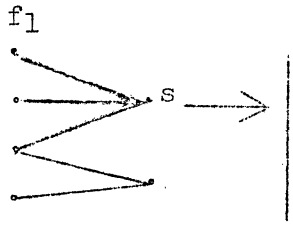
INTRODUCTION A LA JOURNEE DE LINGUISTIQUE

A. LEROY

En documentation nous sommes essentiellement intéressés par le contenu des textes scientifiques. En effet, les questions qui sont posées à un système documentaire ne concernent que la signification; toutefois, il est bien évident qu'il n'y aurait pas de signification sans symboles porteurs de cette signification, sans "forme signifiante" et nous sommes donc amenés à nous intéresser en tout premier à la correspondance = signification - forme signifiante. Elle est bien mise en évidence sur le schéma suivant tiré de celui de SESTIER. (1)



Les F_1 sont des formes du langage étudié, juste suffisantes pour être interprétables sans équivoque, abstraction faite du contexte. Les S sont des contenus signifiés.



Les f_1 sont des formes élémentaires du langage étudié, les s les contenus signifiés correspondants.

On conçoit qu'il serait très utile de trouver un langage qui exprime réellement le contenu d'une manière univoque, c'est-à-dire tel que l'on pourrait avoir =

$$\frac{F_1}{f_1} \quad \frac{S}{s}$$

On voit qu'il s'agit surtout de tenir compte des problèmes de synonymie et de polysémie - Mais c'est en fait ici que commencent nos difficultés car pour dire que deux expressions ont des sens identiques, voisins ou totalement différents, il faut pouvoir déterminer les sens en question. Or, ce problème est extrêmement complexe. L'analyse du processus de la connaissance permet de s'en faire une idée . P. CHAUCHARD dit à ce sujet que les organes sensoriels envoient les signaux sous les diverses influences du monde extérieur et l'ensemble de ces signaux permet à l'intérieur de l'homme une reconstitution plus ou moins précise de ce monde. Il construit ainsi une vision du monde qu'il cherche, bien sûr, à rendre la plus objective possible, se basant sur les conséquences pratiques de ses pensées. Mais il y aura toujours forcément une part subjective, les messages sensoriels étant déjà une déformation des structures réelles et leur intégration dans les structures cérébrales productrices d'idées et d'abstractions étant encore plus déformante.

(1) SESTIER A. "La traduction automatique".
Ingénieurs et Techniciens, mars, avril,
mai, juin 1959.

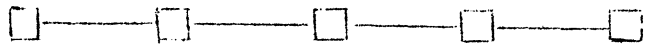
Une des conséquences en est que chacun "voit" un monde différent et on comprend facilement que les mots que l'on emploie pour rendre compte de ce que l'on voit ou ressent puissent donner lieu à des confusions.

On peut penser que ces différences individuelles viennent de ce que l'homme constitue une sorte de "compromis" réalisé en vue de lui permettre avant tout de subsister dans la grande variété des conditions régnant sur la terre : notre constitution nous empêche en quelque sorte d'apercevoir le simple qui existe fondamentalement dans la nature mais qui, répété un très grand nombre de fois, n'apparait que sous forme complexe; elle ne nous permet que d'y déceler des choses globales, résultat d'un processus de simplification subjective. On peut dire que c'est à ce stade qu'apparaît la "qualité" alors que la matière en elle-même n'est que "quantité".

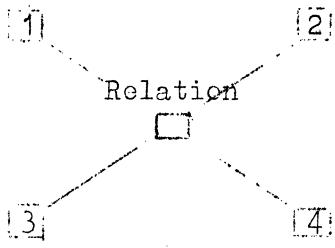
Incapable de s'adapter à la simplicité quantitative du monde extérieur, l'Homme y cherche la simplicité qualitative; les mots du langage reflètent ce désir, englobant dans une seule expression des choses parfois extrêmement complexes, dont la définition fait intervenir plusieurs notions elles-mêmes complexes. C'est ce qui fait dire qu'un mot n'est jamais quelque chose que l'on peut isoler; il n'a de signification que par ses relations avec un très grand nombre d'autres notions, à l'intérieur du "réseau général de la connaissance" -A un même mot, chacun fera correspondre une partie différente du réseau, d'ailleurs très mal délimitée et qui dépend de l'expérience du sujet; on voit là apparaître la notion de "flou sémantique".

Les difficultés rencontrées sont donc notables (Nous verrons dans les prochaines journées comment nous comptons y remédier). Encore n'avons-nous pas considéré jusqu'ici celles qui ont trait à la mise en évidence de la structure du langage. Il est souvent dit que ces questions sont du domaine de la syntaxe, mais on sait depuis longtemps que la sémantique n'est pas indépendante de la syntaxe, par exemple que la signification n'est pas indépendante de l'ordre des mots, du moins pour les langues qui nous intéressent. D'ailleurs très souvent la simple considération de la structure nous permettra de construire des éléments d'un réseau à partir duquel certains problèmes sémantiques seront résolus systématiquement.

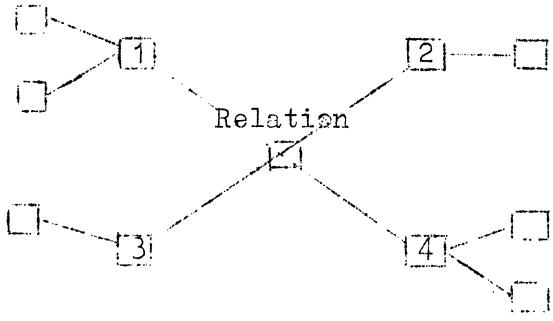
Pour avoir une idée de la notion de structure, demandons-nous ce qu'est la construction d'une phrase. Construire une phrase c'est, en particulier, mettre un certain nombre de mots en relation - Les exigences du parler font que la phrase apparaît sous la forme d'une chaîne linéaire que l'on peut représenter de la manière suivante :



En réalité, parmi cette suite de mots il y en a au moins un qui joue un rôle spécifiquement relationnel *, si bien que l'on peut imaginer un schéma à deux dimensions plus conforme à ce que l'on veut exprimer :

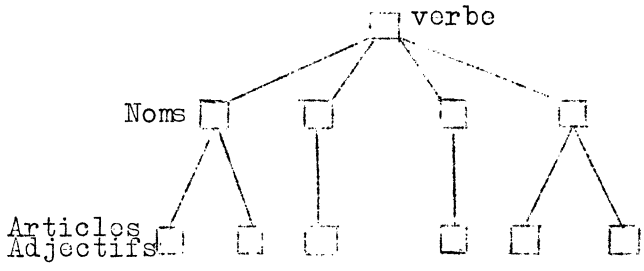


Les mots 1, 2, 3, 4, peuvent recevoir des qualificatifs qui viennent s'ajouter par exemple de la manière suivante :



Les relations qui comprennent essentiellement des verbes, possèdent plusieurs "branches" qui peuvent recevoir des repères spéciaux constitués par les prépositions; les mots 1, 2, 3, 4, sont essentiellement des noms; les qualificatifs sont des articles et des adjectifs ou des expressions jouant ce rôle.

Ce dernier schéma peut être disposé de différentes façons; la disposition en arbre par exemple nous rapproche de ce que TESNIERE (1) a appelé le "stemma"



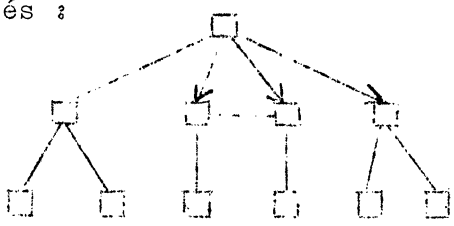
On verra dans les exposés suivants quels sont les différentes figures qui ont été choisies par les spécialistes de la question, et les différentes manières d'en parler. On pourra ainsi étudier notamment la méthode CECCATO, présentée par l'auteur lui-même, la méthode de CHOMSKY, présentée par M. BRAFFORT, et la méthode de HARPER et HAYS présentée par M. IUNG à l'occasion de son exposé sur la traduction automatique. Enfin la méthode GRISA sera illustrée dans plusieurs exposés pendant lesquels on insistera également sur la méthode de TESNIERE.

---

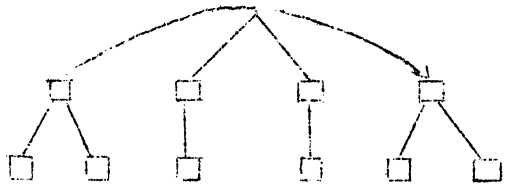
* dans une phrase bien construite.

(1) TESNIERE - Eléments de syntaxe structurale KLINCKSIECK (Paris) 1959

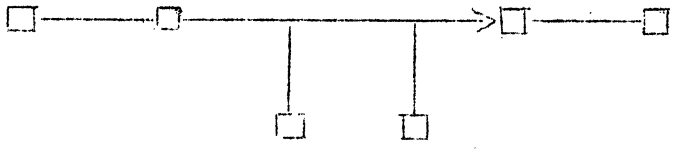
Au cours de ces exposés on aura l'occasion de parler des liaisons d'un autre type, et qui sont appelées "anaphores" par TESNIERE; une liaison anaphorique est, par exemple, celle qui existe entre un nom et le pronom qui le remplace dans une phrase quelconque. On peut la représenter en pointillés :



Du point de vue du traitement en machine, il est commode de représenter l'élément relationnel sous la forme d'une flèche avec des branches latérales



De plus, la liaison anaphorique permettra de simplifier dans certains cas le diagramme (liaison nom-pronom). Enfin il est possible de supprimer certains éléments du langage qui sont souvent redondants; il en est ainsi des articles - Le diagramme pourra finalement prendre la forme suivante :



Quelles sont alors les difficultés qui se posent pour obtenir un tel diagramme d'une façon systématique ? Nous le verrons à propos de l'exposé de M. LECERF et au cours des travaux pratiques de linguistique. Nous pouvons déjà dire qu'elles proviendront des cas d'exception; si, en effet, tous les noms, tous les verbes etc... se comportaient de la même façon, nous n'aurions pas grand mal à construire systématiquement les représentations ci-dessus. Nous pressentons ici qu'une partie de la recherche consistera à essayer de découvrir des règles qui permettront la construction la plus économique possible.

Enfin, il ne pouvait être question de parler de linguistique sans faire mention de l'application de la statistique dans ce domaine, qui représente une aide efficace dans l'étude du comportement des mots et des rapports entre les mots, étude que nous venons de montrer être indispensable. C'est M. BELEVITCH qui développera ces questions.

I PROBLEMI FILOSOFICI DEL LINGUAGGIO

S. CECCATO *

Signore e Signori,

trovandomi a parlare dei risultati di un lavoro ormai ventennale e per la più parte originale, mi dovrò limitare a fissarne alcuni aspetti.

Anzitutto devo prendere posizione di fronte allo stesso tema che mi è stato assegnato dagli amici Braffort e Leroy, "I problemi filosofici del linguaggio". Per chi è a giorno dell'indirizzo delle nostre ricerche, quel titolo sembrerà fatto apposta per scatenare la lunga serie delle dichiarazioni critiche, in quanto noi riteniamo che, per chi ha intenti costruttivi e tecnici, in cui figuri in qualche modo il linguaggio, il primo dovere sia quello di eliminare ogni filosofare.

Una posizione cosi radicale discende da due ordini di considerazioni : critiche e costruttive.

Le considerazioni critiche denunciano in ogni filosofia un presupposto di tipo conoscitivo per cui viene posta fra le cose nominate e le parole una attività conoscitiva. Questa consiste nell'assumere le cose nominate quali incognite date ad un conoscente che se le fa cognite attraverso un loro raddoppio; questo raddoppio, poiché le cose cognite ed incognite devono restare eguali fra loro, deve essere di tipo spaziale e temporale insieme, cioè il raddoppio deve avvenire in un altro posto da quello della cosa incognita originale e quindi anche in un altro momento, per consentire un passaggio dall'una all'altra.

(Sento già le obbiezioni di chi si occupa di filosofia e ricorda i vanti dell'idealismo, che avrebbe eliminato ogni conoscere. Ma, ahimé, almeno per un pezzo, lo spirito, il pensiero, l'atto, è rimasto con la sua conoscitiva datità; e nelle correnti idealistiche, quando si è discesi dal grande programma ai problemi particolari, la trattazione ha ripercorso le classiche vie).

L'atteggiamento conoscitivo ha molte conseguenze, che informano speculazioni dello scienziato che intende commentare filosoficamente la sua tecnica, dalla matematica alla biologia, o farsi metodologo, ed infiorano spesso anche le nostre dissertazioni quotidiane. Per quanto riguarda il linguaggio la principale conseguenza per noi è che le parole avranno un corrispondente soltanto nel caso in cui la cosa designata sia di tipo osservativo (come "tavolo", "bicchiere", "albero", etc.), o si possa supporre "lasciata" da una di queste, "astraendone" alcune

* Centro di Cibernetica e di Attivita Linguistiche - Milano

delle caratteristiche. Infatti, la "conoscenza" deve raddoppiarle per posto e momento, e soltanto gli osservati hanno come costitutivi spazio e tempo. Le parole che indicano, in tutto o con una loro parte, un rapporto, resteranno prive di corrispondente, e quindi per esempio non designeranno alcunché le congiunzioni, le preposizioni, gli articoli, certi suffissi, etc. Con ciò però sarà impossibile rendersi conto del significato di una proposiozione, in cui di necessità figurano designazioni di rapporti. Inoltre si andrà incontro alla contraddizione di avere delle grafie o fonazioni, cioè un certo materiale grafico o fonico, che è considerato linguistico pur mancando di corrispondente, di significato.

Con questi arresti di tipo conoscitivo è escluso che si possa raggiungere una consapevolezza di che cosa è il linguiggio, di che cosa è una lingua, di che cosa è una proposizione e una parola, ed infine di che cosa è una grammatica. Questo almeno se, per consapevolezza del linguaggio, si intende un rendersi conto delle funzioni che esso puo e deve svolgere come espressione, designazione e communicazione del pensiero. Come analizzare un linguiggio in riferimento a cio che designa se per metà non designa niente ? E poiché oggi il linguaggio rappresenta ancora la principale via di accesso al pensiero, è chiaro che questi arresti di tipo conoscitivo hanno impedito anche una consapevolezza di questo. Pensiero e linguaggio sono rimasti cosi qualcosa di magico, su cui lavorare non con la pazienza dell'analizzatore e con il controllo della collaborazione, ma con l'intuizione dell'artista! E vi è forse di più: che l'analisi, non soltanto arrestata ma anche deviata, ha portato a conclusioni affatto inutilizzabili, se queste devono venire applicate nella costruzione di modelli meccanici o nell'ispezione dell'anatomista e del fisiologo.

Certo, un trattato di linguistica si può sempre scrivere lo stesso. Ma sembrerà già una scoperta se in esso si parlerà della proposizione come di una forma o struttura. Soltanto, sarà impossibile precisare quale; e cosi, poiché ogni cosa può essere vista come una struttura, pera, tavolo o città che sia, la scoperta sarà di ben poco aiuto per chi speri di trovare in quel trattato la chiave di come si traduca "proposizione per proposizione", traduttore sia l'uomo o la macchina.

In queste condizioni di inconsapevolezza linguistica è nata anche la sterile branca di studi che si chiama sintattica, con il suo usuale accoppiamento di logica simbolica.

Dalla situazione conoscitiva si esce con due passi :

a) cogliendo e denunciando l'equivoco da cui origina, e con cio si ridà ad ogni parola e ad ogni espressione la possibilità di avere un corrispondente analizzabile;

b) trovando criteri di analisi adatti per la descrizione delle cose nominate.

La Scuola Operativa Italiana, che qui io rappresento, ha sostituito la posizione conoscitiva con una posizione operativa, che permette appunto ,in quanto non conoscitiva, di cercare per ogni parola ed espressione la controparte designata, ed ha programmaticamente svolto una analisi di questa controparte in termini di operazioni, quello cioè che noi

bambini stavamo facendo quando abbiamo appreso ad adoperare le parole, e quello che poi ripetiamo quando le comprendiamo.

Che una congiunzione, "e", od una disgiunzione, "o", corrispondano ad operazioni si può subito comprendere; ma, come si vedrà meglio in seguito, in operazioni si può vedere facilmente anche un tavolo, un uovo, etc. Basta pensare queste cose osservative come il risultato di una operazione di differenziazione, con cui si separi, per esempio per durezza o colore, il legno dall'aria, et di una operazione di figurazione, con cui si tracci una figura, seguendo la linea di differenziazione.

Con queste analisi operative si troveranno vari tipi di operazioni, ma in ogni caso la controparte del linguaggio riceverà una sua trattazione esauriente ed omogenea.

(Prima di passare ad una descrizione, sia pure sommaria, dei vari tipi di operazioni, ritengo di dover indicare qui alcuni scritti in cui i pochi cenni ora dati di critica al conoscere ricevono un certo sviluppo. Si vedano di S. Ceccato : "Il linguaggio con la Tabella di Ceccatieff", testo italiano e inglese, Ed. Hermann & Cie, Paris 1951; "L'Ecole opérationnelle et la rupture de la tradition cognitive", Mars-Mai, 1952-53, Ed. Librairie A. Colin, Paris; "Contra Dingles, pro Dingler", testo italiano e inglese, Methodos, anno IV, 2953; "Comment ne pas philosopher", Actes du XIème Congrès International de Philosophie", Vol. I, Ed. North-Holland Publishing Company, Amsterdam, 1953; "Le definizioni sviate", Atti del XVI Congresso Nazionale di Filosofia, Ed. Fratelli Bocca, Roma-Milano, 1953; di S. Ceccato e V. Somenzi, "Operazionismo e tecnica operativa"; di G. Vaccarino, "L'origine del conoscitivismo greco", Methodos, anno X, 1958).

Nel tema della presente esposizione, quale sostituzione dei risultati di una analisi in operazioni delle cose nominate al filosofare sul linguaggio, rientra però almeno una scorsa fra queste operazioni e risultati, ed anche fra le applicazioni cui essi dànno luogo.

Ricordo intanto che l'analisi in operazioni, oltre al programma, richiede un criterio di analisi, e che questo non puo venire imposto da alcuna Natura o Realtà conosciute. La scelta del criterio è stata determinata dall'intento di giungere ad "atomi operativi" tali da poter vedere ogni cosa nominata o corrispondente ad uno di questi o ad un loro composto. A questo intento si è poi aggiunto quello di giungere ad atomi operativi riproducivili nel funzionamento di organi di un modello meccanico, già più o meno nei limiti della possibilità costruttive della tecnica attuale.

Quale primo tipo di operazioni indichiamo la differenziazione. Essa ha per risultato dei termini polari, come caldo e freddo, luce e buio, etc., o dei termini fissi, come i vari colori, i vari suoni della scala musicale, etc.

Una difficoltà potrebbe presentarsi alla domanda di quali differenziati siano elementari e quali composti. Il "siano" trae in

inganno. Anche a questo proposito bisogna decidere, perché l'elementarità o meno non risulta da osservazione; e non vale rifarsi alla individualità degli organi, perché fisiologicamente ed anatomicamente questi sono ricavati dalla funzione. La scelta dei differenziati da considerare elementari è stata suggerita anche qui dalla lingua e quindi dalla storia, seguendo partizioni ormai entrate universalmente in uso.

L'ambito della differenziazione si potrebbe accostare infatti a quello della sensasione, (badando però a sottrarre subito questa dalle funzioni ad essa assegnate dalla tradizione conoscitiva).

Un altro tipo di operazioni è la figurazione. Da essa risultano le figure, o forme.

Anche a proposito delle figure si è posto un problema di elementarità et di composizione, ed è dovuta intervenire una decisione. Sono state scelte come elementari tre tipi di linee e cinque tipi di angoli.

Il terzo tipo di operazioni è costituito dalle categorie mentali. Figurano fra queste, per esempio, il soggetto e l'oggetto, il tempo e lo spazio, l'eguale e il differente, l'e, l'o, il ma, il non, la causa e l'effetto, l'elementare el il composto, etc.

Quale categoria elementare è stata scelta qui quella corrispondente al qualcosa.

Naturalmente, la maggior parte delle cose nominate non rappresentano i risultati puri di queste classi di operazioni, bensi loro composti.

Per esempio, se il resistente ed il cedevole si possono considerare semplici differenziati, già per ottenere il duro ed il molle bisogna aggiungere la figurazione ed un confronto fra due figure.

Fra le composizioni più comuni abbiamo gli osservati, con la percezione quando alla differenziazione si aggiunge la figura, e con la rappresentazione quando alla figurazione si aggiunge il differenziato.

Per la comprensione del linguaggio, l'analisi più importante ed i risultati più decisivi riguardano tuttavia il pensiero.

Quando i tre tipi di operazioni già considerati hanno porto i loro risultati, isolati od in composizione fra loro, non abbiamo ancora il pensiero, bensi soltanto i suoi possibili, i suoi futuri contenuti. Affinché essi diano luogo ad un pensiero, affinché con essi si pensi, bisogna dare ad essi un particolare ordine temporale, che è l'ordine caratteristico della correlazione, per cui le cose cosi ordinate sono sempre almeno tre.

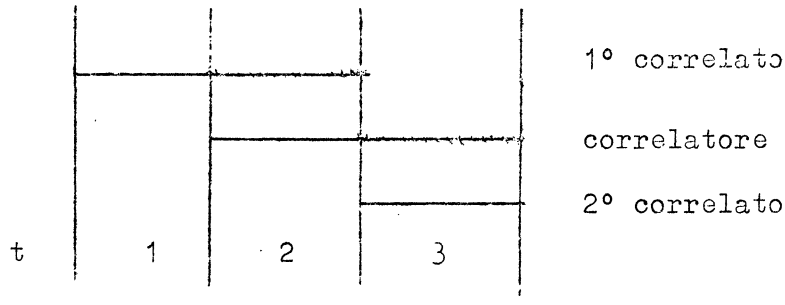
Descrivo la situazione nei suoi termini dinamici :

Una delle tre cose viene costituita in un intervallo di tempo 1, e viene mantenuta in un intervallo di tempo 2. Durante questo intervallo

2 viene costituita un'altra delle tre cose, e viene mantenuta in un intervallo di tempo 3. Durante questo intervallo 3 viene costituita la terza delle tre cose.

Ecco un esempio che illustra e conferma questo ordine temporale correlazionale. Chi legga a fine di pagina la frase "ieri ho mangiato carne e", mantiene cio che corrisponde all' "e", ed attende appunto la terza cosa, diciamo, "pesce". Chi legga invece a fine di pagina "ieri ho mangiato carne", e voltata la pagina trovi "e pesce", deve tornare indietro, appunto con il pensiero, per riprendersi "carne", e cosi poter ordinare le tre cose nella correlazione; altrimenti proprio non capirebbe quello che è scritto. Cosi come anche le note musicali, soltanto concatenate in un certo modo compongono i temi. Per rendersi conto della differenzi fra la cosa isolata e la cosa come contenuto del pensiero si provi a presentarsi alla mente per esempio cio che corrisponde a "cane", una prima volta con la semplice categoria del nome, e la seconda con la categoria del sostantivo o del soggetto. Sarà facile accorgersi che con le due ultime categorie già si appresta una struttura di cui "cane" viene ad essere un elemento, ancor prima di farlo seguire da un aggettivo o da un verbo.

Per maggior chiarezza si veda anche questa rappresentazione dell'ordine temporale caratteristico del pensiero :

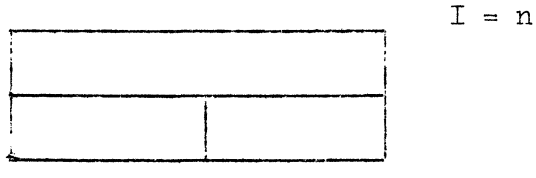


_____ = tempo in cui la cosa è fatta

..... = tempo in cui la cosa è mantenuta

I diversi tempi di presenza delle tre cose nella struttura fanno dell'una, come è indicato, l'elemento correlatore, e delle altre due i correlati, correlato primo e secondo.

Un altro modo di rappresentare la correlazione è il seguente, meno corretto e suggestivo, ma più comodo (e di esso ci serviremo in seguito) :



Con la consapevolezza della correlazionalità del pensiero è stato possibile rendersi conto di quali e quante indicazioni siano necessarie per indicarlo, e quindi di come deva essere fatta una lingua al fine di poter assolvere le sue funzioni. Appare infatti subito chiaro che la designazione di una correlazione richiede almeno cinque indicazioni. Tre sono per indicare le tre particolari cose messe in correlazione, i correlandi, una quarta per indicare quale di esse funge da correlatore - e da questa è possibile ricavare che le altre due cose fungono da correlati -, ed infine una quinta per indicare il posto di uno dei due correlati, se il primo od il secondo - poiché da questa si può ricavare anche il posto dell'altro.

A prima vista si potrebbe supporre che occorrano meno indicazioni per designare una correlazione, in quanto certe cose già fungerebbero di per sé da correlazione, come gli "e", gli "o", etc. Ma le cose non stanno cosi se non quando noi classifichiamo quelle parole come "congiunzioni" o "preposizioni", cioè già dando loro la funzione di correlatori. Infatti si puo benissimo trovare anche espressioni come "e ed sono rapporti" etc., ove quelle cose fungono anche da correlati.

Qualsiasi cosa, cioè, può essere adoperata come correlato, anche se questo è un caso molto raro per certe cose; mentre non vale in contrario : i correlatori sono una classe particolare di categorie mentali, e quindi per esempio un osservato non potrebbe mai prenderne il posto.

Per fornire queste indicazioni le lingue si servono sia del particolare materiale sonoro o grafico, sia dell'ordine di successione delle parole. Si dànno i casi estremi, a questo proposito, per esempio del latino, in cui l'ordine di successione è di significato minimo, e del cinese, ove invece è massimo. In ogni caso, le cinque informazioni vengono distribuite fra due o tre parole. Il nostro esempio di "carne e pesce" mostra una distribuzione in tre parole; il latino che dicesse "caro piscesque" le distribuirebbe in due. Di solito con due parole sono indicate le correlazioni che ricorrono più di frequente, ed allora una delle due contiene, per esempio in un suffisso, la designazione della particolare correlazione in gioco, cioè il correlatore, ed il suo posto di correlato in questa. Tali sono i nostri aggettivi, i nostri verbi personali, etc.

Naturalmente, il pensare non si limita alle singole correlazioni. Per lo più i nostri pensieri sviluppano una intera rete correlazionale, ove le singole correlazioni figurano come correlati o come correlatori di altre. Per esempio, "mangiare carne" è una correlazione, "mangiare pesce" è un'altra, "mangiare carne e mangiare pesce" mostra una rete in cui una correlazione più ampia, con il correlatore "e", contiene come correlati le altre due, cioè è una rete costituita da tre correlazioni.

Per lo meno da vari millenni l'uomo è in grado di svolgere pensieri con reti correlazionali di decine di correlazioni.

Questa rete correlazionale, caratteristica del pensiero, è forse il patrimonio più comune dell'umanità, ed in essa, con differenze certo minori delle eguaglianze, tutti gli uomini si ritrovano. Né ciò deve stupire, perché essa corrisponde ad una costruzione del mondo che deve conservare la massima varietà pur risultando dalla composizione di elementi il più possibile sempre eguali e di rapida fattura.

Il sostantivo-aggettivo ci permette di fissare una cosa arricchendola poi di caratteristiche, il soggeto-verbo ed il verbo-oggetto di fare la storia delle cose, seguendole nel tempo, etc. Sarebbe strano se queste strutture non comparissero presso ogni popolo ed in ogni epoca.

Alcune varietà fra popoli, e quindi fra lingue, si notano nel diverso modo di raggruppare le singole operazioni, che forniscono i contenuti del pensiero, in unità correlazionali. Per esempio, ove in una lingua si trova soltanto il "camminare velocemente", in un'altra si trova il "correre", ed una terza offre entrambe le possibilità; ove una lingua mi permette di indicare con una parola il movimento tenendo conto del posto di chi parla (come quando noi adoperiamo l'"andare" ed il "venire") un'altra deve riccorrere al'avverbio "qui" e "là" (in russo, etc.)

Altre varietà riguardano il modo di presentare insieme le cose. La espressione in questi casi ha da un punto di vista oggettivo la stessa portata comunicativa, ma differenzia il comportamento del parlante, sia come suo atteggiamento generico, sia come storia della particolare composizione. Inoltre, questa varietà, proprio perché legata ad elementi che esulano dalla sfera più strettamente dell'osservazione, assumono una particolare importanza come rivelatori delle varie psicologie dei popoli. Per esempio, una tavola e dei colori si possono mettere in rapporto in diversi modi. L'italiono ne conta almeno sei : tavola colorata, tavola con colori, tavola a colori tavola di colori (per esempio, vivaci), tavola dai colori (vivaci) tavola in colori (vivaci); e potremmo aggiungere, con l'apposizione, "tavola, i colori vivaci,...". Me un'altra lingua puo usare soltanto una di queste possibilità, e proprio quella e non un 'altra (credo che l'italiano sia a questo proposito una delle lingue più ricche).

Si è già parlato della ricchezza inesauribile del pensiero come risultato della combinazione di un numero limitato di elementi. Anzitutto, isolando i correlatori si è visto che questi non superano il centinaio. Era dunque possibile cominciare ad individuare le correlazioni servendosi di questi, assunti appunto quali individui. Per i correlati, più numerosi
questi, assunti appunto quali individui. Per i correlati, più numerosi, la possibilità di fissarli in un ordine, si poteva ottenere considerandoli in classi.

In tal modo si sono ottenute varie centinaia di correlazioni, la cui originalità è in funzione dei tre correlandi, l'uno come individue e gli altri due come classi.

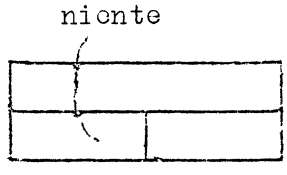
Queste "semiartificiale" unità del pensiero sono ora alla base dei quattro progetti attorno ai quali si articola il lavoro del Centro di Cibernetica et di Attività Linguistiche dell'Università di Milano.

Per numero di persone impegnate viene in primo luogo la traduzione meccanica.

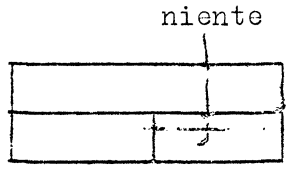
La realizzazione di questo progetto è sostenuta dal Governo Americano, con un contratto che ha avuto inizio nel febbraio del 1959 e si estenderà sino al maggio del 1962. Contempla la traduzione principalmente dal russo all'inglese, per un totale di circa 50.000 voci russe, cioè di un dizionario previsto sufficiente per la traduzione anche di un giornale quotidiano. Nei prossimi mesi, un aiuto dell'Euratom permetterà di aggiungere le uscite dal russo sia in tedesco che in italiano (ma già in questa prima fase le analisi linguistiche hanno sempre tenuto presenti queste lingue, il francese ed il latino, ed in un primo momento il confronto è avvenuto anche con il cinese).

Cercherò di mostrare con un esempio molto elementare come funziona il nostro procedimento di analisi del pensiero e del linguaggio in vista della traduzione meccanica.

Sia la proposizione "Niente egli fece". Dalla parola "niente", per quanto riguarda la funzione correlazionale, è possibile sapere che esso a) non può indicare un elemento correlatore, b) può indicare il correlato primo a secundo di un certo numero di correlazioni. Questo sapere discende dalla parola "niente" anche del tutto isolata, per cui risulta segnato in una matrice che accompagna la parola e che viene preparata preliminarmente. Ricorrendo ad uno dei nostri modi di raffigurare la situazione, scriveremo pertanto sia lla forma correlazionale aperta



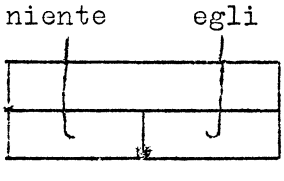
di cui "niente" occupa il posto di primo correlato sia la forma correlazionale.

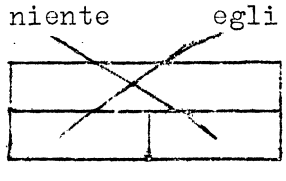


di cui "niente" occupa il posto di secondo correlato.

Nel testo fi ingresso si incontra ora la parola "egli". Anche questa non può indicare un correlatore, ma può indicare il correlato primo o secondo di un certo numero di correlazioni, ed avrà pertanto la stessa rappresentazione di "niente".

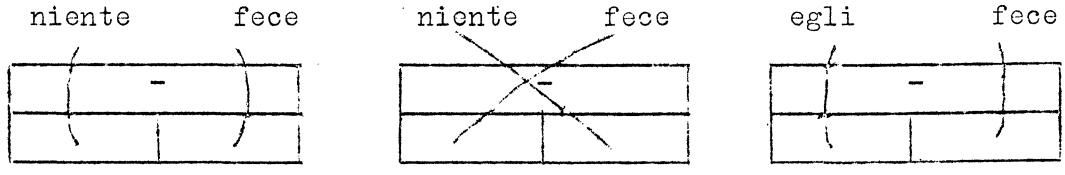
Non appena entrata la seconda parola, si deve cercare di accoppiarle costituendo una correlazione. Da un punto di visto di apertura correlazionale questo sarebbe possibile, dando anzi luogo a due alternative :



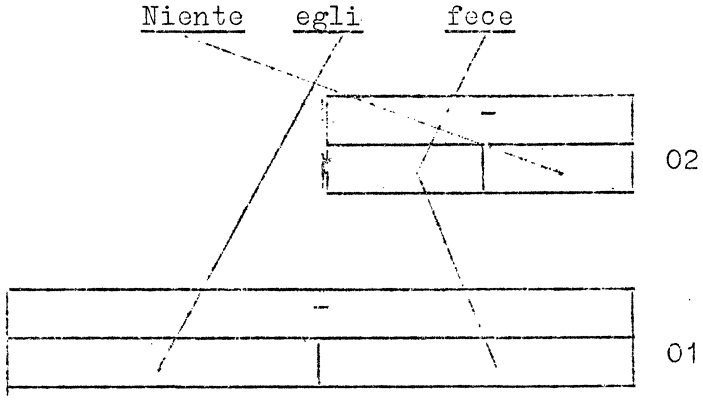


ma entrambe vanno scartate per un insieme di regole che suonano pressapoco cosi - L'incontro di due correlati senza correlatore esplicito, cioè indicato da una parola isolata, può generare certe correlazioni, circa una dozzina, a patto però che i due termini appartengano a certe classi, per esempio l'uno appartenga alla classe dei nomi usabili come sostantivi e l'altro a quella degli aggettivi, etc. Ma nessuno di questi casi è presente con l'incontro di "niente" con "egli".

Per poter effettuare una correlazione bisogna dunque attendere l'ingresso di un'altra parola, "fece". "Fece", come le altre due parole, puo indicare il correlato primo o secondo di un certo numero di correlazioni, e non isolatamente un correlatore. Ma all'accoppiarla con le altre due parole, questa volta le cose vanno diversamente, perché "fece" si accoppia con "niente" sia assumendole come un soggetto sia assumendolo come un oggetto, "niente fece" e "fece niente", e si accopia con "egli", assumendolo pero soltanto come un soggetto, "egli fece". Ecco le varie rappresentazioni :

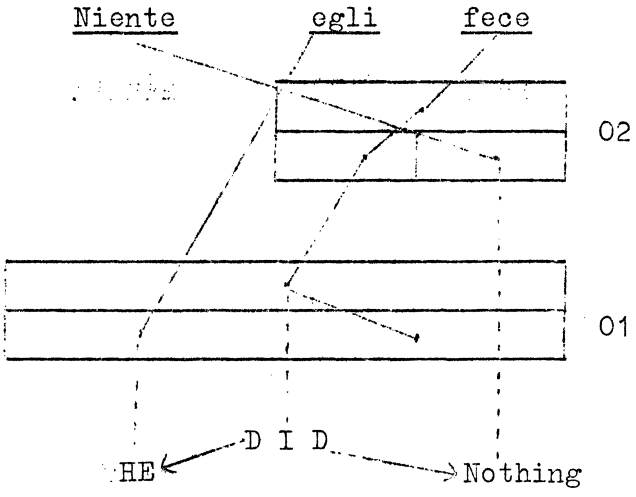


Si tratta ora di accoppiare le correlazioni già costituite, formando una rete correlazionale, cio che è possibile per sovrapposizione, cioè utilizzando quale perno la parola comune, "fece", che figura in due correlazioni diverse :



Una volta costituite le singole correlazioni esse, come ho detto, ricevono un numero indice che è in funzione dei tre elementi che le compcigono. A titolo illustrativo abbiamo scelto qui i numeri 01 per "egli fece" e 02 per "fece niente".

Questi numeri servono ora per l'uscita in un'altra lingua, essendo riferite ad essi le regole semantiche proprie di questa. Per esempio l'uscita inglese sarà :



Nonostante la fecondità del procedimento analitico, che assicura, se ben applicato, di esaurire tutte le indicazioni contenute nelle espressioni, cioè i due ordini di informazioni, riguardanti le particolari cose messe in correlazione e la funzione correlazionale da queste assolta, devo pero dichiarare che nessuna analisi linguistica potrebbe far apparire in certi casi la nostra comprensione di un discorso. Questo avviene perché le lingue sono fabbricate contando sulla nostra facoltà rappresentativa delle cose nominate e su una integrazione che scaturisce da tutte le esperienze vissute e dal sapere che in esse è racchiuso. L'espressione linguistica, cioè, non cade su una tabula rasa, ma incontra situazione già costituite che la integrano e sulle quali può giustamente contare chi si esprime.

Sia per esempio la frase : "sulla sedia a sdraio lacera ed abbattuta sedeva la giovane donna". A chi riferire il "lacera ed abbattuta" ? alla sedia od alla donna ? Questa frase può colpire, perché noi stessi per un momento forse saremmo incerti; mentre non lo saremmo se fosse "sulla sedia a sdraio piangente sedeva la giovane donna". Ma è chiaro che da un punto di vista strettamente grammaticale anche la seconda frase permette più interpretazioni, legando "piangente" sia alla sedia ed alla donna, e sia addirittura a "sdraio".

Noi uomini decidiamo questi casi, più meno immediatamente, con la rappresentazione della situazione descritta.

Per una macchina che manchi di facoltà rappresentativa si deve cercare di sostituirla con una serie di classificazioni che articoli ogni cosa nominata negli elementi che ne costituiscono l'inconfondibile originalità.

Questo complementare tipo di analisi deve intervenire ogniqualvolta l'analisi correlazionale lascia sussistere una alternativa di interpretazione, o cosi ogniqualvolta le parole della lingua di ingresso si rivelino polivoche in rapporto a quelle della lingua di uscita.

Illustriamo con un esempio queste classificazioni in una applicazione. Noi italiani, e cosi i francesi, gli inglesi, etc., indichiamo distintamente un "aldilà" ed un "dietro", a seconda che due cose siano separate fra loro semplicemente da un intervallo spaziale o da una

terza cosa per cui l'una risulti nascosto mettendosi al posto dell'altra. Ma il russo non ha che una parola per le due situazioni, "nad", e lascia che la differenza nasca dalla rappresentazione delle cose messe in quei rapporti, a seconda che l'una, piatta, orizzontale, lasci scorgere l'altra, o verticale ne occulti la vista. Cosi nel caso del fiume l'uscita sarà "aldilà", ma nel caso del muro, o del cespuglio, etc. sarà "dietro".

Ecco allora la necessità di classificare tutte le cose anche per questi loro eventuali aspetti. Ma questo non è ancora sufficiente. Per esempio le Alpi si inframmettono fra due cose certamente con la loro verticalità; tuttavia di una casa si direbbe che essa si trova non "dietro", ma "aldilà" delle Alpi, perché, anche se si togliessero la casa non comparirebbe lo stesso, data la loro dimensione anche in senso orizzontale. La classificazione dovrà dunque tener conto anche di un aspetto dimensionale globale. E cosi via.

Al progetto per la traduzione meccanica è legato un progetto per il riassunto meccanico. Questo progetto è al suo primo stadio e sinora non è impegnato in alcun contratto.

Esso si basa sullo sviluppo di un procedimento che già talvolta deve venire applicato nella traduzione meccanica, quando la rete correlazionale risultante dal testo di ingresso deve venir modificata al fine di trovare la sua espressione nella lingua di uscita. Ma mentre nella traduzione si cerca che la rete correlazionale rimanga la stessa, cioè si cerca di avere un'unica rete correlazionale, nel riassunto è invece programmata la sua modificazione; e per far questo si enuncia un complesso di regole, per cui si ripeta quanto fa uno di noi quando riassume.

Si tratta, naturalmente, di un riassunto vero e proprio, e non di un "abstracting" che individui di un testo le parole chiave, per esempio di un riassunto al 30%, o al 50%, o al 70%.

Come già per la traduzione, anche ora il linguaggio viene esaminato sino ad ottenerne la rete correlazionale corrispondente, ed è su questa, quando ogni dubbio interpretativo del testo di ingresso è stato eliminato, che si opera, servedoci dei numeri indice che caratterizzano le singole correlazioni ed i pezzi di rete correlazionale da esse costituiti.

Un accenno ai due principali ordini di regole puo illustrare come sia diretta la modificazione della rete correlazionale.

Il primo comprende regole che fanno riferimento direttamente all'espressione linguistica : a) sia perché si avvalgono della forma delle espressioni di ingresso, b) sia perché fissano la forma da usare in ogni caso per quelle di uscita. Una regola di tipo a) è per esempio quella che, quando il testo di ingresso mostri due proposizioni principali di cui una avversativa, suggerisce di mantenere solatanto la avversativa (per cui "le rose sono belle, ma sfioriscono" diviene "le rose sfioriscono"),; un'altra regola di questo tipo impone di sopprimere le specificazioni che non siano ulteriormente richiamate (per cui "un fazzoletto di lino di batista", se lino e batista non ricompaiono più diventa "un fazzoletto" e basta; una regola di tipo b) è per esempio quella generalissima che chiede di preferire per l'uscita, qualunque sia l'ingresso, la forma soggetto, verbo e predicato.

Il secondo ordine di regole riguarda le cose nominate ed i rapporti fra queste dovuti al loro contenuto. A questo psoposito le due regole forse più importanti prescrivono, l'una di sostituire le specie con il genere, e l'altra di eliminare l'esplicito ogniqualvolta possa figurare implicito. Per esempio, "una campagna ricca di meli, peri, ciliegi, castagni, susini" diventerà "una campagna ricca di alberi da frutta"; "il popolo di navigatori che commerciava con la sua flotta sui mari, etc." diventerà "il popolo che commerciava sui mari".

Come si vede, qui il sapere, compare in primo piano e deve essere inserito nella macchina attraverso tante sfere di nozioni affinché sia possibile supplire alla mancanza di facoltà rappresentativa, che permetterebbe di ricavare sens'altro cio che è eguale (genere) nel differente (specie), o cio che è implicito nell'esplicito (nell'esempio, la navigazione e le navi nel commercio per mare).

Il terzo progetto si trova in uno stadio piuttosto sviluppato, tenendo conto dell'enorme complessità delle operazioni in gioco, e concerne la costruzione di un modello meccanico che svolga le nostre operazioni di osservazione e di categorizzazione mentale e le accompagni verbalmente, cioè di un modello che si comporti come un cronista, sia pure in miniatura.

Gli studi per questo progetto sono stati condotti in seno al Centro di Cibernetica e di Attività Liguistiche ed hanno già permesso di costruire quattro anni fa un primo modello meccanico di operazioni mentali (cfr., di E. Maretti, "Adamo II", Civiltà delle Macchine, n. 3, 1956). La realizzazione del cronista meccanico inizierà nei prossimi mesi con l'aiuto dell'Euratom, del Governo Italiano e della IBM italiana.

Sarà probabilmente la continuazione di questi studi che permetterà di superare nella traduzione e nel riassunto le varie difficoltà incontrate pretendendo da una macchina lo svolgimento di una attività che nell'uomo avviene attraverso la rappresentazione, mancante invece nei calcolatori.

Per il momento la costruzione deve essere mantenuta in un ambito operativo molto stretto ed ha intenti più teorici e dimostrativi che pratici. Si cerca anche di non dover superare difficoltà di tecnica costruttiva, per esempio nella differenziazione ottica e nel tracciare delle figure a distanza, che esulano dagli scopi del progetto.

La memoria permanente del modello conterrà una settantina di cose, fra oggetti familiari, come pere, mele, qualche tipo di stoviglie, alcuni colori, ed una trentina di categorie mentali fra le più usate; il dizionario disporrà di un centinaio di parole, con cui formulare la descrizione di ciò che al modello si mostra.

Questo progetto si differenzia dai precedenti in primo luogo perché i dati di partenza non sono qui linguistici, ma percettivi, ed è da questi che si deve passare al linguaggio.

Ciò comporta organi del tutto nuovi, a parte l'apparecchiatura osservativa. Si deve porre una dipendenza fra le operazioni di osservazione e quelle mentali, ed in genere fra tutti i tipi di operazioni che il modello è in grado di svolgere, cio che comporta uno studio estremamente particolareggiato di carattere psicologico. Questa dipendenza fra operazioni diventa una vera e propria soglia di coscienza quando il progressivo svolgersi del pensiero nella macchina determina quali risultati percettivi devono essere accettati e quali scartati come suoi contenuti.

Illustro con un esempio una dipendenza del mentale dal percettivo.

Il modello debba descrivere ciò che vede sul tavolo esprimendosi con la frase "un bicchiere e una bottiglia" nella situazione in cui noi ci esprimeremmo cosi. Tralasciamo di occuparci qui delle operazioni cui corrisponde l'uso dell'articolo in determinativo, per la complessità dell'analisi, e limitiamoci all'"e" con cui sono congiunte le due cose percepite. Anzitutto è chiaro che con le sole operazioni di percezione, cioè la differenziazione che in questo caso è ottica e con le figure tracciate non si potrebbe mai avere quale risultato qualcosa che corrisponde all'"e".Quando e come dunque entra in gioco questo "e"? Noi lo inseriamo in una situazione percettiva quando : a) l'esplorazione della situazione è stata continua, cioè non è mai stata interrotta e poi ripresa (altrimenti diremmo "anche"); b) le cose percepite sono due sin dall'inizio, cioè non sono due come successiva divisione di un unico percepito (altrimenti diremmo "con"); nel percepire le cose hanno funzionato gli stessi organi (altrimenti tenderemmo a lasciarle staccate, senza rapporto fra loro); le due cose sono differenti fra loro (se fossero eguali adopereremmo il plurale). Le condizioni ora indicate possono non essere sufficienti, ma bastano per dare un 'idea delle regolarità psicologiche da studiare, affinchè il modello possa formulare anche le espressioni più elementari.

In tema di filosofia, vale la pena di richiamare a questo punto come se si vuole conservare il mentale come antimeccanico e magico, questa magia non comincia con le cause finali e la probabilità, ma proprio con le cose tanto comuni da aver fatto pensare al pigro realista che si nasconde in ognuno di noi che per esempio il singolare ed il plurale, od un "e" ed un "con" si vedano con gli occhi.

L'esame delle dipendenza mostra come il modello debba avere fra i suoi organi anche degli elaboratori di dati, per esempio dei confrontatori delle operazioni eseguite dei risultati ottenuti, dei contatori, degli ordinatori, etc.

L'esempio dell'"e" può servire anche per accennare alla funzione della soglia di coscienza. Si è visto che i due correlati del correlatore "e" devono rispondere ad alcuni requisiti che li pongono un dipendenza fra loro; e questo vale per la maggior parte dei correlatori. Se, quindi, nel pensiero è già entrato come contenuto il primo dei due, e nel caso di uno svolgimento spontaneo è la macchina stessa che ha messo in gioco il correlatore, il secondo correlato sarà accettato soltanto se risponde a certi quel quisiti : qualsiasi altra cosa comparisse nell'orizzonte percettivo della macchina non potrebbe entrare nella correlazione

che cosi si è cominciata a costituire, e con ciò resterà nella macchina in forma latente in una sua memoria di transito.

Al tri organi caratteristici di questo modello sono gli eccitatori ed inibitori che stabiliscono i passaggi fra le varie matrici delle cose depositate nella memoria permanente.

Per esempio, ogni espressione negativa od avversativa, come "non è una mela", "ma è piccola", e simili, indica una differenza risultante dal confronto fra qualcosa di presente a noi, e quindi alla macchina, e qualcosa che ci si rappresenti. Questa rappresentazione deve quindi venire sollecitata dalla macchina stessa. Se questo basta poi affinché si abbia il "non", per il "ma" la differenza, come già si ebbe occasione di accennare, la differenza risultante dal confronto deve trovarsi già semplicemente con la cosa presente a noi, ma con una seconda cosa da noi associata a quella.

Il quarto ed ultimo progetto in cui trovare applicazione la consapevolezza raggiunta con le nostre analisi del pensiero e del linguaggio è il più ambizioso. Esso rappresenta un po' il contrapposto del progetto per il riassunto meccanico, in quanto ora si intende invece ottenere lo svolgimento meccanico di un tema, di un soggetto, cosi come modestamente avviene nelle scuole o come più ampuamente si propone ogni artista.

Anche in questo caso è però sempre la possibilità di disporre dei risultati delle analisi in operazioni che assicura la realizzabilità in principio della macchina. Nella macchina che sta operando, ogni cosa compare come una costellazione, e gli elementi comuni fra le varie costellazioni diverse segnano una rete di strade. Per quanto strano possa sembrare a prima vista, la maggiore difficoltà sarà qui, non di insegnare alla macchina come andare avanti, ma proprio il contrario, come limitare le possibili alternative che ad essa si presentano.

Nel corso di questa esposizione forse mi sono dimenticato che lo scopo sarebbe dovuto essere principalmente, od almeno secondariamente, filosofico. Riassumerò in due parole la presa di posizione iniziale a questo proposito. Il filosofare, di necessita conoscitivo, se non si vuole cambiare il dizionario, arresta la nostra consapevolezza dell'operare umanoe attribuisce ad una Natura o Realtà date una volta per tutte ciò che è possibile vedere come opera della nostra storia, dei padri e individuale, e comea tale ripetere e far ripetere da una macchina. Chi estrometta questo filosofare si riapre la strada per studiare l'uomo, una strada che non può incontrare se non i limiti di ogni indagine tecnica, limiti di volta in volta da superare con la pazienza.

Senonché, come ebbe a dire Eberhard Rogger, "Filosofare e difficile, ma non filosofare è più difficile ancora".

ELEMENTS DE LINGUISTIQUE MATHEMATIQUE (1)

par P. BRAFFORT *

SOMMAIRE

On montre comment les nécessités d'élaborer des modèles formels se sont imposées peu à peu en linguistique. On présente ensuite un court exposé des éléments mathématiques nécessaires à cette formulation, puis on examine quelques types de formalisation qui ont été beaucoup développés depuis quelques années. On présente alors une critique de la linguistique mathématique en précisant les limitations et l'on s'efforce de voir comment ces limitations peuvent, dans certaines conditions, être surmontées, grâce aux nouveaux modèles actuellement en cours de développement.

1. INTRODUCTION

1. 1. Pour voir comment on a été amené à formaliser la linguistique, quels sont les processus internes et externes qui ont amené les linguistes à se tourner vers l'outil mathématique, il faut se reporter à la linguistique d'il y a 50 ans. Cette linguistique était encore en grande partie expérimentale, c'est-à-dire qu'on y décrivait des langages en établissant des lexiques en étudiant les règles de grammaire, exprimées comme dans les grammaires que nous avonseuesentre les mains lorsque nous étions enfants. On utilisait donc des notions qui sont déjà des notions formelles, mais à un niveau très élémentaire ; les notions de parties de discours : noms, adverbes, conjonctions, prépositions etc... Et puis, le progrès des connaissances linguistiques aidant, on a fait de la linguistique comparée, c'est-à-dire qu'on a mis en face les uns des autres les vocabulaires et les structures grammaticales des langages de plus en plus éloignés les uns des autres. On est sorti du cadre purement indo-européen pour étudier les langues de l'Afrique, les langues des Indiens de l'Amérique du Nord, les langues asiatiques et on s'est aperçu qu'il n'était plus possible de se servir des mêmes catégories grammaticales. Pour comparer les langues, pour en faire par exemple une classification, pour essayer du point de vue historique de retrouver la genèse des différents langages, il fallait dégager des classes, des groupes de langues et on a utilisé des notions

* avec la collaboration de J. LARISSE, Y. LECERF et A. LEROY

(1) Contrairement aux autres, le présent texte tient compte de développements apportés depuis la date de l'enseignement.

formelles appartenant à un niveau d'abstraction déjà plus élevé, telles que les notions de langues analytiques, de langues synthétiques, agglutinantes, polysynthétiques, etc.. Mais on a montré (notamment E. SAPIR) qu'aucune de ces classifications n'était satisfaisante, et c'est cet échec des méthodes de classifications traditionnelles dans l'étude comparée des langages qui a été un des éléments moteurs pour l'élaboration d'une linguistique formalisée.

1. 2. Il est classique de dire que la linguistique générale commence avec F. de SAUSSURE. C'est lui en tous cas qui a indiqué le plus clairement la possibilité d'une science autonome du signe (sémiologie). C'est lui aussi qui a distingué les diverses branches de la linguistique et a permis leur essor systématique.
En particulier, la phonologie qui étudie les systèmes de signes phonétiques composant les divers langages s'est développée rapidement comme une science expérimentale basée sur une méthodologie systématique (et abstraite) très perfectionnée (notamment grâce aux travaux du Cercle de Prague (N.S. TRUBETZKOY, R. JAKOBSON). L'étude des phonèmes, délivrée de tout lien "sémantique" (au moins en apparence) conduisait naturellement vers un modèle abstrait, et c'est ce qui a donné la glossématique du Cercle de Copenhague (L. HJEMSLEV). Simultanément BLOOMFIELD, aux Etat-Unis, créait une école de linguistique structurale dont l'aboutissement se trouve dans l'oeuvre de HARRIS.
L'ensemble de ces travaux possède un commun dénominateur qui est le désir de mettre en évidence des structures abstraites dans des systèmes de signes et ceci d'une façon indépendante de la signification de ces signes. Cette tendance (dont le représentant le plus connu en France est A. MARTINET) est tout naturellement conduite vers une formalisation de plus en plus poussée, donc vers une linguistique mathématique.

1. 3. La tendance de la linguistique générale à s'orienter vers les systèmes formels (illustrée d'une façon très claire par l'école de Copenhague, notamment par V. BRØNDAL et S. JOHANSEN), s'accompagne d'une tendance réciproque des mathématiques les plus générales, c'est-à-dire de la métamathématique, à aborder à son tour des problèmes linguistiques. On sait, en effet, que vers la fin du 19ème siècle et de plus en plus vite à partir des années 1900, les mathématiciens se sont posés des problèmes d'existence et de décision. Ils ont été ainsi amenés à voir, que ce qu'ils considéraient comme étant assuré, comme étant pratiquement l'expression spontanée du raisonnement humain, pouvait aussi être sujet à question, qu'il fallait donc formaliser non seulement les objets sur

lesquels on raisonnait, mais les raisonnements eux-mêmes qui pouvaient devenir l'objet d'une enquête formelle. C'est ce qui a donné lieu au développement de la logique mathématique et des disciplines voisines.
Une théorie mathématique, c'est essentiellement un ensemble de symboles constituant ce qu'on pourrait appeler un vocabulaire ; avec les mots de ce vocabulaire on construit des phrases et il existe des règles de formation qui nous disent si ces phrases ont un sens dans la théorie mathématique en question ou si elles n'en n'ont pas. Par exemple, si l'on fait de l'algèbre, on dispose d'un vocabulaire qui comprend les lettres A, B, C, etc..., les symboles d'addition, de multiplication, les parenthèses, etc..., et on dit que la phrase A+B a un sens, que la phrase (A+B)x C a un sens. Par contre la phrase A + (n'a pas de sens, la phrase A x + n'a pas de sens. Donc un certain nombre de phrases possibles seulement ont un sens dans les règles de formation. D'autre part, on a des axiomes qui sont des phrases ayant un sens que l'on pose comme point de départ. Par exemple : A + B = B + A définit une algèbre commutative. Et enfin, on a des règles de déduction qui à partir des axiomes permettent de découvrir les théorèmes appartenant à la théorie. Il y a donc un certain parallélisme entre cette formalisation et certains aspects du langage. Les règles de formation et les règles de déduction sont les équivalents de la syntaxe, puisque la syntaxe nous permet de construire des phrases qui sont correctes, de même que les règles de formation nous permettent de construire des phrases mathématiques qui sont correctes. R. CARNAP a montré le premier l'analogie entre les problèmes syntaxiques des mathématiques et les problèmes syntaxiques du langage naturel.
Mais ce qui est le plus intéressant, c'est que la construction métamathématique elle-même nous oblige à sortir d'une conception purement formaliste d'un jeu de signes dépourvus de sens. Car le formalisme logico-mathématique lui-même pose des problèmes de signification et certaines "antinomies" de la théorie des ensembles couramment qualifiées de sémantiques. Si par conséquent la linguistique recherche, notamment pour poser ses problèmes de syntaxe, un outil combinatoire, du côté des systèmes formels, la théorie des systèmes formels pose à son tour des problèmes linguistiques de fond.

1. 4. Le développement, au sein de la linguistique traditionnelle, d'une linguistique structurale ; le développement, au sein de la logique mathématique, d'une syntaxe et d'une sémantique formelle, ont permis la naissance d'une linguistique mathématique. Mais il est clair que l'élément déterminant pour la promotion d'une telle discipline, c'est le développement des techniques de mécanisation des informations tout d'abord numériques, puis non-numériques et c'est finalement le dernier facteur qui est décisif. En effet, traiter le langage comme

une information, c'est-à-dire enregistrer, les textes, les mots, sur un support permanent, et construire un automate qui traite ces informations, pose évidemment des problèmes de formalisation. C'est ainsi qu'en calcul numérique, il ne suffit pas d'avoir une équation différentielle formalisée pour qu'une machine soit capable de donner la solution, il faut encore décrire la méthode d'approximation et il faut établir le programme complet des différentes étapes que la machine va réaliser pour résoudre l'équation différentielle par approximation. Il faut donc formaliser le processus du calcul numérique lui-même. De même si on veut transformer des informations linguistiques, il faut formaliser tous les processus intermédiaires qui ont lieu dans la machine. La machine ne fait qu'exécuter des instructions complètement énoncées, complètement élaborées. Il est donc certain que nous avons ici un moteur encore plus puissant pour le développement d'une linguistique mathématique.

Mais il est particulièrement intéressant de constater que, tout comme celui de la linguistique, le développement de la métamathématique conduit également à l'étude des automates. Jusqu'ici ces automates avaient été purement théoriques : machine de TURING, réseaux de neurones de Mac CULLOCH et PITTS, etc... Plus récemment, avec GELERNTER et HAO WANG, les algorithmes ont été transportés aussi sur calculateur réel. On voit se dessiner ainsi une discipline nouvelle, celle du traitement automatique de l'information non numérique, dont la linguistique mathématique pourrait bien être la base formelle.

2. OUTILLAGE MATHEMATIQUE:

2. 1. Dans l'optique d'une linguistique mathématique se situant au point de rencontre de la linguistique structurale et de la métamathématique, tout l'outillage de la logistique devient nécessaire. Mais une telle optique éclaire d'avantage les prospections du développement de notre discipline que son état actuel. Nous nous bornerons donc ici à décrire l'outillage nécessaire à la compréhension des textes effectivement disponibles dans la littérature. Cet outillage est essentiellement celui de l'algèbre moderne. Mais pour acquérir des rudiments d'algèbre, il est nécessaire d'avoir présente à l'esprit la notion d'ensemble.

2. 2. Un ensemble sera pour nous une collection d'objets quelconques que l'on peut désigner par n'importe quel symbole graphique, qui ont entre eux un certain nombre de relations qui expriment ainsi la structure du système. On se rend bien compte que, par exemple, des atomes dans un cristal présentent une certaine structure et que les assistants à cette conférence en présentent une autre. Aussi les mots dans une phrase présentent une certaine structure, mais les mathématiciens ont développé une notion plus rigoureuse de la notion de la structure. Pour l'expliciter, nous nous livrerons à la construction d'une échelle dont la base est un certain ensemble E. Si je dispose d'un ensemble E, je peux construire, en partant de cet ensemble, des couples formés de deux éléments appartenant déjà à l'ensemble E. Donc je fais le couple A,B, le couple B,C, le couple A,C,....Et s'il y a d'autres éléments, je forme tous les autres couples de la sorte. L'ensemble de ces couples c'est lui-même un ensemble, on le désignera par l'expression E x E, c'est le produit cartésien de deux ensembles. Maintenant si je prends un certain nombre d'éléments d'un ensemble et que je les mets dans une même boîte, par exemple: (A,B) (A,B,C) (B,C) (D,C,D) (A,C,D) etc., j'ai défini des parties de l'ensemble. Si je considère toutes ces parties, faites à l'aide d'éléments appartenant à l'ensemble E, j'ai défini un nouvel ensemble qu'on appelle l'ensemble des parties de l'ensemble E. Je le désigne par $\mathfrak{P}(E)$. A partir de l'ensemble E et des deux opérations "produit cartésien" et "ensemble des parties", je peux construire toute une échelle de nouveaux êtres mathématiques Je peux construire par exemple : E x E x E, ou bien $\mathfrak{P}(E) \times E$, ou bien $\mathfrak{P}(\mathfrak{P}(C,E))$, etc. Déterminer un élément dans un ensemble appartenant à l'échelle, c'est définir la structure de l'ensemble E. Une structure algébrique, en particulier, c'est la donnée d'une opération entre éléments de l'ensemble E, c'est-à-dire une relation telle que A + B = C.

Cela veut dire qu'à tout couple d'éléments A et B on associe un élément C.

Tous les trois appartiennent à E. Donc un couple d'éléments de E appartient à E x E. Le résultat C appartient également à E. Par conséquent il s'agit d'une application de E x E dans E, une projection de E x E dans E. Cette projection est un élément de l'ensemble $\mathcal{P}[(E \times E) \times E]$. Ceci est bien un ensemble que l'on construit à partir de l'ensemble de base E par les opérations que j'ai définies plus haut: On fait d'abord le produit de deux ensembles multiplié par un troisième et on prend l'ensemble des parties. Définir la structure algébrique comme étant un élément de cet ensemble, c'est bien procéder comme je l' indiqué tout à l'heure pour le choix d'une structure. Pour définir une structure d'ordre, une structure topologique, on procéderait de même, mais l'ensemble au sein duquel on choisirait un élément distingué serait un autre ensemble appartenant à la même échelle de base E. On a donc la possibilité de définir axiomatiquement des structures mathématiques en partant d'ensembles dont on ne sait absolument rien, c'est ce qu'on appelle la définition de structures multivalentes, par opposition aux structures univalentes qui sont celles que l'on définit à partir de notions de nombres entiers qui seraient considérés comme intuitivement connus.

2. 3. En fin de compte nous aurons besoin essentiellement des notions et notations suivantes :

A. Définition :

Un ensemble E est une collection d'objets, de nature quelconque (points du plan, nombres, fonctions, mots de l'alphabet etc..) qui sont par définition les éléments de E.

On désigne habituellement les ensembles par des lettres latines majuscules = A,E, etc.., les éléments par des lettres minuscules = a, e..,si ce sont des éléments déterminés et x,y,z, si ce sont des éléments variables ou arguments.

B. Signes conventionnels :

Symboles:	Signification:	Définition:
$e \in E$	l'élément e appartient à l'ensemble E	
$e \notin E$	l'élément e n'appartient pas à l'ensemble E	
$A \subset E$	l'ensemble A est inclus dans l'ensemble E	Tous les éléments de A sont éléments de E.

$A \not\subset E$	L'ensemble A n'est pas inclus dans E	
$A \cup E$	Réunion des ensembles A et E	Ensemble des éléments appartenant soit à E soit à A.
$A \cap E$	Intersection des ensembles A et E	Ensemble des éléments appartenant à E et à A.
$\emptyset$	Ensemble vide	Ensemble qui n'a aucun élément.
$A + E$	Somme des ensembles A et E	Réunion des ensembles disjoints A et E.
C_E (A)	Complémentaire de A dans E	Ensemble des éléments de E qui n'appartiennent pas à A.
$E\{x, P(x)\}$	Ensemble des nombres x possédant la propriété P (x).	

Quantificateurs:

$(\exists\ x)\ P$ Il existe un $x \in X$ tel que x possède la propriété P.

$(\forall\ x)\ P$ Pour tout $x \in X$, x possède la propriété P.

C) Ensemble des parties de E:

On se donne un ensemble E et une propriété d'un élément de E; ceux des éléments de E qui possèdent cette propriété forment un sous-ensemble ou partie de E.

Ex: 1) la propriété $x = x$ appartient à tous les éléments de E. la partie que définit cette propriété est l'ensemble E lui-même; on dit encore que c'est la partie pleine de E.

2) la propriété $x \neq x$, n'appartient à aucun élément de E, la partie définie par cette propriété est la partie vide $\emptyset$.

Par définition, l'ensemble des parties de E que l'on symbolise par $\mathcal{P}(E)$ est l'ensemble dont les éléments sont les parties de E.

Ex: Considérons l'ensemble $X = \{1,2,3\}$, une partie de X est par exemple $\{1,2\}$, l'ensemble des parties de X sera:

$$\mathcal{P}(X) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{2,3\}, \{3,1\}, \{1,2,3\}\}$$

Recouvrement:

Une famille $(X_i)_{i \in I}$ (I est l'ensemble des indices) constitue un recouvrement de A si

$$A \subset \bigcup_{i \in I} X_i$$

Partition:

On appelle partition de E un recouvrement $(X_\tau)_{\tau \in I}$ de E tel que :

1. $X_\tau \neq \emptyset$ quel que soit $\tau \in I$
2. $X_i \cap X_j = \emptyset$ pour tout $i \neq j$ $i \in I$ $j \in I$

Autrement dit, les parties X_τ sont disjointes et un élément de E appartient à une partie de E et une seule.

Exemple : $X = \{1,2,3,\}$ une partition de X est l'ensemble formé des deux éléments $X_1 = \{1\}$ $X_2 = \{2,3\}$

D) Relation d'équivalence:

C'est une relation binaire (c'est-à-dire entre deux éléments) que l'on note $\equiv$ et qui jouit des 3 propriétés suivantes:

1. $a \equiv a$ (réflexivité)
2. Si $a \equiv b$ et $b \equiv c$, alors $a \equiv c$ (transitivité)
3. Si $a \equiv b$, alors $b \equiv a$ (symétrie)

La relation d'équivalence jouit de la propriété intéressante qu'elle définit sur l'ensemble X une partition et inversément une partition définit une relation d'équivalence.

Exemple : soit l'ensemble $X = \{1,2,3,4,5,6,\}$ définissons la relation d'équivalence de la manière suivante: tous les éléments de X de la forme 2n (où n'est entier positif) sont équivalents, de même pour les éléments de la forme 2n + 1. Dans cet exemple simple on distingue alors deux classes d'équivalence

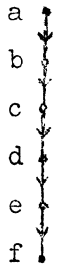
$\{1,3,5\}$ d'une part $\{2,4,6\}$ d'autre part.

On peut aisément vérifier les trois propriétés précédentes, et le fait que l'on a défini une partition sur X.

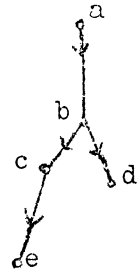
2. 4. Par ces préliminaires , un mathématicien peut voir que dans le linguistique mathématique actuelle il n'est pas nécessaire de faire appel à la théorie des ensembles dans sa totalité. Il suffit que l'on puisse disposer des structures les plus simples, celles qui sont définies en prenant des éléments dans les ensembles qui sont les plus bas dans l'échelle des ensembles de base E, à savoir les structures algébriques et les structures binaires et notamment les structures d'ordre.

Une structure d'ordre est définie sur un ensemble E, si l'on peut, étant donné deux éléments appartenant à E, établir entre eux une relation que nous écrivons par exemple (A plus petit que B) et que nous pouvons représenter graphiquement par deux points avec A et B et une flèche, et on peut montrer qu'une structure d'ordre possède certaines propriétés de symétrie, de transitivité, etc..

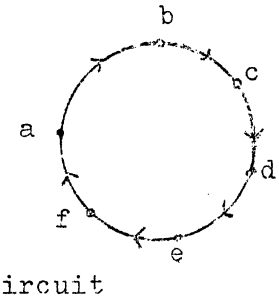
Il est commode d'utiliser une représentation graphique des relations binaires (et notamment des relations d'ordre). Pour cela on joint les éléments x et y de l'ensemble E par une flèche chaque fois que le couple (x,y) appartient à la structure en question. Il est alors facile de visualiser les différents types de graphes (équivalents aux relations binaires). On les voit sur les figures quivantes, qui présentent divers graphes et relations d'ordre définis sur un ensemble de six éléments:



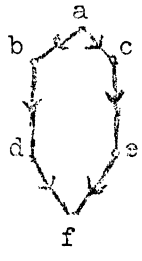
a) Chemin hamiltonien



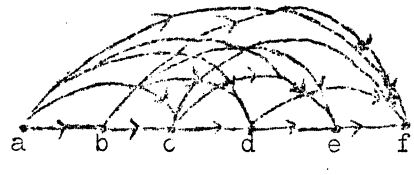
b) arborescence



d) circuit



c) treillis



e) ordre total

3. LES MODELES BINAIRES

3.1. Dans ces modèles, on considère des "ensembles linguistiques" (ensemble de phonèmes, de morphèmes etc..) et l'on s'efforce d'établir l'existence d'une structure de relation binaire soit sur l'ensemble E lui-même, soit sur $\mathcal{P}(E)$ en définissant des parties du texte étudié (qui respectent, en général, l'ordre linéaire de la chaîne parlée).

En réalité c'est l'existence d'un tel ordre linéaire qui justifie la recherche de structures d'ordre (mais alors non totalement ordonné). Dans cet esprit on distingue essentiellement deux tendances.

3.1. LA TENDANCE BAR-HILLEL CHOMSKY

A. L'idée de base chez Bar Hillel est que tous les mots d'un langage donné appartiennent à un ou plusieurs membres d'une hiérarchie infinie de catégories syntaxiques, dont deux sont considérées comme fondamentales, à savoir les catégories de chaînes nominales et des phrases, notées (n) et (s). Les autres sont des catégories d'opérateurs, dont les membres, les opérateurs, sont considérés comme se trouvant à côté de leur "arguments" se trouvant toujours immédiatement à leur gauche ou à leur droite (c'est pourquoi le "modèle" théorique correspondant est appelé: "modèle à constituant immédiat")

exemple: John slept

John = Chaîne nominale

slept= verbe intransitif, c'est-à-dire opérateur qui, avec une "nominale" à sa gauche, forme une phrase; nous noterons la catégorie de la façon suivante:

n\s (lire: n sous s)

little= adjectif, c'est-à-dire opérateur qui, avec une "nominale" à sa droite forme encore une nominale. Nous aurons donc affaire à la catégorie:

n/n (n sur n)

Little John slept soundly

soundly: adverbe (auprès d'un verbe intransitif)

(n\s) (n\s) ou n\s\\n\s

Problème:

Une certaine phrase a donné lieu à la transcription, à partir d'un dictionnaire de catégories, d'une suite de symboles:

n/n n n\s n\s\n\s

Trouver toute la structure de la phrase.

Méthode:

① n/n n

↓

opérateur qui avec une nominale à droite donne une nominale

d'où: l'échelon supérieur est: n

① ' n\s n\s\n\s

opérateur qui avec un opérateur n\s à sa gauche donne une n\s

d'où: échelon supérieur: n\s

2 n n\s

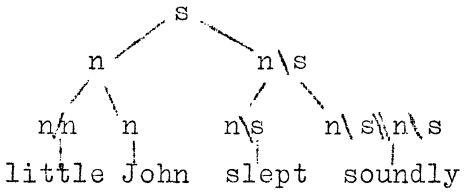
↓

opérateur qui avec une nominale à sa gauche donne une phrase

s

Conclusion:

La phrase proposée était réellement une phrase bien formée: elle correspond à la structure suivante:



exemple: little John slept soundly

Les règles d'opération sont simples: elle peuvent s'exprimer de la manière suivante:

$$X / \beta \frown \beta \longrightarrow \alpha$$
$$X \frown X \setminus \beta \longrightarrow \beta$$

B. Point de vue de Chomsky

Dans l'analyse à constituant immédiat, les mots d'une phrase sont groupés dans des chaînes constituantes plus petites et ainsi de suite jusqu'aux constituants ultimes. Ces chaînes sont alors classées comme chaînes nominales (NP), verbales (VP) etc.

Par exemple la phrase "the man took the book" peut être analysée de la façon suivante:

the man	took	the book
N P	verbe	N P
	VP	

On dit alors qu'on a affaire à un modèle à structure de chaîne.

En fait on peut aller encore plus loin dans l'analyse (afin de réaliser une plus grande économie de description, ce qui est le but de toute analyse structurale). En effet, on n'a considéré ici qu'une seule manière de traiter le verbe "took". Mais d'autres formes auraient pu apparaître; exemple: takes, has taken, has been taking, is taking, has been taken, will be taking etc... On peut donc considérer le verbe comme une suite d'éléments indépendants. Exemple: pour la chaîne "has been taking" nous pouvons séparer les éléments "have ....en" "be...ing" et "take" et nous pouvons alors dire que ces éléments se combinent librement.

Nous avons ainsi:

(1) # ⌢ the ⌢ man ⌢ Verb ⌢ the ⌢ book ⌢ #

(2) # ⌢ the ⌢ man ⌢ Auxiliary ⌢ V ⌢ the ⌢ book ⌢ #

(3) # ⌢ the ⌢ man ⌢ Auxiliary ⌢ take ⌢ the ⌢ book ⌢ #

(4) # ⌢ the ⌢ man ⌢ C ⌢ have ⌢ en ⌢ be ⌢ ing ⌢ take ⌢ the ⌢ book ⌢ #

(5) # ⌢ the ⌢ man ⌢ pas ⌢ have ⌢ en ⌢ be ⌢ ing ⌢ take ⌢ the ⌢ book ⌢ #

Définissons maintenant la classe AF comme contenant "en" "ing" et "C" et la classe V comme comprenant V, M, have, be. Nous pouvons alors convertir la ligne 4 en une suite ordonnée de morphènes par la règle suivante:

$$Af \frown v \longrightarrow v \frown Af \frown \#$$

⑥ the⌒man⌒have⌒past⌒#⌒be⌒en⌒#take⌒ing⌒#⌒the⌒book⌒#

Enfin en appliquant des règles morphémiques telles que les suivantes:

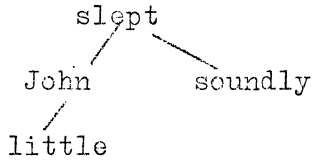
have ⌒ past → had
be ⌒ en → been
take ⌒ ing → taking

Nous arrivons à:

⑦ the man had been taking the book.

3.3. La conception que Tesnière se fait du langage diffère notablement en apparence de celle de Bar Hillel ou de Chomsky. La phrase, dit-il, est un ensemble organisé dont les éléments constituants sont les mots. Entre chaque mot et ses voisins plus ou moins proches, l'esprit aperçoit des connexions dont l'ensemble forme la charpente syntaxique de la phrase. Ces connexions ne sont indiquées par rien, il est indispensable de les matérialiser (automatiquement ou non) si l'on veut exprimer le véritable contenu de la phrase. Tesnière les figure à l'aide d'un diagramme bidimensionnel qu'il appelle le stemma.

En notation de stemma, l'exemple donné plus haut par Bar Hillel s'écrit:

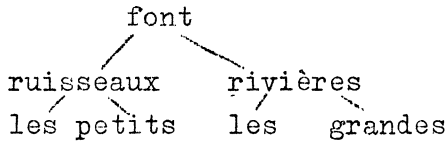


Relever un stemma ou en faire la "mise en phrase", c'est en transformer l'ordre structural en ordre linéaire. Inversement, comprendre une phrase, c'est en rétablir par la pensée les connexions, c'est transformer l'ordre linéaire en ordre structural.

Exemple:

Ordre linéaire: "les petits ruisseaux font les grandes rivières"

Ordre structural:



La forme des graphes destinés à répresenter la structure du langage varie avec le degré de finesse de l'analyse. La hiérarchie fondamentale est sans aucun doute celle des stemmas arborescents utilisés par Harper et Hays. Une étude telle que celle de Tesnière conduit à matérialiser en outre d'autres hiérarchies secondaires dites "anaphoriques" et à mettre en évidence des phénomènes de "translation" et de "jonction". On obtient ainsi des graphes plus compliqués mais aussi plus explicites.

3.4. INTERPRETATION ENSEMBLISTE DE LA DUALITE DES THEORIES LINGUISTIQUES

A) DEFINITION D'UN ENSEMBLE DE REFERENCE

Prenoms comme ensemble de référence la phrase de Chomsky citée plus haut en exemple:

the man took the book.

Il est légitime d'en numéroter les éléments constituants (lettres, phonèmes mots, etc, comme l'on veut) pour matérialiser leur succession dans le temps tout au long du fildu discours. Supposons que les mots soient choisis ici comme éléments de l'ensemble de référence E. Cette numérotation mettra en évidence le fait que les deux "th" sont deux évènements distincts dans le temps, donc deux éléments différentes de l'ensemble E.

the	man	took	the	book
n+1	n+2	n+3	n+4	n+5

B) LINGUISTIQUE DE TESNIERE

L'existence d'un lien hiérarchique entre deux mots constitue une relation binaire entre deux éléments de l'ensemble E, donc un élément de E x E . Un stemma ne lie que certains mots entre eux, l'ensemble de ses liens n'est qu'une partie de E x E, donc un élément de $\mathcal{P}(E \times E)$.

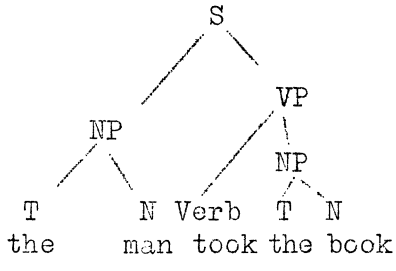
C) DECOUPAGES SELON CHOMSKY.

Dans l'étude de Chomsky, on met en lumière l'existence de chaînes nominales. L'une d'elles par exemple, peut être considérée comme s'incarnant dans la chaîne

" the book "
  n+4 n+5

une autre dans la chaîne" the man". On montre aussi que la séquence
                          n+1 n+2

"took the book" peut être considérée comme l'incarnation
 n+3 n+4 n+5
d'une catégorie syntaxique de "phrase verbale" et ainsi de suite:



Le principe même de ce découpage montre que les catégories syntaxiques sont isomorphes de groupes de mots, c'est-à-dire de parties de l'ensemble E. Par cet isomorphisme, toute catégorie syntaxique peut être assimilée à un élément p (E) de l'ensemble $\mathcal{P}(E)$ des parties de E.

Soit M ensemble des catégories syntaxiques reconnues dans la phrase considérée, M est ensemble dont les éléments sont des p(E), donc un ensemble de parties de $\mathcal{P}(E)$ donc

$$M \subset \mathcal{P}(\mathcal{P}(E))$$

Enfin, l'étude des relations entre catégories syntaxiques se fera dans l'ensemble

$$R \in \mathcal{P}(\mathcal{P}(E) \times \mathcal{P}(E))$$

D) MISE EN CORRESPONDANCE DES DEUX LINGUISTIQUES

La comparaison de constructions théoriques portant sur des niveaux différents de l'échelle des ensembles de base E n'a guère de signification; seule une application d'un niveau sur l'autre peut permettre une mise en parallèle précise des deux linguistiques évoquées plus haut. On voit aisément qu'une application de E sur $\mathcal{P}(E)$ conduirait du niveau $\mathcal{P}(E \times E)$ au niveau $\mathcal{P}(\mathcal{P}(\mathcal{P}(E) \times \mathcal{P}(E))$

4. LES MODELES DE PARTITIONS

4.1. Les modèles décrits précédemment sont "binaires" en ce sens qu'ils manifestent l'existence de relations d'ordre entre éléments d'un ensemble linguistique. Mais chez TESNIERE (et aussi chez HAYS et chez ANDRIEV) ces éléments sont les mots de la langue écrite. Au contraire chez BAR-HILLEL, CHOMSKY (et aussi OETTINGER) ce sont des fragments de phrase; si E est l'ensemble des mots, on a dans un cas une structure appartenant à $\mathcal{P}(E \times E)$ dans l'autre à $\mathcal{P}(\mathcal{P}(E) \times \mathcal{P}(E))$. Il est intéressant d'étudier systématiquement les familles de mots puisqu'elles définissent dans la linguistique traditionnelle les catégories. C'est ce que différents auteurs ont récemment tenté. On a alors des structures qui appartiennent à $\mathcal{P}(\mathcal{P}(E))$ ou à $\mathcal{P}(\mathcal{P}(\mathcal{P}(E)))$.

4.2. KULAGINA

A) But:

Il s'agit d'élaborer une grammaire spéciale, valable pour toutes les langues jouissant de toute la rigueur d'une théorie mathématique et évitant le caractère descriptif de notions insuffisamment définies telles qu'on en trouve constamment dans les grammaires existantes. Pour cela KULAGUINA propose une méthode de définition des notions grammaticales à l'aide de la théorie des Ensembles.

B) Ensemble de base:

L'ensemble de base est l'ensemble des mots de l'alphabet. Sur cet ensemble on définira une partition Γ et les sous-ensembles qui lui correspondent seront appelés les environnements des éléments qu'ils contiennent.

Exemples: 1) Soit S un substantif, on prendra comme environnement S au singulier et S au pluriel. Dès lors:

$$\Gamma\begin{pmatrix}\text{arbre}\\ \text{tapis}\end{pmatrix} = \begin{Bmatrix}\text{arbre, arbres}\\ \text{tapis, tapis}\end{Bmatrix} \quad \text{par convention.}$$

2) Soit un adjectif A, l'environnement de A est l'ensemble: A au masculin singulier, au féminin singulier, au masculin pluriel, au féminin pluriel.

Γ(abstraite)= {abstrait, abstraite, abstraits, abstraites}
Γ(conforme) = {conforme, conforme, conformes, conformes }

Phrases repérées:

Une phrase est une suite de mots de l'ensemble Σ de base. Parmi toutes ces phrases, on distinguera le sous-ensemble Θ = {A} des phrases dites repérées et qui pourront être par exemple le sous-ensemble de phrases correctes du point de vue grammatical; il est bon de remarquer qu'une phrase repérée peut avoir ou non un sens.

Grossissement:

Soit x un élément de l'ensemble Σ; dans une partition -$B^{(1)}$ x appartient à un sous ensemble qui est lui-même un élément de la partition -$B^{(1)}$ et que nous appellerons $B^{(1)}(x)$; soit alors deux partitions $B^{(1)}$ et $B^{(2)}$, on dira que $B^{(1)}$ est un grossissement de $B^{(2)}$, si on a:

$$\begin{cases} \forall x \in \Sigma \; B^{(1)}(x) \supset B^{(2)}(x) \\ \exists x \in \Sigma \; B^{(1)}(x) \supset B^{(2)}(x) \end{cases}$$

Structure B d'une phrase A:

Soit A = {x_1 x_2 ... x_n} une phrase. Nous savons qu'à chaque x correspond dans une partition B l'élément B (x). La structure B d'une phrase A sera par définition l'image de A lorsqu'on applique chacun des x sur B. Autrement dit:

$$B\ (A) = (x_1)\ B\ (x_2) \ldots B\ (x_n).$$

Et une structure B sera dite repérée s'il existe au moins une phrase répérée qui s'applique sur la structure B donnée.

Exemple: A = {mignonne fillette brune .}

Prenons comme partition -B, la partition nommée Γ et définie ci-dessus; on aura:

Γ(A) = Γ(mignonne) Γ(fillette) Γ(brune) soit:
Γ(A) = [{mignon, mignonne, mignons, mignonnes}, {fillette, fillettes,}
{brun, brune, bruns, brunes .}]

Cette structure -Γ de A est bien repérée puisqu'il existe une phrase repérée {mignonne fille bune} dont la structure -Γ est celle donnée précédemment.

-B équivalence:

Deux éléments quelconques B_i et B_j d'une partition B de Σ (qui sont donc des sous-ensembles de) sont dits B -équivalents, ce que l'on note par $B_i \underset{B}{\sim} B_j$ si quelles que soient les phrases A_1 et A_2, les structures -B

$B(A_1)\ B_1\ B(A_2)$ et $B(A_1)\ B_j B(A_2)$ sont simultanément repérées ou non repérées. On peut vérifier que cette équivalence jouit des 3 propriétés de l'équivalence mathématique définie dans le chapitre précédent. En effet:

1) $B_i \underset{B}{\sim} B_i$ (réflexitif)

2) $B_i \underset{B}{\sim} B_j$ et $B_j \underset{B}{\sim} B_k$ alors $B_i \underset{B}{\sim} B_k$ (transitivité)

3) $B_i \underset{B}{\sim} B_j$ alors $B_j \underset{B}{\sim} B_i$ (symétrie)

Exemple: Si nous choisissons comme partition B la partition telle que pour tout x de Σ $B(x) = \{x\}$; alors le fait que la substitution, dans une phrase correcte grammaticalement, d'un substantif masculin singulier par exemple, par un autre substantif masculin singulier ne change pas le caractère repéré de la phrase, nous permettra de déduire que deux substantifs au masculin singulier sont -B équivalents.

En effet, la phrase A = "Je veux un livre neuf" donne dans la structure B telle que $B(x) = \{x\}$ l'image:

$$B(A) = \underbrace{\{Je\}\ \{veux\}}_{B(A_1)}\ \{un\}\ \underbrace{\{livre\}}_{B_i}\ \underbrace{\{neuf\}}_{B(A_2)}$$

Substituons à $B = \{livre,\}$ $B_j = \{crayon\}$ on obtient

$$B(A') = \underbrace{\{Je\}\ \{veux\}\ \{un\}}_{B(A1)}\ \underbrace{\{crayon\}}_{B_j}\ \underbrace{\{neuf\}}_{B(A_2)}$$

qui est une structure repérée. On pourrait facilement vérifier que B(A) non repérée entrainerait B (A') non repérée:

Je veux des crayon neuf entrainant
Je veux des livre neuf

Grossissement régulier

On dira que la partition -B (1) est un grossissement régulier de B (2) si:

1) B (1) est grossissement de B (2)
2) la condition $B^{(2)}(y) \subset B^{(1)}(x)$ entraine que $B^{(2)}(y) \underset{B^{(2)}}{\sim} B^{(2)}(x)$

Autrement dit en opérant le grossissement, on a réuni dans un même élément $B^{(1)}(x)$ de la partition $B^{(1)}$ tous les éléments $B^{(2)}(y)$ qui sont -$B^{(2)}$ équivalents à $B^{(2)}(x)$.

Partition dérivée:

Soit une partition -B de Σ ; réunissons dans un même sous-ensemble B'(x) de Σ tous les éléments B_i de B qui sont B équivalents à B (x) nous définissons une nouvelle partition de Σ que l'on appelle la "partition dérivée" (ou première dérivée) de B.

Notons qu'une partition dérivée B' de la partition -B est un grossissement régulier B' de B mais inversément un grossissement régulier n'est pas nécessairement une partition dérivée du fait que la condition 2) qui définit le grossissement régulier ne comporte pas sa réciproque.

Structures subordonnantes et structures subordonnées.

Soit une partition $B^{(1)}$ grossissement d'une partition $B^{(2)}$. Chaque structure $B^{(1)}$ est d'une part l'image de la phrase dans l'application $x \longrightarrow B^{(1)}(x)$ et d'autre part l'image de la structure $B^{(2)}$ dans l'application $B^{(2)}(x) \longrightarrow B^{(1)}(x)$.

Par définition, la structure $B^{(1)}(x)$ obtenue à partir de $B^{(2)}$ sera la structure subordonnante de la structure $B^{(2)}$, et même on remarquera que plusieurs $B^{(2)}$ distincts peuvent donner un même $B^{(1)}$. Ces structures $B^{(2)}$ seront les structures subordonnées à la structure B^1. Si le grossissement est régulier alors il se produit le fait intéressant qu'une structure $B^{(2)}$ repérée (ou non repérée) donne une structure B^1 repérée (ou non répérée) et on démontre le théorème suivant:

"La dérivée seconde d'une partition coincide toujours avec la prémière dérivée de la partition, c'est-à dire que B" (x)=B' (x) pour $x \in \{$

Partition unique:

C'est la partition où E (x)= {x}
Exemple: {Je} {veux} {un} {livre} {neuf}

Partition en famille:

C'est la partition dérivée de E que l'on notera S. On obtient alors:

1) Quatre familles de substantifs
   Famille des substantifs masculins singuliers
   " " " masculins pluriels
   " " " féminins singuliers
   " " " féminins pluriels

2) quatre familles d'adjectifs

Familles des adjectifs au masculin singulier
" " " masculin pluriel
" " " féminin singulier
" " " féminin pluriel

Langue simple:

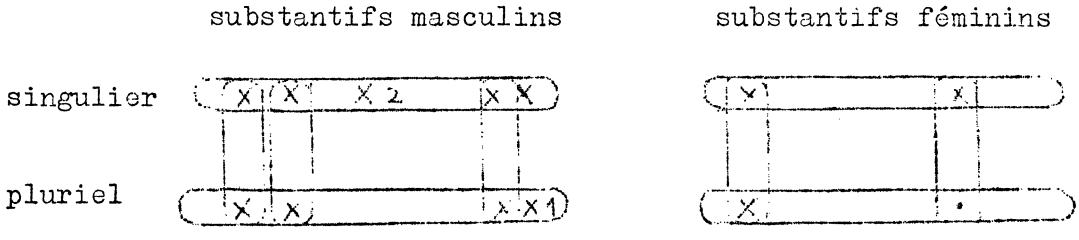
Par définition, une langue définie par la donnée de son ensemble Σ de sa partition Γ, et de l'ensemble des phrases repérées Φ soit Φ est dite simple s'il existe entre les partitions S et Γ les relations suivantes:

1) $\forall x \in \Sigma \rightarrow \Gamma(x) \cap S(x) = \{x\}$

2) $\forall x_1 \in \Gamma(x)$ et $\forall x_2 \in S(x)$
$S(x_1) \cap \Gamma(x_2) \supseteq \{y\}$

Exemples: Avec les définitions de Γ et S on peut montrer que la langue française est une langue simple tandis que le russe ne l'est pas. On peut penser que d'autres définitions de Γ et S inverseraient les conclusions. L'important est de noter que dans les cas d'une langue simple relativement aux Γ et S définis, on peut tirer des conclusions particulièrement intéressantes.

En français, on a ainsi:



Ici nous avons schématisé
la partition - Γ par les sous-ensembles $\{x\}$

La partition -S par les sous-ensembles $\{x\ x\ x\ x\ x\ x\ x\}$

On vérifie aisément les deux conditions d'une langue simple. En russe, du fait que les adjectifs au pluriel ne changent pas avec les différents genres, on a les familles suivantes.

subst.	pluriel	nominatif	subst. masc.	sing.	N	subst. fém.	sing.	N	subst. neutre
"	"	génitif	"	"	G			G	
"	x2	accusatif	"	"	A			A	
"	"	datif	"	"	D			D	
"	"	prépositionnel	"	"	P			P	
"	"	instrumental	"	"	I			I	

Si on prend:

x_1 = substantif masculin singulier au génitif $\in f(x)$

x_2 = " féminin pluriel à l'accusatif $\in S(x)$

On vérifie aisément que:

$S(x_1) \cap \Gamma(x_2) = \emptyset$

puisque $S(x_1)$ est l'ensemble de tous les substantifs masculins singuliers au génitif et $\Gamma(x_2)$ l'ensemble des formes du substantif féminin x_2 au singulier et au pluriel à tous les cas.

Classe:

On appelle classe $K(x)$ du mot x d'une langue Σ (Γ) l'ensemble des mots x' tels que, ou bien $\Gamma(x)$ et $S(x')$ se coupent ou bien $\Gamma(x')$ et $S(x)$ se coupent. Pour une langue simple, ces deux conditions sont équivalentes, et de plus les classes forment une partition de Σ.

Exemple:

1) Les substantifs forment deux classes:

substantifs masculins
substantifs féminins

2) Les adjectifs constituent une classe

Types:

C'est la partition dérivée d'une partition par classe.
1) les deux classes de substantifs forment un type.
2) la classe des adjectifs constitue un type.

Configurations

Soit B une partition de Σ. On appelle configuration -B du premier ordre une structure -B notée $\hat{B}(1)$ telle que:

1) $\hat{B}(1)$ possède au moins deux éléments.

2) Il existe au moins un élément $B_{\times 1}$ de la partition -B tel que, quelles que soient les phrases A_1 et A_2, $B(A_1)B_{(1)}B(A_2)$ et $B(A_1)\,B_{\times 1}\,B(A_2)$ sont simultanément repérées ou non repérées.

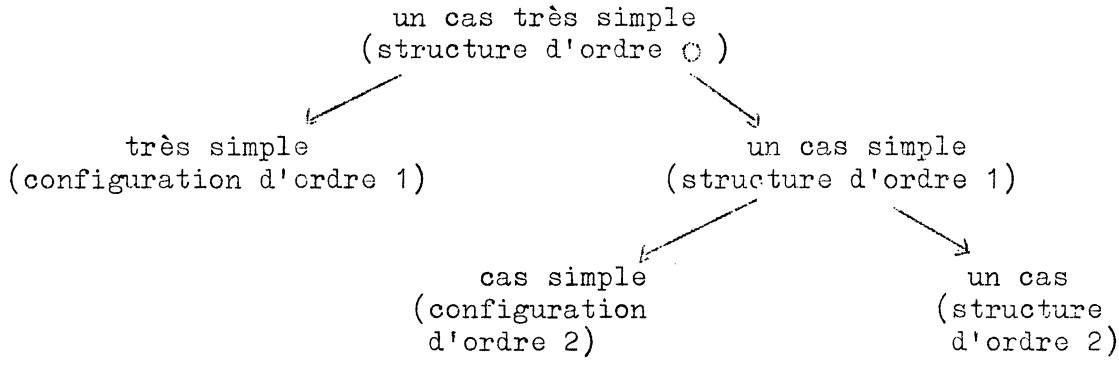
$B_{\times 1}$ est l'élément-résultant de $\hat{B}_{(1)}$

On appellera structure -B du premier ordre, une structure -B ne contenant pas de configuration -B du premier ordre. Les configurations B d'ordre k se définissent par récurrence.

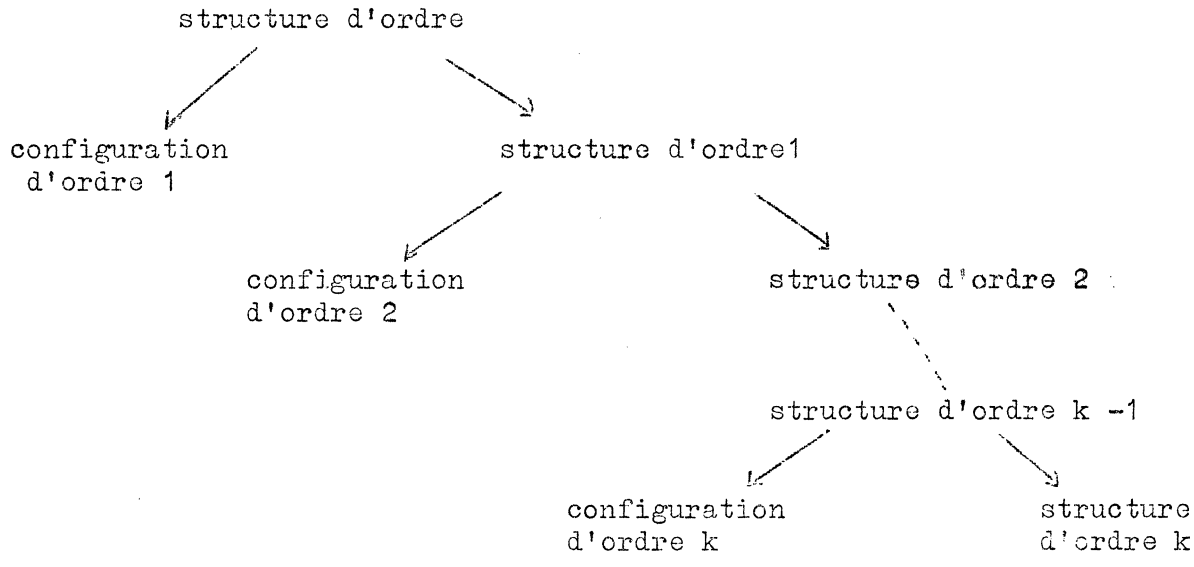
On appelle configuration -B d'ordre k, une structure -B notée $\hat{B}(k)$ telle que:

1) $\hat{B}_k$ possède au moins deux éléments.
2) Il existe au moins un élément $B_{\times k}$ de la partition -B tel que les structures -B d'ordre k-1, $B_{(k-1)}(A_1)\,\hat{B}_k\,B_{(k-1)}(A_2)$, $B_{(k-1)}(A_1)B_{\times k}\,B_{(k-1)}(A_2)$ pour des phrases A_1 et A_2 quelconques, se trouvent être simultanément repérées ou non repérées.

Exemple: Soit la phrase "un cas très simple". Repérons les phrases suivantes: "un cas très simple" "un cas simple", "un cas". On peut donc titer de cette phrase la configuration d'ordre 1 "très simple" avec l'élément résultant "simple". Il reste alors la structure d'ordre 1 "un cas simple" d'où on peut de nouveau tirer la configuration d'ordre 1 "cas simple" avec l'élément résultant "cas":



Le schéma général étant:



L'ordre le plus élevé d'une structure sera par définition le "rang" de cette structure.

On peut alors étudier la manière dont se comporte l'ordre d'une configuration lorsque l'on passe d'une structure suivant une certaine partition -D à la structure suivant une partition -B grossissement de D. On remarque en particulier que l'ordre de la configuration se conserve si le grossissement est régulier. Si cet ordre se conserve sans que le grossissement soit régulier, on dira que la configuration -D est subordonnée à la configuration -B ou que la configuration -B est subordonnante par rapport à la configuration D.

On peut définir des relations entre les éléments d'une configuration -B donnée et la partition D. Grâce à cette notion assez délicate à interpréter dans la première publication de Kulaguina, on peut montrer par exemple l'équivalent mathématique de l'accord grammatical qui serait une relation de premier genre entre le type des adjectifs et le type des substantifs. On peut également aboutir à une définition du régime et de l'accompagnement dans la langue russe.

4.3. SESTIER

Le travail récent de A. SESTIER est intéressant en ce qu'il se situe en un point central par rapport au positions des écoles du type TESNIERE comme de celles du type CHOMSKY ou KULAGINA. SESTIER étudie en effet les relations existant entre deux ensembles

E ensemble de mots et
F ensemble de contextes

si $x \in E$ et $y \in F$ la relation
x a y signifie x est accepté par y.

Pour garder plus de généralité, on ne spécifie pas les conditions linguistiques de définition du contexte. Mais il est clair qu'en général F peut-être identifié à une partie de $\mathcal{P}(E)$ ou même de $\mathcal{P}(\mathcal{P}(E))$.

Si X est une partie de E et Y est une partie de F et si on désigne par a (x) l'ensemble des contextes qui acceptent le mot x (c'est une partie de F) et par a^{-1} (y) l'ensemble des mots acceptés par y (c'est une partie de E) on peut associer à toute partie X et à toute partie Y les parties

$a[X] = \bigcap_{x \in X} a(x)$ coupe étroite directe associée à X

$a^{-1} Y = \bigcap_{y \in Y} a^{-1}(y)$ coupe étroite inverse associée à Y

Les X et les Y sont des familles de parties entre lesquelles il existe une correspondance de Galois par les fermés.

$\underline{\underline{X}} = a^{-1}[a \quad X])$ $\qquad$ $\underline{\underline{X}} = X$

$Y = a \quad [a^{-1} \; Y])$ $\qquad$ $\bar{\bar{Y}} = Y$

l'ensemble des $\underline{\underline{X}}$ et l'ensemble des $\bar{\bar{Y}}$ sont des treillis complets ordonnés par l'inclusion.

On peut alors définir des classes d'équivalences du type

$X' \lesssim^{y} x \iff a(x') = a(x)$

$X' \lesssim^{x} x \iff a^{-1}[a(x')] = a^{-1}[a(x)]$

entre des éléments de E qui ont, par exemple, même fermeture par la relation a.

On définit alors un ordre partiel entre relations d'équivalence qui peut également servir à l'étude des fermés (c'est-à-dire, pour le linguiste, des catégories, grammaticales ou autres).

De la même façon il est possible de définir une relation d'équivalence (indicée dans F) χ_y telle que l'expression
$X \chi_y X'$ signifie que x est substituable à X' dans le contexte y. On a évidemment $\chi_y(X) = a^{-1}(y)$

On aura alors la relation

$X \gtrsim X'$ qui signifie que X est substituable à X'
dans tous les contextes y qui acceptent x, c'est-à-dire dans tous les $y \in a(x)$.

$$\gtrsim(x) = \bigcap_{y \in a(x)} \chi_y(X) = a^{-1}[a(X)]$$

toute $\gtrsim(x)$ est une fermeture de x et réciproquement.

Muni de cet arsenal, on peut s'efforcer de construire systématiquement un jeu de catégories à partir d'un corpus donné. Le choix de la méthodologie (treillis des fermés, relation d'équivalence etc...) depend alors des possibilités de mécanisation de l'algorithme.

4.4. LECERF

Comme A. SESTIER, Y. LECERF travaille simultanement sur E et $\mathcal{P}(E)$. Mais, au lieu de rechercher les catégories sur un corpus, il se sert d'un dictionnaire comportant les indications de catégories pour énoncer en termes algebriques les règles de la grammaire (et ceci afin d'aboutir à une analyse grammaticale automatique.)Son dessein est donc plus proche de ceux de BAR-HILLEL et de HAYS. Mais son avantage sur ces derniers auteurs qui travaillent l'un sur $\mathcal{P}(E)$ l'autre sur E est de definir une structure plus riche, qui les englobe l'un et l'autre.

A) DEFINITION DES G-SYNTAXES

a) Construction de syntagmes à partir d'un dictionnaire.

Etant donné un ensemble D, ou dictionnaire, d'éléments m_j, ou mots, on se propose de l'utiliser pour construire des éléments s_k, ou syntagmes, d'un autre ensemble S qui pourra éventuellement être infini. Le procédé de construction étant fixé, on conviendra que S est l'ensemble des syntagmes s_k que ce procédé permettrait d'obtenir en agissant suffisamment longtemps. Mais à tel ou tel instant, il se peut que tous les s_k ne soient pas encore construits. On différenciera les syntagmes déjà construits en les écrivant entre deux crochets: $[s_k]$. Dans la suite, et en l'absence d'autre indication, toute expression entre crochets représentera un syntagme déjà construit au moment considéré.

b) Procédés de construction mettant en jeu une famille d'opérateurs.

Dans de tels procédés, on définit une famille F d'opérateurs notés (op_i), et on les utilise comme suit:

1°) Le résultat, s'il existe, et il n'existe pas nécessairement, de l'application à un s_k quelconque d'un (op_i) est encore un élément de S. On se trouve ainsi avoir construit un syntagme, et on le note entre crochets $[s_k] \cdot (op_i)]$

2°) Un tel procédé ne peut démarrer que si l'on a un stock initial de syntagmes. Aussi, on décide d'intégrer initialement à S tous les mots du dictionnaire. En tant que syntagmes déjà construits, ces mots s'écrivent entre crochets $[m_j]$

Tout élément déjà construit de S se présentera finalement soit comme étant un certain $[m_j]$, soit comme étant le résultat de l'application successive d'opérateurs à un certain $[m_j]$. Dans les deux cas, le mot en question sera appelé "tête" du syntagme.

Deux syntagmes seront déclarés différents s'ils ne résultent pas de l'application des mêmes opérateurs dans le même ordre au même $[m_j]$.

Rien ne s'oppose à ce que la famille F d'opérateurs soit elle-même définie comme une fonction de l'état t d'avancement de la construction des $[s_k]$. On l'écrit alors F (t).

c) Cas particulier: G-syntaxe sur un dictionnaire D.

Par définition, le procédé de construction défini au paragraphe A)b) ci-dessus sera appelé G-Syntaxe si, à un instant donné de la construction, il est toujours possible d'associer de façon unique, à un $[s_j]$ déjà construit, un opérateur $\cdot([s_j])$ à droite et un opérateur $([s_j])_\circ$ à gauche, l'ensemble de ces opérateurs constituant la famille F (t).

Deux opérateurs seront déclarés différents s'ils sont associés à des $[s_j]$ différents, ou si, associés à un même $[s_j]$, ils sont l'un opérateur à droite et l'autre opérateur à gauche.

L'expression $([s_j])$ sans point d'opération est appelée "opérateur neutre" associé à $[s_j]$

d) Exemples

opération à droite:

$$[s_j] \cdot ([s_i]) \longrightarrow [[s_j] \cdot ([s_i])]$$

opération à gauche:

$$([s_i]) \cdot [s_j] \longrightarrow [([s_i]) \cdot [s_j]]$$

Ces notations permettent de reconnaître quelle série d'opérations a permis de construire tel ou tel syntagme donné.

exemples de syntagmes:

[([([the]) • [man]]) • [[hit] • ([[the]) • [ball])]]

[([un]) • [([[très]) • [gros]]) • [chien]]]

B) RAPPEL: SEQUENCES CANONIQUES DE SIGNES D'OUVERTURE, SIGNES DE FERMETURE ET SIGNES NON ORIENTES

a) Séquences canoniques de signes d'ouverture, signes de fermeture et signes non orientés.

Par commodité, nous appellerons parenthèses les signes orientés, et marquants les signes non orientés. Une séquence de parenthèses et de marquants est dite canonique s'il est possible de mettre en correspondance biunivoque deux à deux respectivement l'ensemble des marquants, l'ensemble des parenthèses de gauche, l'ensemble des parenthèses de droite, de façon à satisfaire aux conditions suivantes:

1°) les deux parenthèses associées à un même marquant se font face selon leurs concavités, délimitant un segment qui est dit domaine de ce marquant.

2°) un marquant est toujours à l'intérieur de son propre domaine

3°) deux domaines distincts peuvent avoir des rapports d'inclusion ou d'exclusion, mais jamais d'intersection.

4°) si A et B sont des marquants distincts, et que le domaine de A contient B, alors le domaine de B ne contient pas A.

5°) il existe un domaine, et un seul, qui contient tous les autres.

b) Ordre structural et ordre linéaire dans une séquence canonique

Par hypothèse, la succession des signes dans l'expression étudiée établit entre eux une relation d'ordre, que l'on appelle ordre linéaire. Mais on montre facilement que la relation "X est inclus dans le domaine de Y", où X et Y sont deux marquants, est aussi une relation d'ordre que l'on appelle ordre vertical ou parfois aussi ordre structural. Elle est définie sur l'ensemble des marquants seulement. Le marquant dont le domaine contient tous les autres est le plus grand élément selon cet ordre vertical.

Ordre linéaire et ordre vertical ne sont pas indépendants l'un de l'autre. Si l'ordre linéaire des marquants est donné, leur ordre vertical n'est pas quelconque, mais doit être choisi parmi un nombre nettement plus restreint de possibilités. La relation de liaison entre les deux ordres s'écrit:

A non $\wedge$ B V C impose que l'on n'ait
ni B < A < C ni C < A < B

c) arborescence associée à une séquence canonique

On utilise très couramment une arborescence pour représenter la relation d'ordre vertical entre les marquants. Cette arborescence est dite: arborescence associée à la séquence canonique. La relation entre l'ordre vertical et l'ordre linéaire entraîne certaines propriétés géométriques (absence de croisements de certaines lignes).

C GRAPHES ASSOCIES UN G-SYNTAGME QUELCONQUE

a) Théorème 1.

Si, dans l'expression de l'opérateur neutre associé à un G-syntagme, on efface tous les crochets et tous les points opératoires "à droite" ou "à gauche", alors, l'ensemble résiduel de mots et de parenthèses est une séquence canonique.

L'arborescence associée, dont tous les sommets sont des mots, est dite "G-stemma".

Corollaire:

Le stemma a toutes les propriétés des arborescences associées à une séquence canonique. La relation entre l'ordre linéaire et l'ordre structural est en particulier satisfaite.

b) Théorème 2

Si dans l'expression de l'opérateur neutre associé à un G-syntagme, on efface toutes les paranthèses, alors, l'ensemble résiduel de mots, point opératoires et crochets est une séquence canaonique, où les mots et les points jouent le rôle de marquants.

L'arborescence associée est bifurcante. Les mots y jouent tous le rôle de sommets pendants. Les points opératoires occupent les sommets d'où partent deux arcs descendants. Cette arborescence est dite "G-graphe de structure". (On peut dire aussi: G-diagramme de dérivation).

Corollaire

Le G-graphe de structure a toutes les propriétés des arborescences associées à des séquences canoniques. La relation entre l'ordre linéaire et l'ordre structural est en particulier satisfaite (principe du constituant immédiat continu).

c) Complémentarité, compatibilité, indépendance.

Un G-graphe de structure et un G-stemma bâtis au hasard sur les mots $M_1\ M_2\ \ldots M_n$ ne sont en général pas associables à un même G-syntagme. On dit qu'ils ne sont pas compatibles.

On montre facilement que la donnée du G-stemma et du G-graphe de structure issus d'un même G-syntagme dont les mots sont $M_1\ M_2\ \ldots M_n$, suffit à définir ce syntagme, c'est-à-dire son expression en mots, parenthèses, crochets, points opératoires.

Au contraire, la donnée de G-stemma ne suffit pas, et celle du G-graphe de structure non plus. Supposons par exemple donné le second: à un G-graphe de structure portent sur les mots $M_1 M_2 \ldots M_n$ correspondent 2^{n-1} possibilités de G-stemmas compatibles, d'où autant de syntagmes possibles différents.

D) APPLICATION A LA SYNTAXE DES LANGUES NATURELLES

a) Hiérarchies observées par les linguistes.

Des observations accumulées depuis des siècles par un grand nombre de linguistes, puis systématisées, en particulier par l' cole structuraliste américaine, ont concouru à faire apparaître les phrases et les syntagmes comme des ensembles hiérarchisés, représentables par des arborescences bifurcantes où les mots jouent le rôle de sommets pendants (cf. Bar Hillel, N. Chomsky, R. Harris, V. Yngve etc.)

Un autre série d'observations, moins nombreuses, rassemblées par d'autres linguistes, ont conduit aussi à l'idée de hiérarchies représentables par des arborescences; mais dans celles-ci, tous les sommets sont des mots. Le fait que, par exemple, les stemmas de Tesnière et les arborescences de Hays aient été construits de façon indépendante, avec une assez bonne convergence, renforce la valeur de cette seconde série d'observations.

La question de savoir laquelle des deux sortes de graphes convenait pour représenter les hiérarchies syntaxiques a donné lieu, parfois, à une polémique, où les tenants de chaque camp s'efforçaient de présenter les observations adverses comme entachées d'erreur et de préjugé.

Or les arborescences à mots pendants utilisées par les linguistes de la première tendance possèdent en général toutes les caractéristiques des G-graphes de structure décrits au paragraphe C)b).

Les arborescences de Tesnière, Hays etc...ont en général toutes les propriétés des G-stemmas décrits au paragraphe C)a).

En outre, les arborescences proposées pour des mêmes phrases par l'une et l'autre tendance linguistique sont, dans l'ensemble, compatibles. Il est possible, en se fondant sur les unes et les autres simultanément, de construire des G-syntaxes de langues naturelles, alors que chacune des tendances linguistiques considérée isolément n'apporte pas une information suffisante dans ses graphes représentatifs.

b) Portée de ces coincidences

Ces coincidences ne peuvent être accidentelles.

1°) Le fait que les "graphes de structure" utilisés par Chomsky Bar Hillel, Oettinger, Yngve... soient des arborescences associables à des séquences canoniques, a été décrit sous le nom de principe du constituant immédiat continu. De tels graphes ne sont pas du tout quelconques.

2°) Le fait que les "stemmas" utilisés, par exemple, par Tesnière et Hays, soient des arborescences associables à des séquences canoniques a été décrit sous le nom de projectivité. Il implique une contrainte importante. Exemple: plus de 96% des arborescences que l'on peut construire avec, comme sommets, 7 mots donnés dans un certain ordre, ne sont pas associables à des séquences canoniques.

3°) La condition de compatibilité entre un G-graphe de structure et un G-stemma est également difficile à satisfaire. C'est en elle que se réalise la coincidence la plus remarquable.

Etant admis le caractère systématique de ces coincidences, on doit montrer qu'elles n'expriment pas des tautologies. Or ce second point résulte du fait que les graphes de structure et les "stemmas" ne sont pas réductibles les uns aux autres.

On notera que la "projectivité" n'est pas réductible au principe du constituant immédiat.

5. CONCLUSION

efforts déployés afin de créer des algorithmes pour la traduction à l'aide de machines de textes d'une langue dans une autre ont montré que les grammaires existantes ne conviennent pas toujours à l'élaboration des algorithmes, il y a donc insuffisance des grammaires. Les règles grammaticales créées jadis dans tout autre but que la traduction mécanique utilisent souvent des notions insuffisamment définies ayant un caractère descriptif et faisant appel le plus souvent au sens logique. Les algorithmes pour la traduction mécanique par contre sont fondés sur l'analyse d'indices formels et exigent des définitions rigoureuses, dans lesquelles on ne peut pas se permettre de tenir compte des rapports logiques. Par rapport logique il faut plutôt entendre d'ailleurs rapport sémantique, rapport de signification. Mais en fait on est obligé d'en tenir compte, il n'y a absolument pas moyen d'y échapper. Le problème qui se pose par conséquent est la création d'une grammaire spéciale valable pour toutes les langues, plus exactement qui soit suffisamment générale pour que les grammaires particulières des langues en soient des cas particuliers et construite de la même façon que le sont les théories mathématiques. Il faut nécessairement placer à la base d'un tel système un ensemble bien délimité de notions et dont les définitions ne font pas partie du système considéré, c'est-à-dire de notions dont les définitions sont fournies de l'extérieur sous forme de données. Toutes les autres notions dont on se sert doivent être rigoureusement définies et les assertions que l'on peut émettre à leur sujet doivent être prouvées. L'élaboration d'un tel système grammatical demande un long travail durant lequel des modifications et des compléments seront inévitables. A ce jour on n'a effectué que les premiers pas dans cette direction".

Ce texte est extrait de l'important travail de O. KULAGINA auquel nous avons fait allusion dans le paragraphe précédent. Il définit bien l'objet de la linguistique mathématique "certaine".

Cette linguistique mathématique, ainsi basée sur la théorie des ensembles, pourra être subdivisée en linguistique algebrique, topologique etc... Des structures de plus en plus riches pourront ainsi être définies, que l'on s'éfforcera d'utiliser soit pour décrire les langages matériels, soit pour les manipuler.

Mais le choix de l'outil mathématique n'est pas arbitraire, il est commandé par la nature concrète de langage qui est un phénomène qui nous est donné par la nature, tout comme les phénomènes de la physique et de la chimie. On le voit bien en étudiant les trois modèles linguistiques développés par Chomsky et en suivant le déterminisme qui le conduit du premier au second, puis au troisième.

La "Finite state grammar" travaille sur l'ensemble E des mots. Mais la phrase "structure grammar" travaille sur l'ensemble E et sur une partie de l'ensemble $\mathcal{P}(E)$. La "language transformation structure grammar" travaille sur l'ensemble E, l'ensemble $\mathcal{P}(E)$ et l'ensemble $\mathcal{P}(\mathcal{P}(E))$. On s'est donc élévé encore une fois dans la complexité structurale, et il est normal que toute une série d'ambiguités soient à leur tour levées. Mais il n'est pas besoin d'entrer dans les détails de l'analyse de Chomsky pour sentir qu'on n'a pas levé les difficultés, et que si on veut les lever, on va être amené peu à peu à monter de plus en plus loin dans l'échelle des ensembles de base E, car il n'est pas de structure grammaticale finie décrite de la sorte, qui ne fasse finalement pas appel à un moment ou à un autre au contexte, c'est-à-dire à l'ensemble de tout le langage, ensemble qui n'est pas défini au sens de la théorie des ensembles, et par conséquent qu'on ne peut pas formaliser complètement. Nous aboutissons par conséquent avec les travaux de Chomsky à une clarification des difficultés, mais non pas à une résolution des difficultés. Mais par contre on aboutit avec la technique des transformations linguistiques à mettre à jour certaines possibilités de reclassement systèmatique des phrases qui permet de les traiter peut-être aisément par les machines.

Le caractère nécessairement limité de la linguistique mathématique certaine nous conduit donc à sortir du cadre purement "certain" et à examiner les rapports avec la linguistique aléatoire. Si vous réfléchissez sur la façon dont vous explicitez une notion que vous ne connaissez pas en vous servant d'un dictionnaire, vous avez un très bon point de départ pour voir la limitation de la linguistique mathématique certaine et voir aussi le moyen d'en sortir. En effet quand on ne connaît pas le sens d'une notion, on se reporte à un dictionnaire, et ce dictionnaire qu'est-ce que c'est, sinon une transformation qui fait partir d'un mot A et qui donne une certaine phrase B, cette phrase étant elle-même un ensemble de mots. A partir des différents mots de cette phrase et en laissant tomber éventuellement les mots qui n'ont qu'un aspect syntaxique, on a encore un développement possible et le tout est exprimental bien entendu; c'est ce qu'on peut appeler le principe de l'amplification sémantique. Cette amplification est en principe finie puisque le nombre de notions, de mots, contenus dans un dictionnaire, est fini. En fait elle ne l'est pas réellement, puisqu'il y a des bouclages et cela fait l'objet de plaisanteries bien connues sur les dictionnaires. On définit les notions les unes au moyen des autres et au bout d'un moment on a bouclé, et la boucle peut être extrêment compliquée. Elle peut contenir un très grand nombre d'éléments.

Prenez deux mots qui sont complètement différents du point de vue morphologique et effectuez le développement sémantique de ces deux mots. Il est clair qu'au départ les deux diagrammes de développement sont fort différents et il est possible qu'au bout d'un moment ils deviennent de plus en plus voisins et que à partir du moment où ils ont des milliers d'éléments ils soient pratiquement voisins à certaines fluctuations près. L'existence de cette fluctuation correspond justement au flou dont est entaché finalement toute notion sémantique et qui fait qu'on ne peut pas par des procédés finis et certains avoir à la fois la structure et le sens, et comme la structure dans un certain nombre de cas ne se détermine complètement qu'en fonction du sens, tout système fini ne peut pas cerner le langage ni du point de vue sémantique, ni par conséquent du point de vue syntaxique. Mais sera-t-il possible de formaliser aussi ce flou, cette fluctuation? Sera-t-il possible de mécaniser la recherche de cette information compte tenu d'un certain flou? Je crois que oui, nous essayerons de donner des arguments en faveur de cette thèse; de toute façon les travaux de la linguistique mathématique certaine seront utilisés comme un des cas extrêmes d'étude de ces structures fluctuantes, l'autre cas extrême étant alors la considération du langage comme étant quelque chose de complètement statistique et de dépourvu de structure, pratiquement dépourvu de structure. En somme la situation du langage naturel est un peu la situation disons de l'état solide de la matière ou plutôt de l'état colloïdal de la matière, alors que la linguistique mathématique certaine correspondrait à l'hypothèse que la matière se trouve cristallisée au zéro absolu, sans aucune espèce de fluctuation et la linguistique aléatoire que le langage se trouve dans l'état d'un gaz parfait. C'est un bien en effet des notions de la thermo-dynamique des gaz parfaits qui sont utilisés en linguistique aléatoire dans les travaux de Zipf, Mandelbrot de Belevitch et d'autres auteurs. C'est la structure algébrique figée qui est étudiée par Chomsky-Kulagina. Les véritables langages se trouvent entre les deux, et bien entendu pour cerner ce problème, il est bon d'attaquer des deux côtés, d'essayer de se rejoindre, c'est ce que nous essayerons de faire dans les années qui viendront

BIBLIOGRAPHIE COMMENTEE

(dans l'ordre des paragraphes qui précédent)

1. On trouve un exposé très complet sur la linguistique générale dans ses développements les plus modernes dans:

A. MARTINET — Eléments de linguistique générale
Armand Colin Paris 1961

On pourra consulter aussi, pour la déscription de la "glossématique":

B. SIERTSEMA — A study of glossematics
Martinus Nijhoff Den Haag 1955

2. Les éléments mathématiques se trouvent dans le traité bien connu de N. BOURBAKI

Hermann Paris 1940 - 1961

Mais l'essentiel de ce que l'on aura à utiliser ici se trouve dans

C. BERGE — Théorie des graphes — Dunod Paris 1958
C. BERGE — Fonctions multivoques
Espaces topologiques — Dunod Paris 1959

3. Pour les modèles binaires, les textes fondamentaux sont:

N. CHOMSKY — "Syntactic Structures", Mouton and Co. 'S-Gravenhage. 1957

L. TESNIERE — "Eléments de syntaxe structurale. Klincksiek, Paris 1959.

D.G. HAYS — "Basic principles and technical variations in sentence structure determination". 4th London Symposium on Information Theory, 2 sept. 1960

Un important modèle binaire est également le suivant:

S. CECCATO — "Principles and classification of an operational grammar for Mechanical Translation". International Conference for Standards on Common Language for Machine Searching and Translation, Cleveland, September 1959.

4. Ici les textes fondamentaux sont

O. KULAGINA — Sur une méthode de définition des notions grammaticales à l'aide de la théorie des ensembles

Problème Kibernetiki Moscou 1958 p -203

A. SESTIER — Contribution à une théorie ensembliste des classifications linguistiques.

Communication au 1er congrès de l'AFCAL Grenoble 1960

Y. LECERF — "Une représentation algébrique de la structure des phrases dans diverses langues naturelles".
Notes aux comptes rendus de l'Académie des Sciences, Paris 1961, 232 n°2 page 232

Y. LECERF — "Programme des conflits, modèle des conflits", bulletin bimestriel de l'Association pour l'étude et le développement de la traduction automatique et de la linguistique appliquée (Atala) n°4 (octobre 1960) n°5 (décembre 1960)

5. La linguistique statistique "thermodynamique" est illustrée notamment par

B. MANDELBROT — Linguistique statistique macroscopique Institut Henri Poincaré. Paris 1957

Voir également

G. HERDAN — Language as choice and chance

La limitation de ces modèles est bien soulignée par

W. MEYER - EPPLER — Analogies physiques des structures linguistiques Physikalische Blätter 1955 p.445

ON THE STATISTICAL LAWS OF LINGUISTIC DISTRIBUTIONS (*)

by V. BELEVITCH (x)

1. RANK-FREQUENCY DIAGRAMS

It is an experimental fact that by counting frequencies of occurrence of various elements (letters, phonemes, words) in homogeneous texts written in a given language, and dividing them by the total number of elements of the same nature in the text, one often obtains relative frequencies that are stable (independant of the length of the text for sufficiently long texts). These relative frequencies define the a priori probabilities of the elements. Having used texts to obtain probabilities of various linguistic elements, one constructs catalogues of elements of identical nature (i.e. alphabet for letters, lexicon for words, etc.) in which each element is listed with its probability, by order of non-increasing probabilities.

In normal statistical practice one defines discrete distributions by specifying the number N_i of elements having a dimension or some other measurable characteristic x_i. If $N = \Sigma N_i$ is the total number of elements, the ratio $f_i = N_i/N$ is the relative frequency, or probability of finding the value x_i for the dimension x. The cumulative probability is defined by

$$\varphi_i = \sum_{k=1}^{i} f_k \qquad (1)$$

and distribution curves are obtained by plotting the cumulative probability φi versus the dimension x_i.

Most linguistic elements (f.i. letters) have no measurable characteristic (dimension) according to which they could be ordered, except their probabilities themselves. As there is no point in defining a distribution curve by a text probability versus a text probability,

---

(*) M. Belevitch nous a donné l'autorisation de reproduire le texte qui était précédemment paru dans "Annales de la Société Scientifique de Bruxelles, Tome 73, n° III, p. 310-326" et qui correspond à ce qui a été exposé lors de l'enseignement.

(x) Comité d'Etude et d'Exploitation des Calculateurs Electroniques "C.E.C.E.", Bruxelles.

the only alternative is to plot the probability in the catalogue f_i versus the probability in the text p_i. The probability in the catalogue is defined by reference to experiments with an urn containing once each element i, marked with its text probability p_i: the catalogue probability f_i is the probability of drawing from the urn the text probability p_i. For a catalogue of N elements, N_i of which have the text probability p_i, f_i is the ratio N_i/N, and the cumulative probability (1) in the catalogue is

$$\varphi_i = \frac{1}{N} \sum_{k=1}^{i} N_k \qquad (2)$$

But the sum $\sum N_k$ in (2), i.e. the number of elements of text probabilities $\leq p_i$, il precisely the rank i in the catalogue, since elements are ranged in order of non-increasing frequencies. As a consequence (2) becomes $\varphi_i = i/N$ and gives the relative rank in the catalogue. The distribution curves thus defined only differ from the usual rank-frequency diagrams, were the rank of each element in the catalogue is plotted versus its probability in the text, by the factor $1/N$ transforming absolute rank into relative rank. As a conclusion, rank-frequency diagrams can be interpreted as ordinary distribution curves giving the cumulative probability in the catalogue versus the probability in the text. This establishes a relation between the paradigmatic and syntagmatic aspects of the language.

Information theory attributes to each element of probability p_i an information measure (negative entropy) $x_i = - \log p_i$, and it is therefore convenient to use logarithmic scales for text probabilities. When a logarithm of base 2 is used in the above definition, the information is measured in bits. In the analytical expressions, it is more convenient to use natural logarithms; this is equivalent to adopt $\log_2 e = 1{,}44$ bits as natural unit of information (binit).

In general statistics, the range of the independent variable x is unrestricted. In statistical linguistics, x is related to a probability by

$$x = - \log p \qquad (3)$$

and is essentially positive. An additional restriction results from the closure condition

$$\sum_{1}^{N} N_i \, p_i = 1 \qquad (4)$$

which is transformed into

$$\sum N_i \, e^{-x_i} = 1 \qquad (5)$$

or, since the probability in the catalogue was defined as $f_i = N_i/N$, into

$$\sum f_i \, e^{-x_i} = \frac{1}{N} \tag{6}$$

As a conclusion, the closure condition determines the absolute number of elements in the catalogue from their relative distribution law.

2. MEAN VALUES

When dealing with mean values, one should clearly distinguish between averages over the text and averages over the catalogue. The mean m and the variance σ^2 as defined by

$$m = \frac{\sum N_i \, x_i}{N} \; ; \; \sigma^2 = \frac{\sum N_i \, (x_i - m)^2}{N} \tag{7}$$

are averages over the catalogue. In the text, the number N_i of elements of measure x_i must be weighted proportionally to their probability of occurrence p_i, and the average information is

$$h = \frac{\sum N_i \, p_i \, x_i}{\sum N_i \, p_i}$$

Since the denominator is unity, this gives the mean information

$$h = - \sum N_i \, p_i \log p_i \tag{8}$$

in the sense of information theory. Finally (8) becomes

$$\frac{h}{N} = \sum f_i \, x_i \, e^{-x_i} \tag{9}$$

We now consider the case where the range of x, for which f takes significant values, is sufficiently small, i.e. narrow distributions, so that the function e^{-x} can be approximated by the Taylor expansion

$$e^{-x} = e^{-m} \left[1 - (x - m) + \frac{1}{2} (x - m)^2 \right] \tag{10}$$

around the mean m. Condition (6), combined with (7), yields

$$e^{-m} \, (1 + \sigma^2/2) = 1/N$$

Similarly, in (9), the expansion of $x\,e^{-x}$ is deduced from (10) by writing $x = (x - m) + m$ and neglecting the third order term. One finds

$$xe^{-x} = me^{-m}\left[1 - \left(\left(1 - \frac{1}{m}\right)(x - m) + \left(\frac{1}{2} - \frac{1}{m}\right)(x - m)^2\right]\right.$$

and (9) becomes

$$\frac{h}{N} = me^{-m}\left[1 + \left(\frac{1}{2} - \frac{1}{m}\right)\sigma^2\right] \tag{12}$$

Since σ^2 is small for a narrow distribution, (11) becomes approximately

$$m = \log N + \frac{1}{2}\sigma^2 \tag{13}$$

On the other hand, the ratio of (12) and (11) gives, with the same approximation,

$$h = \log N - \frac{1}{2}\sigma^2 \tag{14}$$

or

$$h = m - \sigma^2 \tag{15}$$

If all N elements have the same text probability $1/N$, the information is the constant $\log N$. By comparison, formulae (13) and (14) show that the effect of a small spread in the probabilities is to reduce the average information in the text (and this is well known from information theory) and to increase by the same amount the average over the catalogue. Furthermore, the variance of the distribution is precisely the difference between both averages.

3. TRUNCATED NORMAL DISTRIBUTIONS

It is often convenient to approximate discrete distributions with a large number of elements by continuous distributions. The number of elements of dimension comprised between x and $x + dx$ is $N\,d\varphi(x) = N\varphi'(x)\,dx = N\,f(x)\,dx$, where $\varphi(x)$ is the distribution function and $f(x) = \varphi'(x)$ the probability density. A Gaussian, or normal, distribution of mean m and variance σ^2 id defined by the probability density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{(x - m)^2}{2\sigma^2}} \tag{16}$$

The range of the normal distribution is $-\infty \leq x \leq \infty$ and, since x is essentially positive in linguistic applications, the distributions cannot be rigorously normal. In the following we will consider truncated normal distributions where (16) is restricted to some positive interval $x_a \leq x \leq x_b$, the density assuming the value 0 outside. We will start, however, by the case where the truncation is made at points sufficiently far away from the mean, so that the tails of the distribution are negligible anyway.

For a continuous distribution, conditions (6) and (9) are replaced by

$$\frac{1}{N} = \int_{x_a}^{x_b} f(x) e^{-x}\, dx; \quad \frac{h}{N} = \int_{x_a}^{x_b} x e^{-x} f(x)\, dx \qquad (17)$$

In these expressions also, we first replace the integration limits by $+\infty$, neglecting the truncation. Absolute convergence is still ensured, for the increase of $f(x)$ and e^{-x} for $x = -\infty$ is less rapid than the decrease of (16).

For the density (16), the integrals (17) are reduced to the error integral by transforming the exponent according to

$$x + \frac{(x - m)^2}{2\sigma^2} = \frac{(x + \sigma^2 - m)^2}{2\sigma^2} + m - \frac{\sigma^2}{2} \qquad (18)$$

By the linear transformation

$$z = \sigma + \frac{x - m}{\sigma} \qquad (19)$$

the first equation (17) becomes

$$\frac{1}{N} = \frac{e^{-m + \sigma^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2}\, dz = e^{-m + \sigma^2/2}$$

and is equivalent to (13). The second equation (18) becomes

$$\frac{h}{N} = \frac{e^{-m + \sigma^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (m - \sigma^2 + \sigma z)\, e^{-z^2/2}\, dz$$

The term in Z in the integrand is an odd function and does not contribute to the integral, so that (21) reduces to

$$\frac{h}{N} = (m - \sigma^2)\, e^{-m + \sigma^2/2} \tag{22}$$

and, by comparison with (20), one obtains (15).

The remaining part of this section is devoted to the derivation of the rigorous formulae replacing (20) and (22) and taking the truncation into account. For the error integral, we will use the notation

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-u^2/2}\, du \tag{23}$$

so that the distribution function corresponding to the density (16), with neglected truncation, is

$$\varphi(x) = \Phi\left(\frac{x - m}{\sigma}\right) \tag{24}$$

If truncation is taken into account, the density cannot be (16), but must be corrected by a factor $a > 1$ in order to normalize to unity the integral in the finite range. Similarly the distribution function is no longer (24) but is of the form

$$\varphi(x) = a\,\Phi\left(\frac{x - m}{\sigma}\right) - b \tag{25}$$

deduced from the density (16) multiplied by a, with an integration constant denoted - b. These constants are determined by the conditions $\varphi(x_a) = 0, \varphi(x_b) = 1$, and one obtains

$$a = \frac{1}{\Phi\left(\frac{x_b - m}{\sigma}\right) - \Phi\left(\frac{x_a - m}{\sigma}\right)}\,; \quad b = \frac{\Phi\left(\frac{x_a - m}{\sigma}\right)}{\Phi\left(\frac{x_b - m}{\sigma}\right) - \Phi\left(\frac{x_a - m}{\sigma}\right)} \tag{26}$$

For the truncated distribution, the integrals (17), applied to the density a f(x), become

$$1/N = ace^{-m + \sigma^2/2} \tag{27}$$

$$\frac{h}{N} = ae^{-m + \sigma^2/2} \left[c(m - \sigma^2) + \frac{\sigma}{\sqrt{2\pi}} \left(e^{-z_a^2/2} - e^{-z_b^2/2} \right) \right] \qquad (28)$$

where z_a and z_b have the values resulting from (19) with $x = x_a$ or x_b, and where c denotes the error integral with the limits z_a and z_b, thus

$$c = \Phi \left(\sigma + \frac{x_b - m}{\sigma} \right) - \Phi \left(\sigma + \frac{x_a - m}{\sigma} \right) \qquad (29)$$

In (25), the term b introduces a correction at high probabilities, which is negligible in most applications because $x_a - m$ is negative and equals several times σ. On the contrary, the effect of the truncation at low frequencies is often not negligible, for statistics do not generally extend sufficiently far above the mean. The practical form of (27) is thus deduced from the approximation $x_a = -\infty$, and is

$$N = e^{m - \sigma^2/2} \frac{\Phi \left(\frac{x_b - m}{\sigma} \right)}{\Phi \left(\sigma + \frac{x_b - m}{\sigma} \right)} \qquad (30)$$

In a complete statistical count, the lowest frequency corresponds to a single occurrence in the text, thus to probability $p_b = 1/L$, where L is the length of the text. By (3), one has

$$x_b = \log L \qquad (31)$$

and (30) gives a relation between the length of the text and the extension of the vocabulary. It is obvious that (30) reduces to (13) for $x_b = \infty$

4. APPROXIMATIONS TO NORMAL DISTRIBUTIONS

We consider first an arbitrary distribution function $\varphi(x)$ in the neighbourhood of a point x_0. The Taylor expansion is

$$\varphi(x) = \varphi(x_0) + (x - x_0) f(x_0)$$

where f (x) is the corresponding probability density. The logarithmic slope of distribution function is approximately

$$\log \frac{\varphi(x)}{\varphi(x_0)} = \log \left[1 + \frac{f(x_0)}{\varphi(x_0)} (x - x_0) \right] \cong (x - x_0) \frac{f(x_0)}{\varphi(x_0)} \qquad (32)$$

If one considers an element $x = x_i$ corresponding to a text probability $p_i = e^{-x_i}$, its rank i is given by $N\varphi(x_i)$. Introducing the similar notations p_o and i_o for the reference point x_o, (32) becomes

$$\log \frac{p_i}{p_o} = -A \log \frac{i}{i_o} \tag{33}$$

with

$$A = \frac{\varphi(x_o)}{f(x_o)} \tag{34}$$

Equation (33) is independent from any assumption on the distribution law, and merely shows that (34) measures the slope of the tangent, at x_o, to the rank-frequency characteristic with logarithmiscales for both coordinates.

Expression (33), or

$$\frac{i}{i_o} = \left(\frac{p_i}{p_o} \right)^{-1/A} \tag{35}$$

is similar to Zipf's law, but with a variable slope. By taking into account second order terms in the Taylor expansion, it is possible to obtain a correction similar to the one introduced in Zipf's law by Mandelbrot, i.e. to arrive at a form

$$\frac{i}{i_o} = s \left(\frac{p_i}{p_o} \right)^{-1/B} - t \tag{36}$$

instead of (35), or, equivalently, to

$$p_i = P\,(i + \rho)^{-B} \tag{37}$$

with

$$P = p_o\,(i_o s)^B ; \rho = i_o t \tag{38}$$

The best values of the parameters are obtained by identifying the second-order Taylor expansion of the right-hand side of (36), i.e.

$$se^{(x_i - x_o)/B} - t = s - t + s\,\frac{x_i - x_o}{B} + \frac{s(x_i - x_o)^2}{2B^2}$$

with the corresponding expansion of the distribution function

$$\frac{\varphi(x_i)}{\varphi(x_o)} = 1 + \frac{(x_i - x_o)\, f(x_o)}{\varphi(x_o)} + \frac{(x_i - x_o)^2\, f'(x_o)}{2\,\varphi(x_o)}$$

This gives

$$s = \frac{[f(x_o)]^2}{\varphi(x_o)\, f'(x_o)}\,;\quad t = s - 1 \tag{39}$$

$$B = \frac{f(x_o)}{f'(x_o)} \tag{40}$$

and one has $t > 0$ as long as the curvature of the characteristic at x_o is positive.

For a normal distribution (truncated or not), (40) becomes

$$B = \frac{\sigma^2}{m - x_o} \tag{41}$$

Simple expressions for the other parameters are only obtained if some approximations are introduced, and truncation neglected. For the linguistic applications it is of special importance to discuss the behavior of the characteristics at high frequencies, where the statistics are the most reliable. We will thus assume $x_o \ll m$ and use the asymptotic expansion ([1])

$$\Phi(-u) = \frac{e^{-u^2/2}}{u}\left(1 - \frac{1}{u^2} + \cdots\right) \tag{42}$$

valid for large positive values of u. When the first term alone of (42) is considered, the value (41) is obtained for the exponent (34) of the Zipf approximation. With two terms in (42), (39) gives

$$s = 1 + \frac{\sigma^2}{(m - x_o)^2}\,;\quad t = \frac{\sigma^2}{(m - x_o)^2} \tag{43}$$

([1]) The possibility of deducing Mandelbrot's law with $B \leq 1$ from the asymptotic expansion of the error integral was mentioned to the writer by A. OETTINGER, and originated the present investigation.

By (15), (41) becomes

$$B = \frac{\sigma^2}{\tau^2 + h - x_0} \qquad (44)$$

When the approximating point x_0 is the point of highest frequency x_a, (44) gives $B \leq 1$, since the average information h is certainly larger than the minimum information x_a. When x_0 increases starting from x_a, B increases and passes through the value 1 for $x_0 = h$.

5. LETTER AND PHONEME DISTRIBUTIONS

Because of the small size of the alphabets, such distributions are relatively irregular, but definite systematic trends can, however, be noticed. On the other hand, thanks to the small size of the alphabets, the statistics are more reliable, and the distributions are well known over their entire ranges. Fig. 1 shows data from a number of languages ([2] the lower scale is logarithmic in p, and

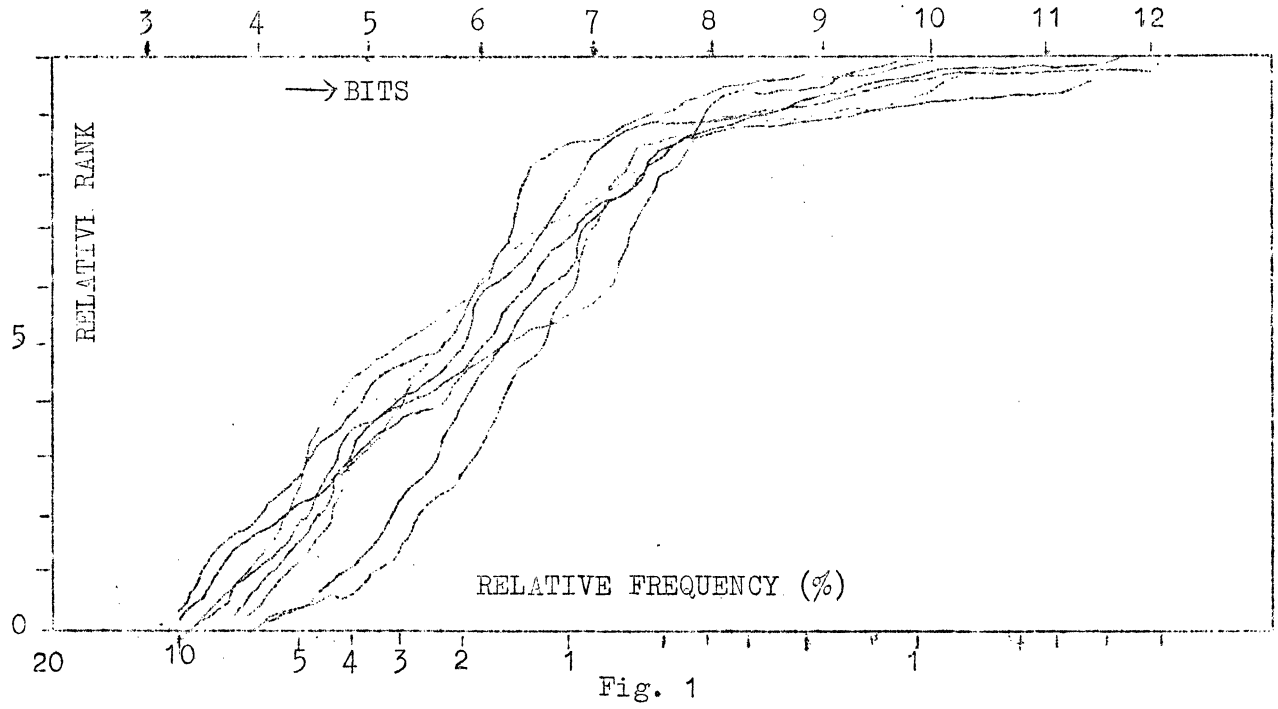


Fig. 1

([2]) Most of these have already been discussed in V. BELEVITCH "Théorie de l'information et statistique linguistique", Bull. Acad. Roy. Belg. (Cl. des Sc.) avr. 1956 pp. 419-436.

the upper scale is linear in bits. It appears immediately that all curves are very similar, and all ranges extend from approximately 2,5 to some 11 bits. All distributions are relatively narrow, and it is therefore expected by (13) that the horizontal shift between the various curves is correlated with the size of the alphabet (varying from N = 21 to 49 in the examples considered). This is indeed the case, as it appears from fig. 2 where each curve has been shifted by log N; in other words, the abscissa in fig. 2 is the relative frequency p/p_m with respect to the equiprobable value $p_m = 1/N$. Fig. 2 clearly shows that all distributions are practically identical and, in particular, have the same variance. Our best estimate of the common value is σ = 1,4 bits. The corresponding normal distribution shifted by $\sigma^2/2$ relatively to the point of abscissa log N in accordance with (13), is the dotted curve of fig. 2.

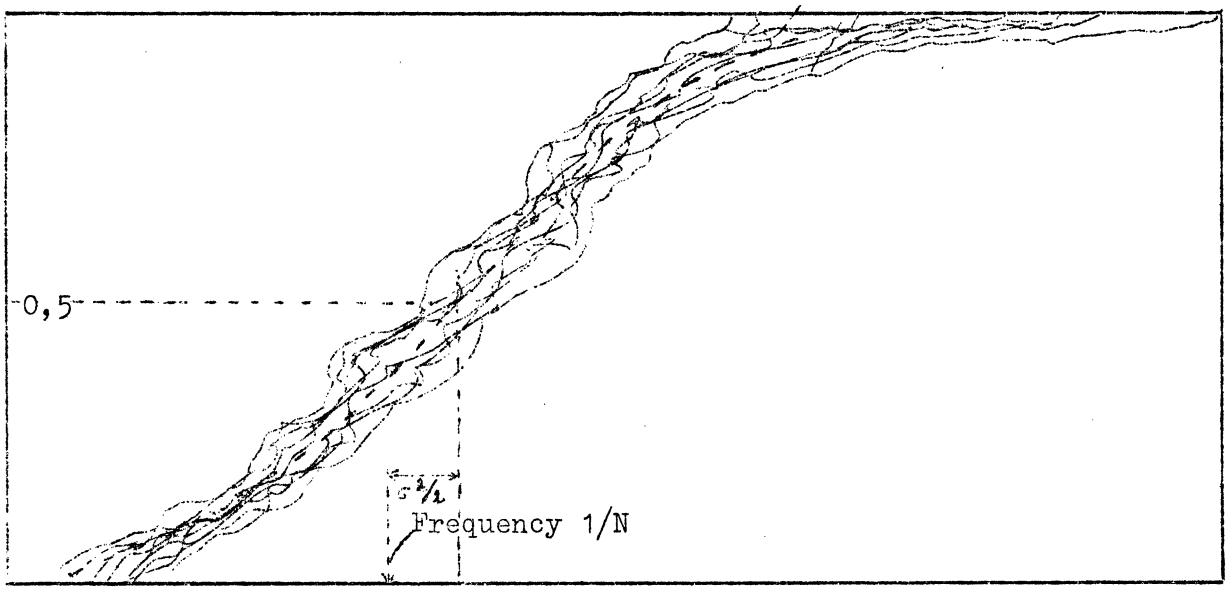


Fig. 2

Formulae (13-15) can be checked on the example of Russian phonemes for which particularly detailed data are available (3). The value of log N is $\log_2 42 = 5,42$ bits, and the published value of h is 4,78 bits. From (15), one obtains σ = 1,33 bits, and from (13), m = 6,1 bits; the last value agrees with the median of the distribution curve deduced from the published data.

---

(3) E.C. CHERRY, M. HALLE, R. JAKOBSON, "Toward the logical description of Languages in their phonemic aspect", Language, vol. 29 n° 1, p. 34, March 1953.

WORD DISTRIBUTIONS

According to Guiraud (4), Zipf's formula, and even Mandelbrot's correction, do not agree with most experimental distributions at low frequencies. The truncated lognormal character of the actual distributions seems to have been suspected by several authors (5). If truncation is neglected, a lognormal distribution

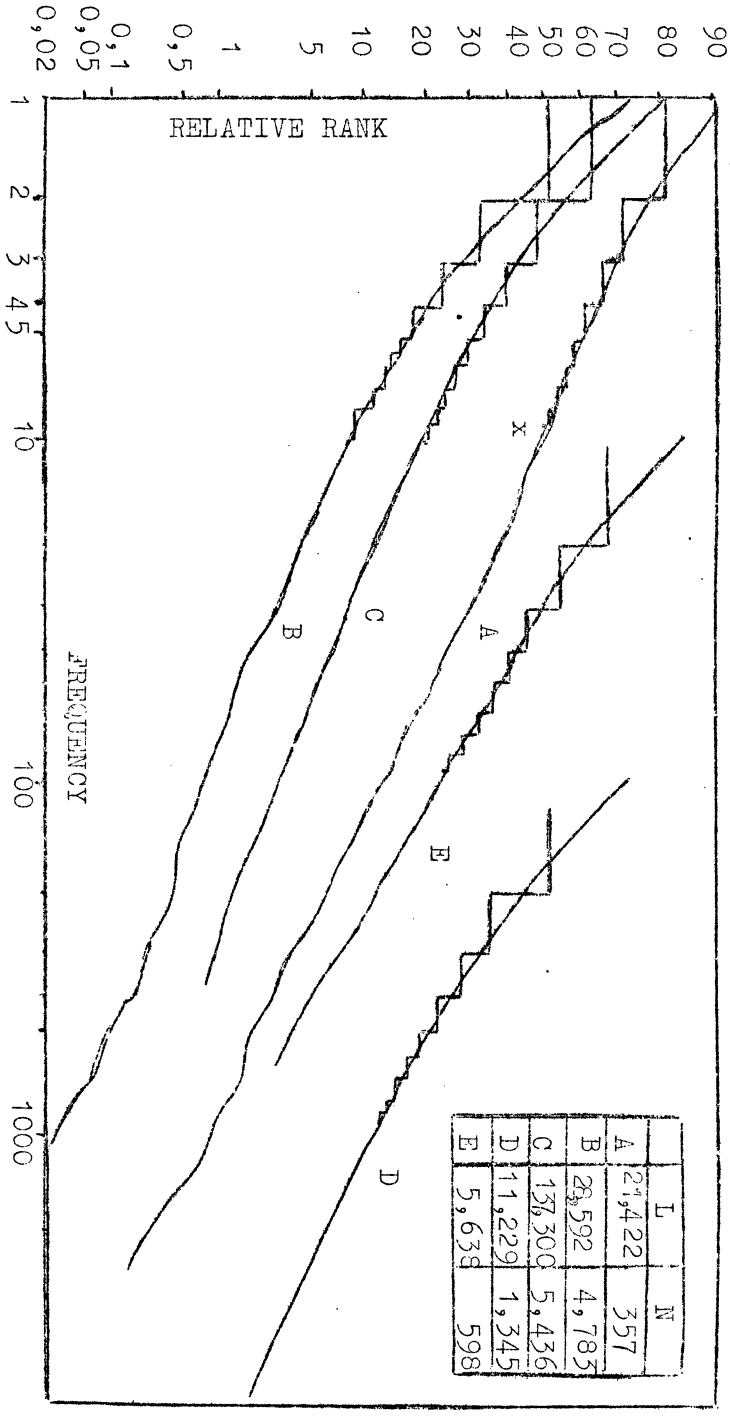


Fig. 3

(4) P. GUIRAUD, Les caractères statistiques du vocabulaire, Paris, Press. Univ. 1954.

(5) See f.i., AITCHISON, J.A.C. BROWN, The lognormal distribution, Cambridge 1957; p. 101.

becomes a straight line on lagarithmic probability paper; this is checked for a few examples on Fig. 3, and it will be established herebelow that the deviations from linearity at low frequencies are quantitatively explained by the truncation effect.

The abscissae in fig. 3 are the absolute frequencies, as published in various sources ( ), but curve E has been shifted by a factor 10 and curve D by a factor 100, to avoid overlapping. The table included in fig. 3 gives the extensions of the vocabulary (N) and of the text (L) mentioned in the source material. The discrete steps at the low frequency ends of the curves arise because the numbers of occurrences of the rarest words are necessarily small integers. Strictly, the usual definition of the rank-frequency relation would yield a continuous curve passing through the top points of the ladder, but the similar relation based on the complementary rank gives a curve passing through the bottom points. A unique smoothed continuous distribution curve can therefore only be defined by joining the vertical mid-points of the ladder, and this has been done in fig. 3. In particular, the first point of the smoothed distribution, corresponding to frequency 1, is $1 - N1/2N$, where $N1$ is the number of hapaxlegomena and N the total number of different words in the sample.

For curve A, the truncation effect is practically negligible because the statistics extends well above the mean. The slope of the straight line (abscissa interval corresponding to a decrease of the ordinate from 50 to 16%) defines the standard deviation as $\sigma = 2,8$ bits. From this value, and the value of N mentioned in the table, one can compute m by (13), and the corresponding median frequency is $L\, e^{-m}$; the value thus obtained is shown by a cross on fig. 3. Fig. 4 shows the curve of fig. 3A in bilogarithmic coordinates, and the dotted straight lines are Zipf's approximations at $x = x_a$ and $x = h$, the values of the slope being computed by (44), and the value of h by (15).

The truncation effect at the low frequency end will be discussed on the example of curve B of fig. 3. The corresponding bilogarithmic representation of fig. 5 shows that the value of the initial slope

---

(6) Curve A is based on data for English function words (G.A. MILLER, E.B. NEWMAN, E.A. FRIEDMAN, "Length frequency statistics for written English", Inform. and Control, vol. 1 n° 4 pp. 370-389; Dec. 1958). Curve B corresponds to Russian words in the Captain's Daughter by Pushkin (H.H. JOSSELSON, The Russian word count, Detroit, 1953). Curves C (New testament) and D (St Mark) are based on R. MORGENTHALER, Statistik des neutestamentlichen Wortschatzes. Zurich 1958. Curve E is for French adjectives from all 8 tragedies of Racine (ref. note 4 p. 31).

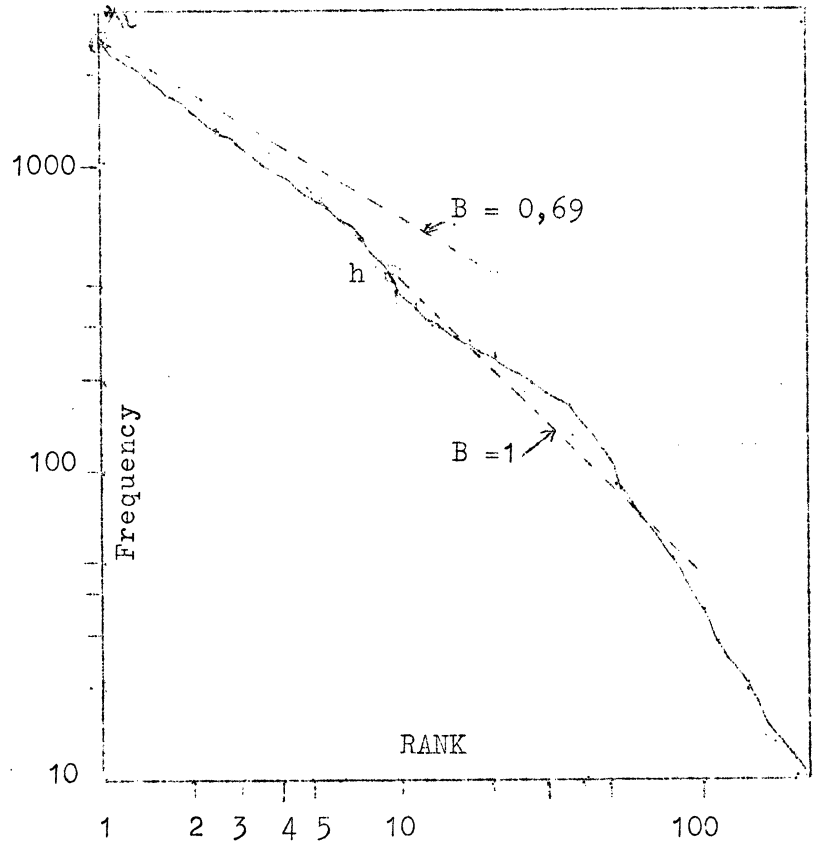


Fig. 4

is 0,48. Since the experimental value of x_a is 3,23 binits, (41) requires

$$0,48\ (m - 3,23) = \sigma^2 .$$

A second relation between m and σ is (30), for N is known and the experimental value of x is 10,9 binits. The solution of these equations is m = 11,12 binits = 16,1 bits and σ = 2,8 bits. The correction factor in (25) is

$$1/a = \left(\frac{x_b - m}{\sigma} \right) = 0,455$$

and this transforms curve B into curve B' as shown on the right hand side of fig. 6. The cross corresponds to the computed value of m. This shows that the truncation effect completely accounts for the curvature at the low frequency end.

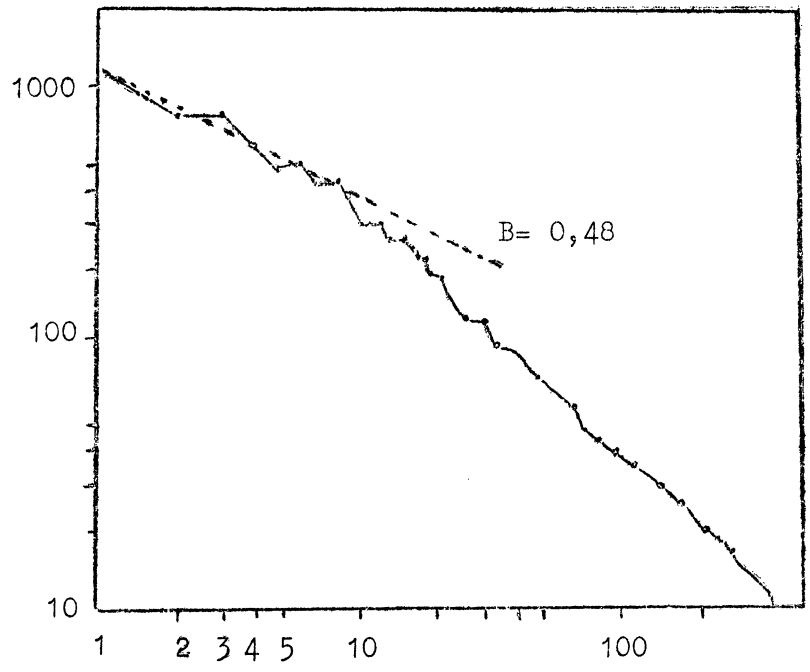


Fig. 5

The various examples of fig. 3 are represented in the left hand side of fig. 6 with an horizontal shift identical to the one discussed in passing from fig. 1 to fig. 2. The smoothed curves are still called A, B ... E, but the discrete steps have been omitted. It is apparent that all curves will become practically identical after correction for the truncation effect. This correction does not alter the slope of the linear part, and it is already obvious in fig. 3 that all distributions have practically the same slope. The common value seems to be 2,8 bits, which is the double of the value found for phonemes. The straight line in fig. 6 is the theorical characteristic corresponding to σ = 2,8 bits.

If one accepts the normal distribution as the general law for words, the fact that Mandelbrot's or Zipf's laws are often satisfactorily confirmed would simply result from the enormous extension of the vocabularies combined with the limitation of many statistics well below the mean rank : for m large, B, as given (41), is constant for all moderate values of x_0. The necessity of a large vocabulary would also explain the insistence of several authors in counting all inflected forms as distinct.

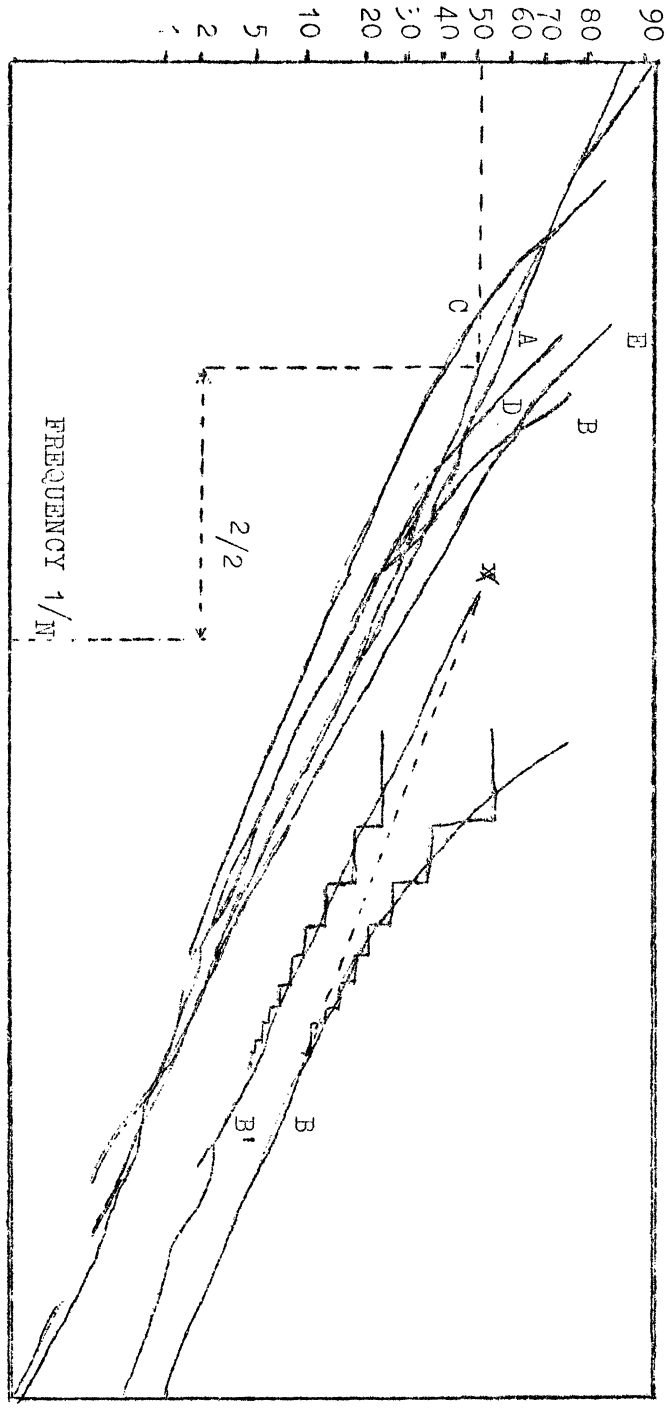


Fig. 6

A number of theorical models have been proposed by Mandelbrot to account for his law (7). The model based on the weakest hypothesis

(7) The publication of B. MANDELBROT, "Linguistique statistique macroscopique" in Logique, langage et théorie de l'information, Presses Univ. Fr., Paris, 1957, gives a non-mathematical account of his theory, and a bibliography of the subject. See also the more recent contribution of B. MANDELBROT, in Info and Control vol. 2 pp. 90-99; april 1959.

assumes that texts are separated into words by a randomly distributed "space" symbol. No special theory is needed to explain a normal distribution, but it must be remarked that randomness has been shifted from the text to the vocabulary. For the high frequency tail of the distribution, where the saturation effect due to the finite extent of the vocabulary is still negligible, both explanations are equivalent, since the same stochastic model can be interpreted as yielding the text by a dilution of the dictionary, or the dictionary by a concentration of the text. But, for finite vocabularies a difference arises, because texts remain potentially infinite by hypothesis.

Finally, the constancy of the standard deviation for phoneme distributions on one hand, and for word distributions on the other, would suggest some common discrete substructure for both linguistic levels, but with a double number of degrees of freedom in the latter case.

ACKNOWLEDGEMENT

The author is grateful to W. Croes and P.G. Neumann who criticized the manuscript.

TRAVAUX PRATIQUES DE LINGUISTIQUE

Y. LECERF

I N T R O D U C T I O N

A) NOTION DE CALCUL D'ADRESSES

Il ne peut y avoir, dans la mémoire d'un ordinateur, d'information sans adresse. Une fois emmagasinée, une information ne peut être utilisée que si l'on connait cette adresse, ou si l'on peut la retrouver indirectement au moyen d'un calcul. Lorsque les informations à exploiter sont nombreuses, on évite autant que possible de dresser une liste explicite de leurs adresses. Il est plus élégant de recourir au second procédé, celui du calcul d'adresses, si c'est possible.

Ainsi, le traitement en machine d'êtres linguistiques fait intervenir en réalité :

1. des êtres linguistiques proprement dits (mots, règles de grammaires, etc...)
2. leurs adresses
3. les repères (explicites ou implicites) en fonction desquels on calcule ces adresses.

B) AMBIGUITES DU LANGAGE ORDINAIRE

On sait que les mots du langage ordinaire sont, pour la plupart, susceptibles de plusieurs significations. Pour départager celles-ci, il faut faire intervenir des renseignements relatifs au contexte, d'une part, et à un certain nombre de règles grammaticales, d'autre part. Par exemple, le mot "voile" prend des significations différentes dans "un voile" et "une voile". Pour savoir de quelle acception il s'agit dans chaque cas, on fera intervenir l'article (c'est-à-dire une certaine règle grammaticale). Dans les phrases "les voiles sont beaux" et "les voiles sont belles", la même ambiguïté peut être résolue en faisant intervenir l'attribut (c'est-à-dire le contexte) et la règle d'accord en genre du sujet avec son attribut (c'est-à-dire une règle grammaticale différente de celle utilisée plus haut). L'expression "les voiles de notre bateau" est interprétable à partir du contexte du mot "voiles" ; d'un corpus de renseignements sémantiques, et d'une règle sémantique d'utilisation de ce corpus. Enfin, lorsque le mot "voile" est utilisé en tant que forme verbale ce sont d'autres règles qui doivent intervenir, à partir d'autres contextes.

Une deuxième source d'ambiguïté du langage ordinaire apparait lors de l'interprétation des rapports entre mots. Même en supposant connues et fixées les interprétations de tous les mots dans la séquence : "les fils de fonctionnaires morts à la guerre" il reste à déterminer qui, des fonctionnaires ou de leurs fils, est mort à la

guerre. La phrase : "les filles de fonctionnaires morts à la guerre" montre que dans certains cas, l'ambiguïté peut être levée par consultation de règles normatives (ici, l'accord en genre de l'adjectif). Des règles sémantiques peuvent également être appelées à intervenir.

C) LA LEVEE DES AMBIGUITES DU LANGAGE ORDINAIRE POSE DES PROBLEMES DE CALCUL D'ADRESSES.

Un premier problème de calcul d'adresses se pose donc dans les termes que voici : étant donnée une phrase du langage ordinaire, d'une part, et un ensemble de règles normatives (grammaire, sémantique etc...) d'autre part, trouver les adresses des règles qui permettent d'interpréter la phrase donnée.

Second problème : les règles grammaticales ou sémantiques sont en général formulées sans indication des adresses des éléments liés par elles. Soit par exemple la règle sémantique selon laquelle les verbes représentant des actions d'êtres animés ont en général des êtres animés pour sujet. Cette règle permet de départager les significations du mot facteur dans des séquences telles que : "le facteur rit", car elle montre qu'il pourrait difficilement s'agir d'un facteur mathématique (en allemand Faktor), et que l'interprétation facteur préposé des postes (en allemand Briefträger) est plus vraisemblable. Mais cette règle sémantique ne donne pas les positions respectives, dans les phrases, de l'agent et du verbe indiquant l'action. Or ces positions sont extrêmement variables ; on peut intercaler une très grande variété de configurations entre ces deux éléments : exemple : "Le facteur, chaque fois qu'on dit une plaisanterie, rit aux éclats". Il est exclu de donner la liste de toutes ces confirmations intermédiaires possibles : elles peuvent comporter un nombre quelconque de subordonnées, elles-mêmes quelconques, et des années ne suffiraient pas à leur énumération. Il serait souhaitable, donc, de pouvoir exploiter la règle ci-dessus dans sa forme naturelle, où les adresses de l'agent et de l'action ne sont pas données. Le même problème se pose à propos des règles grammaticales qui, dans leur forme habituelle, ne précisent pas les adresses des éléments liés. Ce qui vient d'être dit à propos de la règle sémantique : agent animé, verbe à sujet animé, pourrait être repris presque identiquement à propos de l'accord grammatical sujet - verbe. La grammaire précise qu'il y a accord en personne et en nombre, mais n'indique pas les adresses relatives du sujet et du verbe, et donne seulement certaines indications générales à ce sujet. Il en est ainsi pour la plupart des règles normatives concernant les langues naturelles : on ne les exprime simplement qu'à la condition de ne pas donner les rangs, dans l'ordre linéaire de la phrase, des éléments liés. Puisque ces règles sont habituellement énoncées ainsi, et sont plus simples ainsi, il faut nous en contenter ; mais le calcul d'adresses devra suppléer, dans chaque phrase particulière, à cette insuffisance,

car en machine une règle ne signifie rien et ne sert à rien sans les adresses des éléments qu'elle lie. Le calcul devra donc rétablir ces adresses. Tel est le second problème, celui de l'exploitation d'un catalogue de règles sans indication des adresses des éléments liés.

Il n'est pas question de traiter globalement, dans les présents travaux pratiques, ces problèmes de calcul d'adresses.

On se bornera à étudier quelques techniques qui jouent un rôle important dans certaines méthodes de calcul machine.

D) CALCUL D'ADRESSES ET THEORIES LINGUISTIQUES

On pourrait envisager de séparer de façon bien distincte ces deux activités, dont l'une consiste à rassembler des règles normatives (règles grammaticales, règles sémantiques), et l'autre, à exploiter ces règles, grâce à des calculs d'adresses. De la première activité, les grammairiens et sémanticiens auraient le monopole. La seconde serait l'affaire exclusive de mathématiciens, d'automaticiens, de spécialistes de calcul machine.

En fait, ces deux activités sont difficilement dissociables. D'une part, en effet, les techniciens du calcul automatique, qui ont le souci d'économiser le temps machine, sont conduits à surveiller les grammairiens et sémanticiens. Il y a en effet bien des manières possibles d'exposer les règles grammaticales d'une certaine langue, d'exposer des règles sémantiques : peu de sciences connaissent un état de division pire que celles du langage, et peu de discussions sont aussi âpres que celles où s'opposent des grammairiens ou des sémanticiens. Puisqu'il y a plusieurs manières d'exposer les règles normatives du langage, les ingénieurs ont tendance à remarquer que les unes coûtent cher en calcul machine, que les autres sont meilleur marché, et ils feront pression sur les grammairiens et sémanticiens pour réduire les temps et améliorer les rendements. Certains mathématiciens, jugeant que les travaux des grammairiens d'origine n'étaient pas suffisamment économiques ni rentables, se sont eux-mêmes attelés à la construction de grammaires mathématiques, tels Chomsky (1) (2), Kulaguina (3) et bien d'autres ; il existe également des traités de sémantique très mathématiques par leur forme, tels ceux de Carnap (4).

Inversement, des grammairiens et sémanticiens ont tendance à s'occuper de calcul d'adresses. Ils abordent en général ces questions par le biais de l'étude des processus mentaux de la parole ou de l'identification du langage. Si en effet les problèmes de la traduction automatique ont été bien formulés, et dans des termes suffisamment généraux, il est admissible de penser que ces mêmes problèmes ont dû être résolus par l'esprit humain ; que lorsque nous parlons, comprenons, traduisons, notre esprit effectue des calculs d'adresse, puisque ces calculs dont de toutes façons indispensables. L'étude de ces faits touche si profondément au langage, qu'il est naturel de la considérer comme relevant de la linguistique. L'apport des linguistes en matière de calcul d'adresses peut se révéler extrê-

nement précieux, même si ces linguistes, comme c'est par exemple le cas de Tesnière (5) ou Hjelmslev (6), ne pensaient pas à la traduction automatique. La parenté des "stemmas" décrits par le premier, avec les graphes utilisés par Harper et Hays (7) en traduction automatique, montre l'importance des travaux de Tesnière, relativement aux problèmes de calcul d'adresses.

Ainsi, il est bien difficile de tracer une frontière entre les calculs d'adresses en linguistique, qui font l'objet des présents travaux pratiques, et la linguistique elle-même.

P R E M I E R E P A R T I E

ZONE D'INFLUENCE DES MOTS D'UNE PHRASE

OBJET DE LA PREMIERE PARTIE

Dans cette partie, ainsi que dans la suivante, on se propose d'étudier certains procédés utilisables pour énoncer des adresses de groupes de mots, ou, plus précisément, pour énoncer les adresses de ces groupes de mots que les grammairiens et les sémanticiens appellent syntagmes.

Il est certes toujours possible d'énoncer cette adresse directement, en indiquant où le groupe commence et où il finit, mais cette notation conduit à manipuler des couples de nombres dans les calculs, et l'on peut trouver la chose malcommode, en comparaison du procédé qui va être exposé, et qui permet de mettre en correspondance un groupe de mots avec une seule adresse, celle d'un mot reconnu comme important dans ce groupe, et qui n'y occupe pas nécessairement une position d'extrémité.

Il s'agit donc d'un artifice de calcul ; mais c'est par des linguistes qu'il a été proposé ou suggéré en premier lieu.

Ces linguistes sont parfois d'accord sur le choix du mot important dans un syntagme (exemple : chaîne nominale), mais souvent aussi en désaccord (exemple : chaîne associée au verbe). Du point de vue du calcul, on choisira la solution la moins coûteuse, qui peut d'ailleurs varier selon la machine utilisée et selon la nature du problème particulier traité. Le choix d'une méthode de calcul d'adresses parmi plusieurs possibles et susceptibles de fournir le résultat cherché n'a de conséquences qu'à l'intérieur de la machine. Il relève de la seule autorité de l'ingénieur.

Le travail pratique ci-après (1e partie) pose justement le problème litigieux de la chaîne associée au verbe. On examinera les diverses possibilités, dont beaucoup sont suggérées par des linguistes.

Pour simplifier, par contre, les parties suivantes (2,3,etc..) on adoptera pour ces dernières, arbitrairement, les choix correspondant à la linguistique de Tesnière ou Harper et Hays (7).

DOMAINE D'UN MOT

A chaque mot d'un texte, on convient d'associer une zone d'influence s'étendant d'un seul tenant et incluant le mot considéré. On l'appellera domaine de ce mot. Celui-ci sera dit mot fondamental du domaine, par opposition aux autres mots qui pourraient éventuel-

lement y être aussi contenus.

exemple : Le Gouvernement britannique réclame une enclave de 122 milles carrés.

"Le Gouvernement britannique" constitue la zone d'influence du mot "Gouvernement".

"Une enclave de 122 milles carrés" constitue la zone d'influence du mot "enclave".

exemple : Les autorités craignent une nouvelle flambée de terrorisme.

"Une nouvelle flambée de terrorisme" constitue le domaine du mot "flambée".

On admet que l'existence d'une signification introduit des liens de hiérarchie entre les mots d'une phrase, et l'on désire se servir des domaines pour mettre ces liens en évidence.
Pour cela, on posera que le domaine d'un mot contient aussi les domaines des mots directement subordonnés, et l'on admettra jusqu'à preuve du contraire que cette exigence ne contredit pas l'hypothèse de domaines d'un seul tenant.

exemple : dans "une enclave de 122 milles carrés",

- milles est subordonné à enclave, puisque contenu dans son domaine.
- le domaine de "milles" c'est-à-dire : "122 milles carrés" est contenu dans le domaine de "enclave".

Mais l'inclusion et l'infériorité sont toutes deux des relations transitives. De ce fait, la zone d'influence d'un mot contiendra ce mot lui-même et l'ensemble de ses subordonnées directs ou indirects, ainsi que leurs domaines.

On se propose d'étudier la configuration et l'étendue des domaines des mots.

TRAVAIL PRATIQUE

Délimiter les zones d'influence de tous les verbes dans les phrases ci-dessous :

- "Le commissariat du Théâtre des Nations vient d'approuver, nous l'avons dit, les grandes lignes du programme de la saison qui débutera le 15 mars."

- "Les expériences ont montré qu'il n'était pas possible d'obtenir des photomésons sans la présence de L_iH qui absorbe sélectivement les photons de faible énergie".

- "Tandis que la brillance du zinc augmente pendant le processus de polarisation anodique, celle de l'argent ne croit qu'après l'interruption de courant".

- "Cet article contient une description des systèmes les plus courants dans lesquels on utilise des gaz pour dégivrer des appareils de réfrigération".

S E C O N D E P A R T I E

FRONTIERES D'UN DOMAINE

OBJET DE LA SECONDE PARTIE

Etude d'un second procédé utilisable pour énoncer l'adresse d'un groupe de mots en calcul automatique

FRONTIERES D'UN DOMAINE : SIGLES

Les frontières d'un domaine seront matérialisées par des signes tels que parenthèses, crochets, accolades, etc., que l'on appellera plus généralement sigles - Le discours étant linéaire, et les domaines d'un seul tenant, ces sigles seront au nombre de deux par domaine - Leur concavité sera toujours tournée vers l'intérieur du domaine dont ils sont frontières.

exemple : "l'académie de médecine souhaite la diffusion du vaccin antipoliomyélitique". On peut noter que :

a) la phrase entière constitue le domaine du mot "souhaité"; on l'entoure de deux accolades.
{l'académie de médecine souhaite la diffusion du vaccin antipoliomyélitique}

b) on peut entourer de crochets le domaine du mot "académie":
[l'académie de médecine]

c) de même pour le mot "diffusion":
[la diffusion du vaccin antipoliomyélitique]

d) au total il vient :
{[l'académie de médecine] [souhaite la diffusion du vaccin antipoliomyélitique]}

De ce fait, à chaque mot se trouveront associés deux sigles, ceux-là mêmes qui bornent le domaine dont le mot est mot fondamental. - Ils encadreront le mot de part et d'autre, mais sans être nécessairement situés en son voisinage immédiat.-
Leur concavité sera toujours tournée vers le mot associé.

exemple : dans la phrase précédente, on aurait pu associer deux parenthèses à chaque mot. Donnons-les pour quelques mots :
[(l') académie (de[médecine])]

Pour rappeler que les parenthèses sont associées à tel ou tel mot, on s'est servi jusqu'ici de sigles marqués : crochets pour les noms, accolades pour les verbes, parenthèses pour les autres mots.

De telles notations ne sont pas indispensables. On montrera plus loin que le mot associé constitue à lui seul une marque et suffit à caractériser la paire de sigles considérée.

TRAVAIL PRATIQUE

Délimiter par des sigles les frontières des domaines de tous les mots dans les phrases ci-dessous. On associera aux noms des crochets, aux verbes des accolades, et des parenthèses à tous les autres mots.

- "L'effet Faraday dans le semiconducteur est analysé au moyen de la théorie classique de Drude-Zener".
- "Puis on établit l'interconnexion entre l'effet Faraday d'une part et l'effet Hall et les équations de Maxwell d'autre part".
- "La résistance anormale du sodium à son P F peut être attribué à des lacunes".

T R O I S I E M E P A R T I E

SOUS - DOMAINES

OBJET DE LA 3ème PARTIE

Etude de relations d'ordre entre les adresses de groupes de mots énoncées selon le procédé de la 1ère partie.

Ces relations permettront ensuite un classement des adresses en question.

Le classement en tableau, décrit dans la présente partie, est utilisé par l'école de Z. Harris, à l'université de Pennsylvanie (8) (9).

SOUS-DOMAINES

On désire que le domaine d'un mot contienne les zones d'influence des mots directement subordonnés. Celles-ci seront dites sous-domaines d'ordre n+1 du domaine considéré, lequel se verra lui-même attribuer l'ordre n. - L'ordre se trouvera ainsi fixé à une constante additive n près, ce qui s'explique par le fait que le domaine considéré est en général sous-domaine d'un ensemble plus vaste dont la structure nous importe peu ici. - Les sous-domaines d'ordre n+1 contiendront eux-mêmes des sous-domaines d'ordre n+2 et ainsi de suite, de sorte que si l'on fixait la valeur de n pour un domaine quelconque d'une phrase, elle serait fixée ipso facto pour tous les autres domaines de cette phrase. Mais rien ne nous oblige à expliciter la valeur de la constante n.

exemple : "les quarts de finale des championnats du monde sur courts couverts..."
"quarts" est le mot principal de l'expression, soit n l'ordre de son domaine.
Les domaines de "les" et de "de" sont d'ordre n+1 et ainsi de suite, de sorte que l'on peut construire le tableau suivant :

n	n+1	n+2	n+3	n+4	n+5	n+6	n+7
quarts	les de	finale	des	champion- nats	du sur	monde courts	couverts

[(les) quarts (de[finale (des [championnats (du [monde])(sur [courts (couverts...)])])])]

Ainsi la notion d'ordre d'un domaine conduit à mettre le langage sous forme de diagramme bidimensionnel.

Le domaine d'un mot contient :
au moins :
- le mot, qui est mot fondamental ;
- les sigles associés, qui servent de frontière ;

facultativement :
- divers sous-domaines

Le mot fondamental peut-il se trouver à l'intérieur d'un sous-domaine ? Non car alors ce sous-domaine devrait contenir aussi tout le domaine du mot fondamental, et ce ne serait pas un sous-domaine.

exemples : á ć A C č à n'est pas correct
á A ć C č à est correct

[(la) (sécurité matérielle)] n'est pas correct
[(la) sécurité (matérielle)] est correct

TRAVAIL PRATIQUE

Construire, pour les phrases données en seconde partie, des tableaux analogues à celui dessiné ci-dessus pour l'exemple "Les quarts de etc.".

Q U A T R I E M E P A R T I E

SITUATION RELATIVE DE DEUX DOMAINES

OBJET DE LA 4ème PARTIE

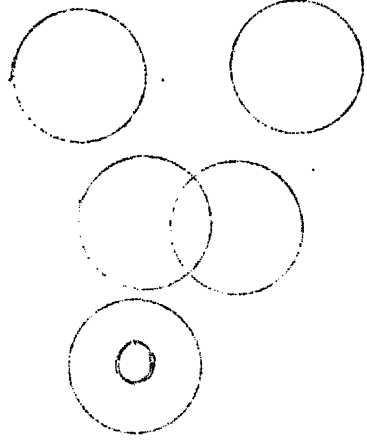
Une convention visant à faciliter les traitements d'adresses de groupes de mots selon le procédé de la 1ère partie.

RAPPEL : SITUATIONS RELATIVES DE DEUX ENSEMBLES QUELCONQUES

DISJONCTION

INTERSECTION

INCLUSION



CONVENTION : DEUX DOMAINES NE SONT JAMAIS SECANTS

Dans un but de simplification, on décide de répartir les zones d'influence des mots sous forme de domaines non sécants. - On admettra, jusqu'à preuve du contraire, que la hiérarchie des mots peut s'exprimer à l'aide de domaines non sécants. - On verra que cette hypothèse apporte de grandes simplifications en échange de peu de difficultés, et que quelques artifices permettent de remédier facilement à ces dernières.

Les seules situations relatives permises pour deux domaines sont donc l'inclusion et la disjonction.
Si deux domaines ont une partie commune, on en concluera que l'un d'eux est entièrement contenu dans l'autre. S'ils n'ont pas de partie commune, ils sont dits extérieurs ou disjoints. Un cas particulier des disjonctions est celui où les domaines, extérieurs l'un à l'autre, ont leurs frontières voisines et simplement séparées par un blanc. Ils sont alors dits mitoyens.

La convention selon laquelle deux domaines ne sont jamais sécants, permet de noter les frontières de domaines à l'aide de parenthèses toutes semblables sans que l'identification d'un couple de parenthèses ne devienne difficile, ni celle de leur mot associé. - Lorsque deux parenthèses de sens opposés se suivent concavité à concavité, elles constituent nécessairement une paire. - Deux parenthèses en regard qui sont séparées uniquement par des paires constituent elles-mêmes une paire.

TRAVAIL PRATIQUE

Vérifier que l'on a respecté la convention de non intersection dans les réponses des parties I - II - et III.
Au cas où il n'en serait pas ainsi, vérifier qu'il est toutefois possible de répartir les zones d'influence d'une autre manière qui soit encore vraisemblable et respecte en même temps la convention de non intersection.

C I N Q U I E M E P A R T I E

CONFIGURATION D'UN DOMAINE QUELCONQUE

OBJET DE LA 5ème PARTIE

On veut réduire le coût du procédé décrit en 2ème partie. Pour celà, on montre qu'il est possible d'utiliser des sigles tous identiques, portant simplement une marque d'orientation.

CONFIGURATION D'UN DOMAINE QUELCONQUE

On sait qu'un domaine quelconque s'étend d'un seul tenant, qu'il est limité par ses parenthèses ; qu'il contient son mot fondamental et aussi éventuellement un ou plusieurs sous-domaines . Peut-il contenir autre chose ? Il ne peut pas contenir de mot isolé supplémentaire, car sinon il contiendrait tout le domaine de ce mot, donc aussi ses parenthèses. - Il ne peut pas contenir de parenthèse isolée, car sinon il appartiendrait à un couple de domaines sécants. - Il ne contient donc rien d'autre.

Un domaine comporte en résumé : obligatoirement le mot fondamental, encadré des parenthèses frontières ; facultativement, des sous-domaines intercalés entre le mot fondamental et les sigles (on a vu que ces sous-domaines ne peuvent contenir le mot fondamental) - ces sous-domaines peuvent être aussi nombreux que l'on veut, à condition d'être introduits entiers entre le mot fondamental et l'une ou l'autre des parenthèses frontières. Ils peuvent contenir eux-mêmes des sous-domaines d'ordres supérieurs.

Un domaine peut ainsi contenir plusieurs mots. Parmi eux, le mot fondamental et lui seul présente cette particularité d'être "nu" à l'intérieur du domaine, c'est-à-dire de n'être pas empaqueté entre des parenthèses de sous-domaines. Il est donc facile de le distinguer des autres mots.

APPLICATION : Un ordinateur électronique a calculé l'étendue des domaines des mots d'une phrase, en disposant des parenthèses entre ces mots. Le listing de sortie signale seulement les nombres de sigles droits et gauches compris dans chaque intervalle entre deux mots, sans préciser s'il s'agit de parenthèses, crochets ou accolades.

0,2 L 1,0 ACADEMIE 0,1 DE 0,1 MEDECINE 3,0

Ce qui signifie :

((L') ACADEMIE (DE(MEDECINE)))

Question A — Est-il possible de déterminer quelles parenthèses sont associées à chaque mot ?

1° Deux parenthèses qui se suivent concavité à concavité sont frontières d'un même domaine. Ceci règle les cas de (L') et de (MEDECINE)

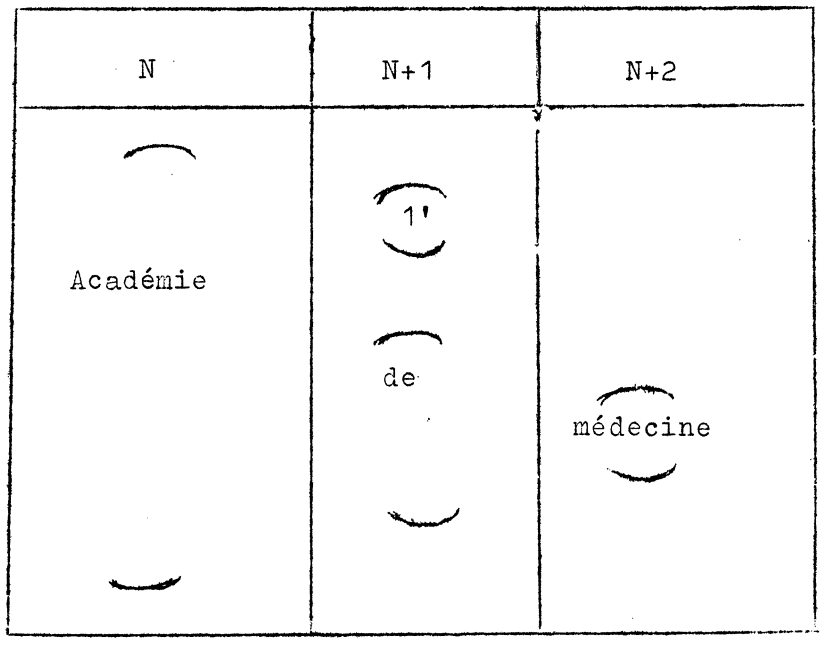
2° Deux parenthèses en regard qui sont séparées uniquement par des paires constituent elles - mêmes une paire.
(DE (MEDECINE)) est donc un domaine, et le seul mot non inclus dans un sous-domaine, c'est-à-dire DE, est son mot fondamental.

(....) ACADEMIE (....) est un domaine et le seul mot non inclus dans un sous-domaine, c'est-à-dire ACADEMIE, est son mot fondamental.

Question B — Est-il possible de faire construire automatiquement le tableau de la deuxième partie ?

Si l'on donne l'ordre de lire tous les signes rencontrés, sigles et mots, avec la consigne :

mots : Ecrire
(avancer d'une colonne puis écrire
) écrire, puis reculer d'une colonne
on obtient le diagramme ci-dessous :



Cette représentation met en évidence le découpage des domaines, la hiérarchie des sous-domaines, et la structure polydimensionelle du langage.

TRAVAIL PRATIQUE :

Un ordinateur électronique a calculé l'étendue des domaines des mots d'une phrase, le listing de sortie signale la répartition de sigles indiquées ci-dessous :

((SI (((l') oxhydrile) est (porté (par ((un) carbone (situé (en (extrémité (de (chaîne)))))))))) ((le) groupement (fonctionnel)) (S') écrit (CH2 OH)).

Question A on demande de retrouver quelles parenthèses sont associées à chaque mot.

Question B on demande de construire un tableau à colonnes analogue à celui donné en exemple plus haut.

S I X I E M E P A R T I E

TRANSFORMATIONS LINGUISTIQUES

OBJET DE LA 6ème PARTIE

Langage ordinaire et langage documentaire ne visent pas à satisfaire les mêmes besoins. Le langage ordinaire doit permettre d'exprimer de façon très souple des nuances parfois très fines. Pour une même idée ou pour des idées voisines il dispose souvent d'un grand nombre de tournures. Un langage documentaire au contraire doit être univoque. Il lui arrivera même d'exprimer sous une forme unique des idées légèrement distinctes mais dont la différence importe peu aux usagers d'un centre de documentation.

Un ordinateur ne pourra traduire le langage courant en langage documentaire que dans la mesure où il saura apprécier l'équivalence sémantique de deux expressions formellement différentes. Nous devons lui fournir des critères à cet effet.

Le procédé qui va être décrit, celui des transformations linguistiques, peut être aussi utilisé à d'autres fins, et notamment pour réduire l'encombrement en mémoire des règles grammaticales et sémantiques: cf. Chomsky (1).

TRANSFORMATIONS LINGUISTIQUES

Pour apprécier l'équivalence sémantique de deux expressions, une première série de critères correspond aux phénomènes de synonymie: équivalence entre deux mots, entre une locution et un mot. Les dictionnaires ordinaires en fournissent des listes concrètes et détaillées.

Une seconde série de critères a été proposée par Harris. Deux morceaux de phrases seront reconnus comme équivalents s'il est possible de passer de l'un à l'autre par une opération abstraite ou une série d'opérations abstraites que Harris a dénommées transformations linguistiques.

Exemple

Les atomes émettent des photons)
Des photons sont émis par les atomes) sont des tournures équivalentes.

N_1 $V_{actif\ transitif}$ N_2 ⇕

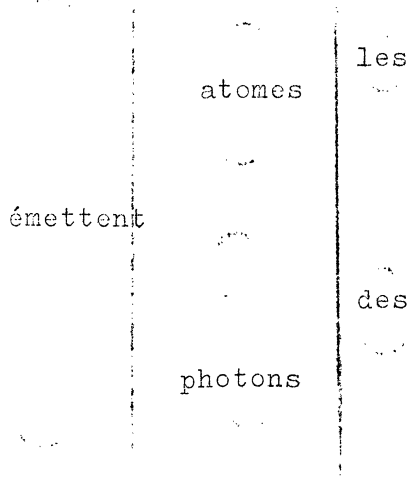
N_2 V_{passif} PAR N_1 ⇕ transformations linguistiques

Exemple

La belle table ;)
La table est belle) équivalence sémantique

Adj N
N EST Adj. Transformations linguistiques

Or, il est possible de commander à une machine de faire ces transformations. Prenons l'exemple de la première phrase



L'ordinateur sait écrire la phrase dans un schéma par colonnes comme ci-dessus. La colonne 1 contient un verbe actif transitif. La colonne 2 contient deux noms qui appartiennent à la zone d'influence du verbe "émettent". La machine peut reconnaitre facilement un cas d'application de transformation linguistique. L'opération n'aura d'influence qu'à l'intérieur du domaine du verbe "émettent". Il faudra successivement :

- intervertir les deux sous-domaines
- remplacer "émettent" par la locution "sont émis par"
- refaire l'analyse de la phrase.

L'avantage de telles consignes réside dans leur caractère abstrait : elles sont générales et pourront servir un grand nombre de fois.

PRODUITS DE TRANSFORMATIONS LINGUISTIQUES

L'équivalence est en principe une relation

- symétrique (A ⌢ B implique B ⌢ A)
- transitive (A ⌢ B et B ⌢ C impliquent A ⌢ C).

Dans la mesure où la relation d'équivalence sémantique mérite de porter ce nom, le produit de deux transformations linguistiques sera aussi une transformation linguistique. Si la transformation inverse peut être définie, elle constituera également une transformation linguistique.

On trouve ainsi que chaque structure est le produit d'une ou plusieurs transformations d'une autre structure, ou la somme de produits de transformations dans le cas où ses parties sont elles-mêmes des transformées.

Exemple

Transformations données :

Les atomes émettant des photons
Les atomes qui émettent des photons — T_1

Les atomes qui émettent des photons
Les atomes par qui sont émis des photons — T_2

Transformation produit =

Les atomes émettant des photons
Les atomes par qui sont émis des photons — $T_1 \times T_2$

TRAVAIL PRATIQUE

- Trouver des exemples de transformations linguistiques
- Donner la liste de consignes permettant de réaliser ces transformations
- Etudier les tranformations inverses.

BIBLIOGRAPHIE

(1) N. CHOMSKY — "Syntactic Structures", Mouton and Co. 's-Gravenhage 1957

(2) N. CHOMSKY — "On certain formal properties of grammars" Information and Control volume 2 n° 2 Juin 1959

(3) O. KULAGUINA — Sur une méthode de définition des notions grammaticales à l'aide de la théorie des ensembles Problemy Kibernetiki 1958,1, page 203 - 214

(4) R. CARNAP — Introduction to semantics and formalization of logic
Cambridge, Massachusetts Harvard University Press 1959.

(5) L. TESNIERE — "Elements de syntaxe structurale". Klincksieck, Paris 1959.

(6) Hjelmslev L. — Prolégomènes pour une théorie du langage. Munskgaard, Copenhague, 1943.

(7) K.E. HARPER and D.G. HAYS — "The use of machines in the construction of a grammar and computer program for structural analysis". Proceedings of ICIP, Paris, June 1959, Unesco, Paris.

(8) Z.S. HARRIS — Higher order substring and well-formedness.
Rapports de l'Université de Pennsylvanie n° 19.

(9) ARAVIND K. JOSHI
Recognition of Local Substrings
Rapports de l'Université de Pennsylvanie n° 18

(10) Z.S. HARRIS — Linguistic transformations for information retrieval
Preprints of papers for the international conference on scientific information.
1958 Area V paper 123-136.

JOURNEE DES SYSTEMES DOCUMENTAIRES

ELABORATION D'UN SYSTEME DOCUMENTAIRE

A. LEROY

Nous nous fixerons ici comme but d'obtenir les documents répondant à n'importe quelle question scientifique le plus exactement possible.

Pour pouvoir effectuer la recherche de ces documents, nous savons que leur contenu doit préalablement être mis en évidence dans une phase nommée analyse.

ANALYSE

Les phrases-clés.- Nous avons donc pour tâche de dégager l'essentiel de ce qui est traité dans chaque document.

Si l'analyse se traduit par l'affichage d'un certain nombre de symboles, c'est que l'on a pu, ou que l'on a cru pouvoir faire correspondre préalablement un sens à ces symboles.

Par exemple : FF14 signifie : "théorie des réseaux electriques linéaires" dans la classification du Commissariat à l'Energie Atomique français. En fait cette symbolisation ne convient pas à nos besoins car elle est beaucoup trop rigide, en ce sens qu'elle ne permet pas d'exprimer des sujets complexes à moins de devenir extrêmement complexe elle-même et le système perd alors tout son intérêt.

On est ainsi conduit à utiliser des mots du texte lui-même

ex : THEORIE RESEAUX ELECTRIQUES RESEAUX LINEAIRES

Et pour sélectionner les articles traitant de la théorie des réseaux électriques linéaires, il suffit de voir si parmi tous les documents il y en a qui possèdent ces trois expressions comme éléments caractéristiques et, à première vue, il ne semble y avoir aucun inconvénient à procéder de cette façon. Si l'on ajoute qu'un système basé sur ce principe est très facilement mécanisable, on s'explique qu'il y ait eu tant d'expériences de ce genre... avec des succès divers, succès dépendant du sujet traité et de la taille de la collection de documents. En vérité, dès que les collections ont atteint une certaine importance,

on s'est aperçu que l'on ne pourrait pas aller loin dans cette direction; la simple juxtaposition des notions-clés entraîne des inconvénients.

En effet, dans ces notions-clés interviennent des actions (actions-clés) et on conçoit que la notion de direction de l'action puisse avoir une grande importance. C'est le cas par exemple lorsque le sujet traite de réactions nucléaires dans lesquelles une particule peut être projectile, cible, ou produit. Plus généralement cela se produit quand les éléments qui "encadrent" l'action sont de même nature par rapport à cette action. On se trouve donc dans l'obligation de faire intervenir des éléments syntaxiques.

D'autre part, le fait d'associer différents mots-clés lors d'une recherche peut conduire à la sélection d'un grand nombre de documents non valables. Par exemple, si un article traite de l'action des rayons gama sur le fer et de l'action de l'acide chlorhydrique sur l'oxyde ferrique, il peut être caractérisé par les mots ACTION RAYONS GAMMA FER ACIDE CHLORHYDRIQUE OXYDE FERRIQUE. Si l'on ne tient pas compte du fait qu'il y a deux sujets différents, on pourra sélectionner - à tort - cet article pour répondre à une question sur l'action des rayons gamma sur l'oxyde ferrique ou de l'acide chlorhydrique sur le fer.

La solution évidente de tous ces problèmes consiste à faire en sorte que les mots-clés concernant un sujet bien déterminé restent groupés et que soient figurés les éléments syntaxiques nécessaires. Cela revient à considérer de véritables "phrases-clés" comme éléments caractéristiques.

Une phrase-clé est donc un ensemble d'éléments "objets" (ex: électron), d'éléments "actions" (ex: bombardement) et d'éléments "relations" (ex: au moyen de), chacun d'eux pouvant par ailleurs recevoir des "qualificatifs". Dans le cas des objets, on dira que ces qualificatifs sont des "propriétés"; par extension, seront également dénommées propriétés les notions qui sont habituellement considérées comme telles dans le langage ordinaire (ex: densité de..., vitesse de...).

On aura ainsi par exemple la suite :

Objet, Action, Objet, Relation, Objet

Dès que la phrase-clé devient complexe, il est nécessaire de placer des repères pour savoir exactement à quel endroit viennent se placer les différentes espèces d'éléments : le moyen le plus simple consiste à tracer un <u>diagramme</u> qui permet ainsi de respecter l'organisation réelle des choses.

Les "objets" sont représentés par des rectangles
" propriétés " " " des cercles
" actions
et relations " " " des flèches.

ex:

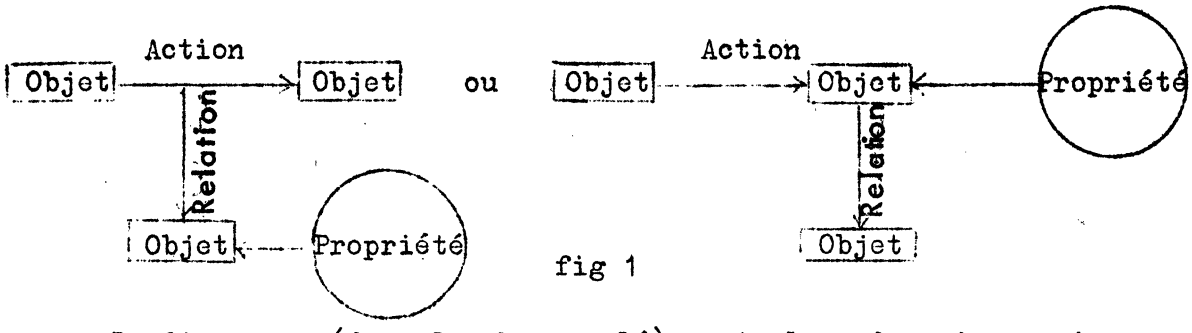


fig 1

Le diagramme (donc la phrase-clé) peut alors devenir aussi complexe qu'on le veut.

Nous rejoignons ici ce que nous avions présenté dans l'introduction à la journée de linguistique. Nous nous sommes seulement placés à un niveau différent d'analyse. Nous avions parlé de langage scientifique, nous nous étions donc situés par exemple au niveau de la phrase ordinaire, alors que nous étudions ici le cas des documents considérés dans leur entier. Les problèmes ne sont en fait pas fondamentalement différents; tout se passe comme dans le premier cas, comme si on voulait mettre en évidence le contenu réel d'un texte réduit à une seule phrase.

On voit que, suivant le désir d'analyse fouillée ou d'analyse plus légère, on se tiendra à un niveau ou à l'autre, ou à un niveau intermédiaire (analyse par groupes de phrases).

Dans tous les cas, la difficulté est évidemment de trouver les concepts adéquats - Ceci ne pourra se faire qu'à la suite d'un certain nombre d'expériences, mais ces expériences nécessitant qu'un choix ait déjà été fait parmi les concepts possibles, il est clair que nous devons procéder par approximations successives, la recherche opérationnelle devant nous guider vers une solution optimale.

ETUDE RELATIVE AUX CONCEPTS

Par suite de ce que nous venons de dire, nous devons prendre en premier lieu les notions telles qu'elles existent dans les textes que nous voulons analyser, et chercher de quelle manière elles doivent être transformées de façon à pouvoir exprimer le plus simplement possible la totalité ou au moins une grande partie de ce qui peut être trouvé de significatif dans le langage scientifique ordinaire. Il sera bien sûr commode au départ de ne considérer qu'une partie du langage scientifique en question, en délimitant de façon conventionnelle le domaine des textes étudiés, mais il est bien entendu que l'ensemble des textes scientifiques doit être considéré comme un tout, c'est-à-dire qu'il n'y a pas lieu de considérer le domaine comme isolé et qu'il faut notamment prendre garde aux problèmes de "frontières" dus par exemple à la polysémie.

Un mot peut en effet avoir plusieurs sens; mais ce n'est jamais qu'un seul d'entre eux qui doit être considéré dans un contexte donné : le sens "contextuel". D'où la nécessité de considérer chaque mot dans son contexte; cela revient à créer autant de mots qu'il y a de sens différents et au contraire à "rapprocher" les mots qui possèdent un même sens.

Il y a donc lieu d'établir un dictionnaire qui associe à chacun des mots qui peuvent être trouvés dans les textes à analyser, les définitions qui correspondent à chacun des sens possibles de chaque mot et éventuellement les nouveaux termes choisis pour exprimer chacun de ces sens.

De cette façon, le contexte d'un mot permettant de lui associer un domaine caractéristique, il suffira de se reporter au dictionnaire pour trouver quelle est la définition correspondant à ce domaine et, finalement, le nouveau terme devant être retenu.

Certains documentalistes pensent que ce nouveau terme doit présenter des qualités spéciales : il doit notamment évoquer quelques uns de ses voisins hiérarchiquement supérieurs et contenir des éléments de définition.

Exemple: le mot associé à rat doit tenir compte de ce qu'un rat est un mammifère, un vertébré etc...

le mot associé à téléphone doit indiquer qu'il s'agit de quelque chose qui agit sur de l'information, que ce quelque chose utilise l'électricité, que c'est un appareil, qu'il est utilisé pour transmettre etc... (1)

Le but de cette complication au niveau de l'analyse est de faciliter la sélection des documents. Ainsi, pour le premier exemple, un article traitant d'une espèce de rats sera obligatoirement sélectionné lors d'une demande concernant les vertébrés. Or ceci ne nous semble pas justifié. Pour répondre à une telle demande, un système documentaire doit, à notre avis, fournir d'abord les documents qui traitent d'une manière générale des vertébrés (et uniquement ceux-là) et des indications montrant que si ces premiers documents ne suffisent pas - ce qui est le cas notamment si la question est mal posée - il est possible d'obtenir des documents traitant des différentes classes de vertébrés etc... Ces indications ne doivent d'ailleurs pas seulement porter sur les liaisons hiérarchiques, mais aussi sur les liaisons "latérales". Nous verrons comment cela peut être réalisé à partir de ce que l'on peut appeler le "diagramme général".

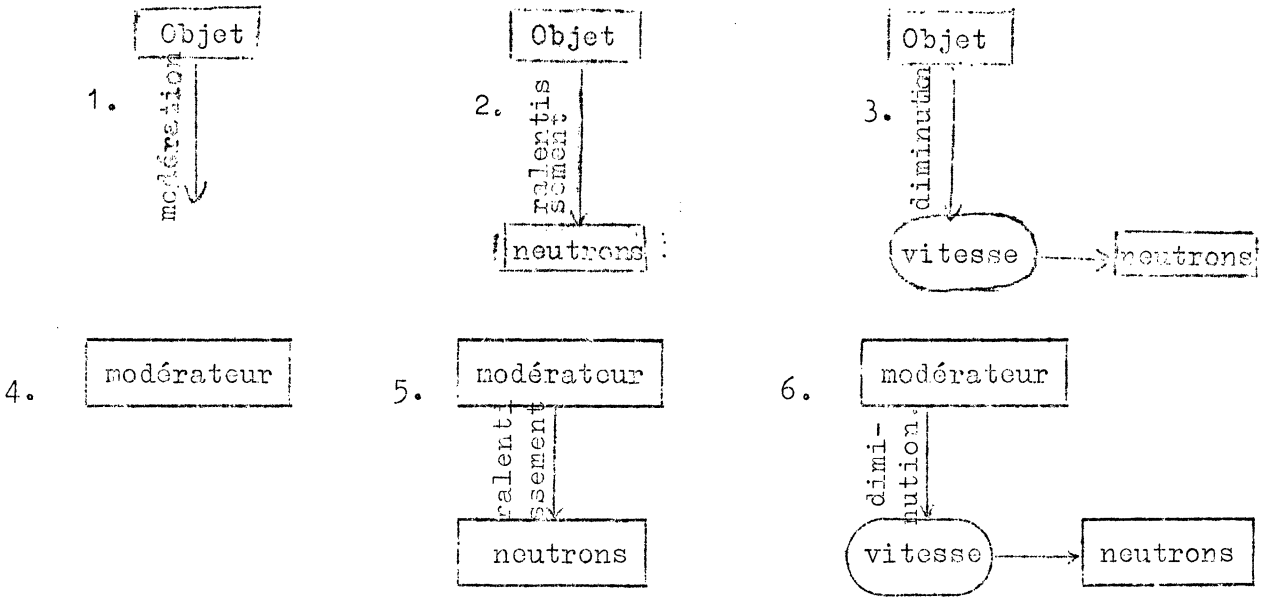
Ceci nous oblige à respecter dans un texte scientifique le "niveau" auquel l'auteur s'est situé, c'est-à-dire à ne transformer les mots et expressions utilisés qu'en restant à un même niveau de description, afin de ne pas modifier le contenu même du document. C'est ainsi que les mots décrivant une structure atomique ne sont pas les mêmes que ceux relatifs à un point de vue macroscopique et une question rédigée en terme d'une structure ne devra se voir correspondre que des documents rédigés "au même niveau", du moins dans une première réponse.

De la même façon, on pourrait être tenté d'attribuer à chaque objet un schéma plus ou moins complexe.

---

(1) voir Perry et Kent. Tools for Machine literature searching. Interscience 1958.

Exemple : Un modérateur peut être nommé de différentes façons; on a donc affaire à des expressions synonymes mais qui mettent en jeu des éléments de plusieurs niveaux :



etc...

fig. 2

D'après ce qui a été dit jusqu'ici, c'est le diagramme 4 qui est le bon si le seul mot modérateur est cité dans le texte analysé. Par contre, le diagramme général permettra de savoir que les autres schémas peuvent réellement exister.

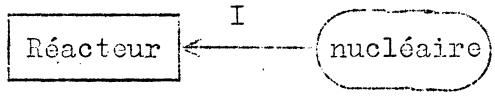
ACTIONS ET RELATIONS

Une action est considérée comme tout ce qui implique un changement d'état.

ex. "fabrication" implique le passage d'un état non fini à un état fini.

Ces actions sont bâties directement à partir de la plupart des verbes du langage ordinaire et elles jouent le rôle des verbes du langage réduit. Toutefois, certains verbes ordinaires, les verbes être, avoir, posséder etc... n'ont pas d'équivalents en tant que verbes du langage réduit, puisqu'ils n'impliquent pas de changement d'état, ils caractérisent en fait chacun un état. Ils trouveront leur équivalent dans les relations d'identité et d'appartenance.(1)

Relation d'identité (I)

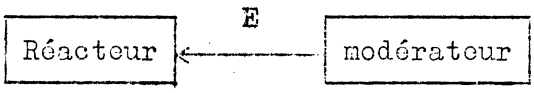


Le réacteur est nucléaire

fig. 3

(1) Le rapport GRISA n°5, publié en août 1960, fait le point des conventions qui ont été adoptées en ce qui concerne les relations.

Relation d'appartenance (E)



Le modérateur fait partie du réacteur.(ou:le réacteur possède un modérateur)

fig. 4

Les relations concernent également les rapports de temps, de lieu, de circonstance, de but, de cause etc...

Nous avons dit qu'une action implique un changement d'état. Or cet état est déterminé par un certain nombre de conditions que l'on peut appeler ses coordonnées : (position, température etc...)

Une ou plusieurs de ces coordonnées sont modifiées au cours du changement. Nous avons donc des schémas du type suivant :

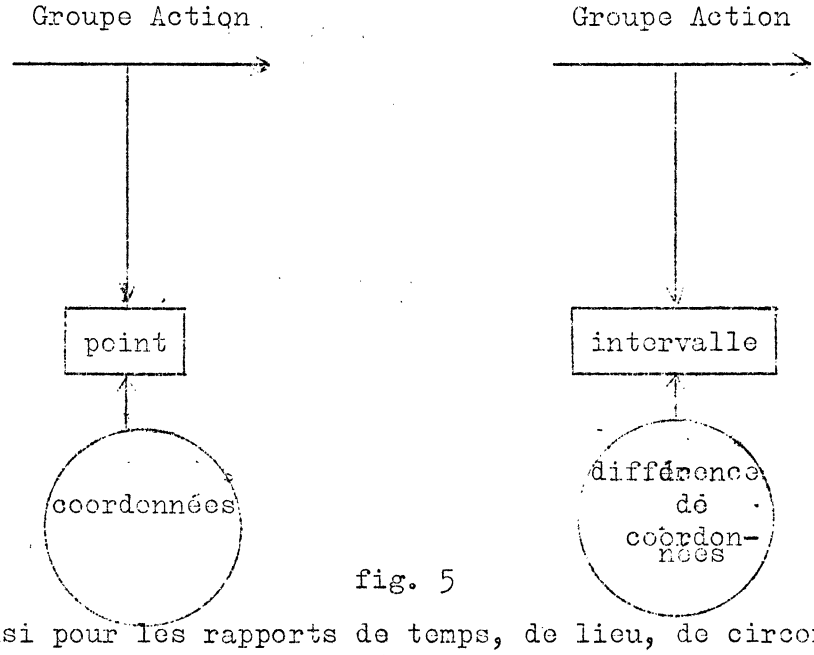


fig. 5

Ainsi pour les rapports de temps, de lieu, de circonstance, les relations correspondront aux prépositions :

à	relation	A
dans		D
vers		V
sur		SU
sous		SO
après		AP
avant		AV
depuis		DE
jusqu'à		JU

ex.

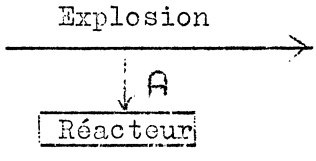


fig. 6

Il n'y a bien sûr pas lieu de prévoir l'utilisation de toutes les prépositions existant dans un langage donné; il faut au contraire rechercher les prépositions de base qui ont seule une raison d'exister du point de vue scientifique. Ceci tient à ce qu'il se pose également ici des problèmes de polysémie et de synonymie dont il y a lieu de tenir compte. Une même préposition peut vouloir signifier plusieurs choses suivant les mots auxquels elle est attachée; au contraire, deux prépositions peuvent avoir le même sens - Pour trouver les relations nécessaires, il sera commode de rechercher quelles sont toutes les possibilités d'utilisation d'un intervalle de coordonnées quelconque.

Considérons par exemple, pour débuter, un intervalle entre deux coordonnées.

On peut s'intéresser : à un point de cet intervalle (A)
à un intervalle compris dans le premier (D)
aux points extérieurs à l'intervalle (AP AV)

et, d'un point de vue dynamique :
à un déplacement vers l'intervalle (V)
à un déplacement depuis un point (DE)
jusqu'à un autre (JU)

Ainsi pour effectuer une analyse, il est nécessaire de bien voir la nature logique des relations pouvant se présenter.

Les relations de moyen, de but, de cause etc... ne présentent pas de difficultés spéciales.

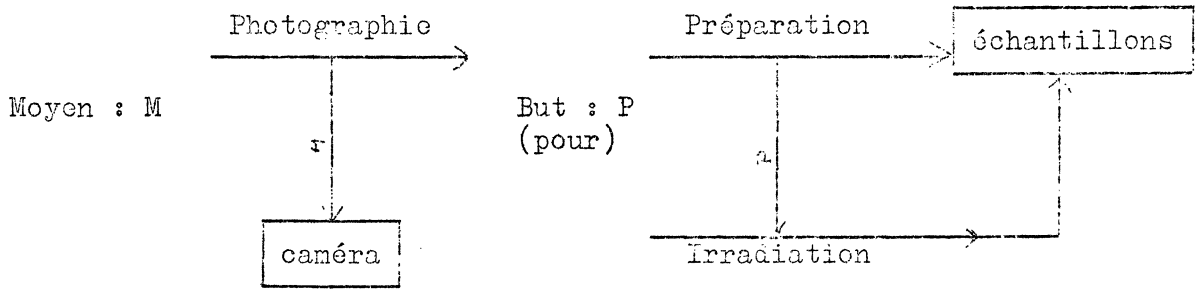


fig. 7

SELECTION

Diagramme général

A la limite, il est tel que tous les contextes possibles d'un objet quelconque y sont indiqués. On y introduit donc tout le contenu des publications; on opère comme si l'on considérait l'ensemble des publications comme un seul document. Ceci est justifié par le fait qu'un document donné est toujours lié à d'autres, la plupart du temps parce qu'il se sert de mots qui ont donné lieu à des explications dans ces autres documents.

Finalement, c'est l'ensemble des diagrammes correspondant à chacun des documents qui constitue le diagramme général. Pour que les diagrammes particuliers puissent se superposer, il faut que les objets dont ils traitent soient les mêmes, c'est-à-dire se situent dans le même contexte.

Exemple de partie de diagramme général d'après une étude de M. Detant.

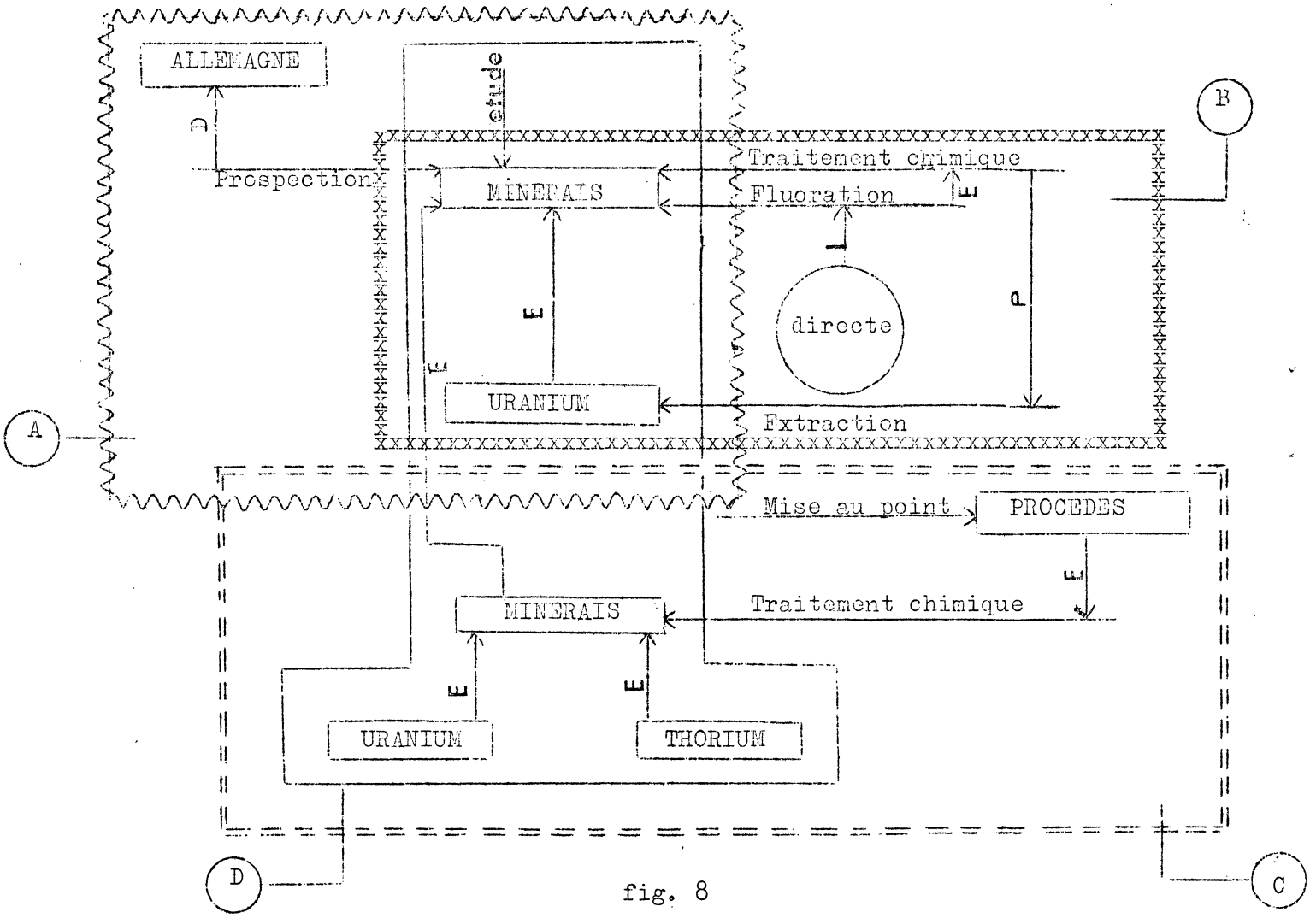


fig. 8

Ce diagramme général nous permet de répondre de façon complète à toute question. En effet, nous y trouvons déjà des indications hiérarchiques et latérales .

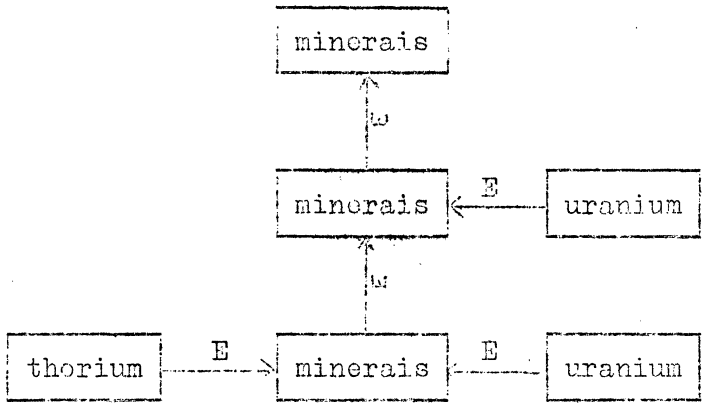


fig. 9 exemple d'indications hiérarchiques.(1)

D'autre part, si nous prenons soin d'y distinguer ce qui a été apporté par chaque document, nous avons la réponse complète, puisque nous procédons par comparaison.

(voir tableau page suivante)

(1) Par la suite il a été jugé nécessaire de distinguer deux sortes de relations d'appartenance = l'appartenance à une "classe" et l'appartenance "physique".

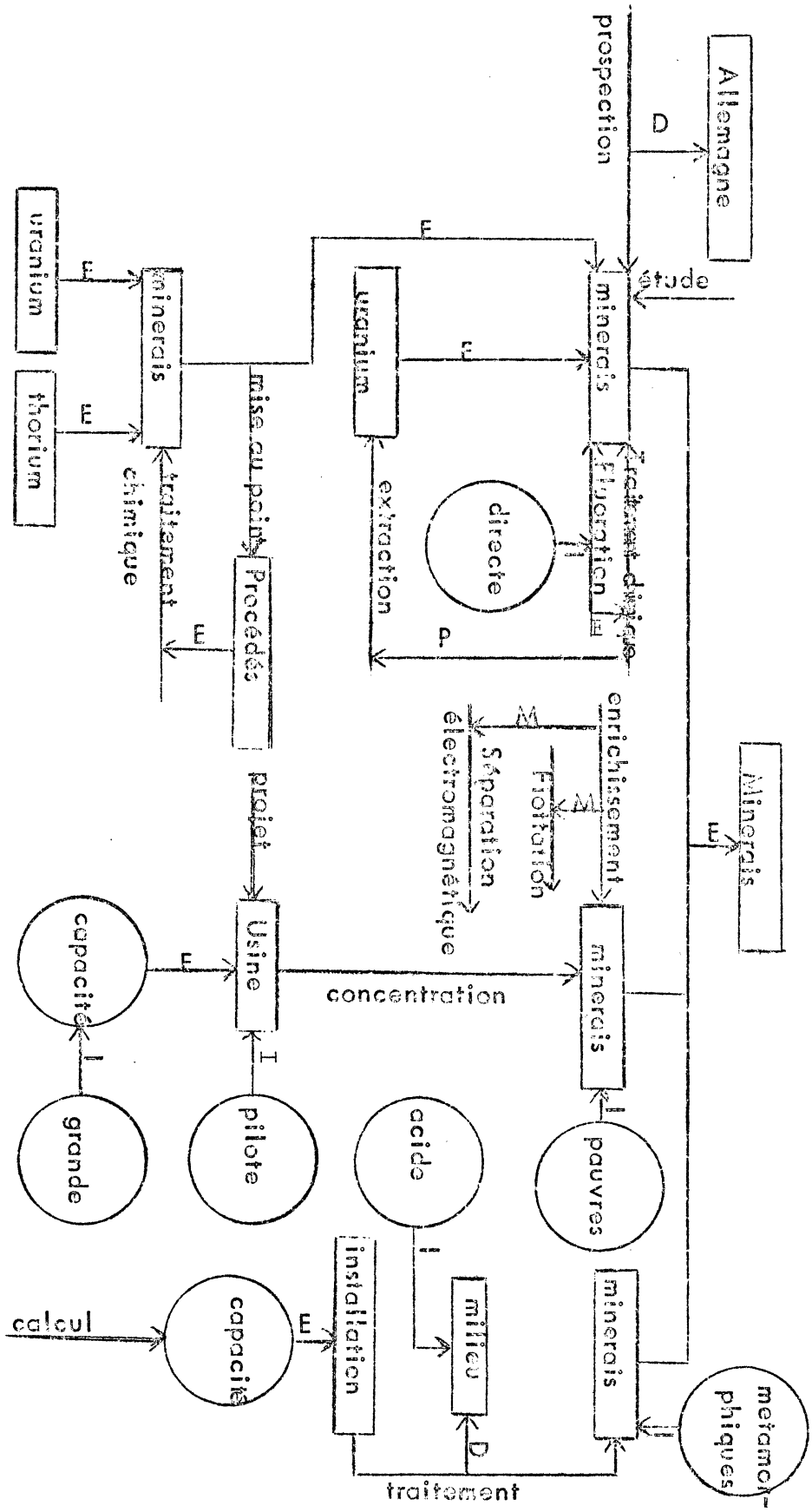

Allemagne
prospection
D
étude
minerais
uranium
thorium
E
extraction
directe
P
Flottation
traitement direct
chimique
traitement
mise au point
Procédés
projet
Usine
concentration
capacité
grande
pilote
I
enrichissement
Séparation électromagnétique
Flotation
M
minerais
pauvres
acide
milieu
D
installation
calcul
capacité
traitement
Minerais
métamorphiques

Figure 10

Ainsi la partie gauche du diagramme a été construite à partir de quatre documents A, B, C et D. (fig.8)

Si une question est rédigée dans les termes suivants : "trouver les documents qui se rapportent aux procédés de traitement des minerais d'uranium et de thorium", le document C devra bien entendu être sélectionné et des indications supplémentaires pourront être données d'après les liaisons hiérarchiques et latérales.

ex. Si les documents sortis ne suffisent pas, il est possible de rechercher également des documents qui se rapportent au traitement des minerais d'uranium en général.

Remarquons que si aucun document ne répond directement à la question posée, il est possible d'obtenir tout de même des documents qui, pris ensemble, y répondent. Il suffit pour cela que la question entière figure dans le diagramme.

ex. Trouver les documents se rapportant à la prospection de minerais qui peuvent donner lieu à un traitement par fluoration.

question :

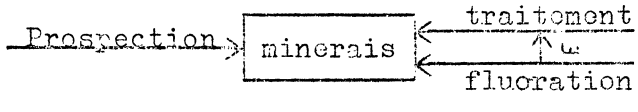


fig. 11

Aucun des documents A, B, C ou D ne répond entièrement à la question, mais la combinaison A + B est satisfaisante.

Réalisation pratique

Il nous faut bien entendu repasser en forme linéaire.

L'intervention des relations apporte évidemment des difficultés d'exploitation par rapport aux simples mots-clés avec des systèmes simples.

Les systèmes du genre peek-a-boo, c'est-à-dire ceux que l'on utilise par transparence, donnent lieu à l'utilisation suivante :

Supposons que l'on veuille noter ARB où A et B représentent des notions-clés et R une relation. Nous pouvons attribuer une carte à chacune des notions et à la relation et obtenir les documents comportant ARB par transparence, à condition de prendre soin de prévoir les flexions nécessaires pour les raisons exposées au début (notamment quant à la direction de la relation).

M. Gardin aura l'occasion de montrer dans l'exposé suivant comment on peut effectivement utiliser un dispositif basé sur ce principe.

Toutefois, la conception et l'utilisation d'une partie suffisamment importante du diagramme général ne peuvent se concevoir qu'à l'aide de calculateurs spéciaux et nous verrons, lors de la dernière journée, comment il est effectivement possible de les envisager.

ANALYSE ET SELECTION DOCUMENTAIRES DANS LES SCIENCES HUMAINES

J.C. GARDIN (*)

Quand j'entends Braffort dire que nous allons de plus en plus vers le concret, et que ce concret est représenté par les sciences humaines, cela me met à l'aise, car le concret, comme vous allez le voir, c'est un peu l'anarchie, de sorte que nous allons lentement de l'ordre vers l'anarchie. Mais c'est peut-être une bonne méthode que de montrer ce que l'on peut faire dans des domaines anarchiques, pour introduire un semblant d'ordre par des procédés très proches de ceux qui viennent d'être exposés par Leroy.

Une première question qui vous est peut-être venue à l'esprit, à la suite des exposés des deux derniers jours, est de savoir où conduisent ces recherches d'analyses formelles - car en fait il s'agissait bien, sauf ce matin, d'analyses formelles, et non pas d'analyses sémantiques - où conduisent, du point de vue de la recherche documentaire, ces procédés de formalisation linguistique. La réponse est déjà contenue dans les commentaires qu'ont donnés les exposants, ici même. En effet, l'analyse purement formelle ne suffit pas à rendre compte de ce qui est spécifique dans un texte; en d'autres termes, elle n'introduit pas une discrimination suffisante, et l'on est appelé, même dans des entreprises qui se veulent purement formelles comme celles de Chomsky, à recourir à des catégories qui relèvent au moins en partie d'une analyse sémantique. Puis, Braffort a montré que, même en introduisant ces notions sémantiques, les formalisations auxquelles on aboutit ne sont pas encore assez fines pour rendre compte de toute la diversité des énoncés linguistiques. Le cas de la littérature propre aux sciences humaines est à cet égard fort bon, parce que, pour la raison que j'ai indiquée au commencement, à savoir l'état anarchique de ces sciences, il est impossible au départ de se fonder sur un ordre quelconque, du point de vue de la formalisation. On se trouve d'emblée en présence de langages extrêmement riches, très peu systématisés, où chaque notion est pratiquement justifiable de plusieurs définitions, selon les auteurs, selon les pays, selon les écoles, de sorte que le problème de la formalisation est immédiatement et essentiellement un problème sémantique. Un premier aspect de cette situation particulière est la fragilité de la dichotomie très fréquemment citée, et encore ce matin par Leroy, entre les éléments d'une part et les relations de l'autre - autrement dit toutes les entités nominales

* Centre d'Analyse Documentaire pour l'Archéologie.

d'une part, et de l'autre, tous les moyens linguistiques qui servent à exprimer les relations entre ces entités. Cette distinction est pratiquement impossible dans les sciences humaines. A la limite, le seul être naturel, la seule entité nominale, c'est l'objet même de ces études, à savoir l'homme, l'individu. Je dirais que toutes les autres entités sont des abstractions, et qui dit abstraction dit au fond, à un certain égard, relation d'une entité abstraite par rapport au système qui la définit. Il n'y a pas d'être naturel, il n'y a pas d'être réel dans les sciences humaines, en dehors de l'homme - et évidemment cet élément-là n'est pas très utile pour la construction d'un système d'analyse formelle. Dans le meilleur des cas, les abstractions dont je parle sont des abstractions à peu près universelles, c'est-à-dire qu'elles sont définies avec relativement peu d'ambiguïté, et qu'elles sont fondées, au moins en apparence, sur une espèce de substrat naturel. C'est l'image, en tout cas, que l'on se fait de certains termes abstraits comme par exemple ceux qui désignent les relations de parenté. On peut imaginer en effet qu'il n'y a aucune ambiguïté sur ce que c'est, par exemple, qu'un "père", un "oncle" ou une "grand-mère". Mais dès que l'on cherche à raffiner la définition de ces termes, on s'aperçoit que ce sont des définitions non pas naturelles mais culturelles et que, même dans un domaine apparemment aussi universel, aussi concret, que celui des systèmes de parenté, les définitions sont des définitions relatives à une culture donnée. Il existe une quantité de sociétés dites primitives par exemple où les oncles ne sont pas du tout des oncles "biologiques", mais plutôt des oncles "sociologiques", c'est-à-dire qu'ils occupent des positions différentes, dans des systèmes de parenté qui sont eux-mêmes différents. De sorte que si l'on emploie le terme dans l'analyse - à savoir oncle, grand-père, grand-mère ou autres - à l'intérieur de cadres sociologiques différents, on introduit une ambiguïté sémantique qui peut être gênante au moment des recherches documentaires. Je prends cet exemple des termes de parenté parce que c'est le plus concret qui soit, et pour vous montrer que dans le meilleur des cas, les notions dont on se sert dans les sciences humaines sont des notions foncièrement abstraites et foncièrement instables de par leur contenu sémantique.

A la limite, un grand nombre de notions propres aux sciences sociales n'ont de sens que dans le cadre d'une certaine théorie particulière dont elles font partie; par exemple, envisagez des notions très fréquentes telles que l'"acculturation", la "volonté de puissance", l'"harmonie sociale", l'"intégration de l'individu au groupe" - autant de concepts dont on se sert constamment dans la littérature, mais dont personne au fond ne sait très clairement ce qu'ils recouvrent.

Entre ces deux extrêmes, se trouvent toute une série de termes qui sont des éléments nominaux et que l'on pourrait être tenté de traiter comme ces entités fondamentales dont parlait Leroy ce matin : par exemple, les "rois", les "prêtres", les "marchands", les "esclaves" etc. Mais pour les mêmes raisons que dans le cas des termes d'origine apparemment biologique comme les termes de parenté, il est très difficile de donner une définition universelle et stable surtout de chacune de ces entités.

On ne peut pas être "roi" dans l'absolu; être roi suppose qu'il y ait quelque part des sujets dont on est le roi. Or, il existe une très grande variété de relations possibles entre un individu privilégié et un groupe, qui font que l'on peut dire ou ne pas dire que l'individu en question est "roi" ou non. Le terme utilisé n'a pas tellement d'importance; on peut imaginer une quantité de périphrases qui, rapportées à un être donné, indiquent d'une façon parfaitement certaine qu'il est "roi", plus, peut-être, que celui auquel on applique le terme de "roi" dans un texte. Il est évident qu'une analyse qui s'effectuerait strictement au niveau de la forme littérale de l'énoncé perdrait par conséquent une large part du contenu sémantique des textes. Du même coup, on voit que cette distinction fondamentale entre les éléments et les relations, dont nous étions tenté de partir, ne peut pas être retenue. Nous avons bien plutôt affaire à des "fonctions" définies par certaines associations récurrentes entre des individus et divers prédicats dont ils sont l'objet. Mais l'on est alors conduit à considérer une quantité de doublets selon que les prédicats désignent des actions particulières, non-récurrentes exceptionnelles - par exemple des "appropriations de biens", des "déplacements de personnes" de "grands travaux", en général - ou au contraire des fonctions générales, récurrentes ou institutionnalisées- par exemple, en correspondance avec les "appropriations de biens", les vols, les saisies, les cautions, actions qui toutes impliquent des transferts de biens, mais qui sont définies dans des univers tout à fait particuliers, juridiques ou politiques; de même, parallèlement aux "déplacements de personnes en général, les exils, les ambassades et plusieurs autres termes qui tous impliquent un déplacement de personnes, mais toujours dans un cadre particulier, également juridique ou politique dans les deux exemples que je viens de donner; enfin, parallèlement aux "grands travaux", certaines notions telles que la corvée, l'artisanat même, qui représentent chacune un cas particulier des travaux en général. L'analyse ne peut jamais s'effectuer sans l'existence dans le code de ces doublets; c'est-à-dire que la notion générale ne dispense pas des variantes spécifiques et que, réciproquement, un catalogue aussi complet soit-il de toutes les manifestations spécifiques d'une notion générale ne permet pas non plus de se passer de la notion générale.

La conséquence d'une situation de ce genre est que le langage même de l'analyse présente encore un grand degré de liberté, et qu'en fait on ne peut pas dire que nous soyons parvenus à aucune formalisation entièrement satisfaisante dans le domaine des phénomènes humains, qu'il s'agisse de textes, d'objets, d'ornements abstraits, qu'il s'agisse enfin d'images, c'est-à-dire de documents iconographiques. Dans aucun de ces cas, nous n'avons pu définir une formalisation qui soit suffisamment rigoureuse pour offrir une certaine garantie d'universalité. En outre, chaque système formel demeure largement subjectif dans la mesure où il nous est difficile de faire abstraction de nos propres catégories linguistiques, c'est-à-dire de nos catégories conceptuelles, au cours de son élaboration. Ces catégories conceptuelles nous sont données par le langage que nous avons appris, et il n'est pas du tout sûr que dans un

même domaine d'autres individus appartenant à des cultures différentes, disons un Malgache et un Indien de plaines d'Amérique du Nord, auraient obtenu le même système de formalisation. Vous voyez à quel point la situation est différente de celle où nous sommes dans les sciences naturelles, puisque, même lorsque nous découvrons certains phénomènes en apparence bien délimités, par exemple les formes de vases, nous ne pouvons jamais avoir l'assurance que les catégories auxquelles nous aboutissons sont des catégories universelles.

Chemin faisant, j'ai déjà fait allusion à une difficulté que l'on rencontre dans l'analyse des textes, ou plus généralement dans l'analyse des documents intéressant les sciences humaines : c'est l'équivalence entre certaines notions définies par des mots simples, et les mêmes notions définies par des périphrases ou par des propositions complexes. Un premier but de l'analyse sémantique est de mettre en évidence ces équivalences. C'est un problème que l'on rencontre dans l'étude des textes, mais aussi dans l'analyse des images, puisque le même thème, autrement dit le même évènement, peut être représenté graphiquement par une série de formes différentes, non seulement en ce qui concerne les acteurs de la scène, mais aussi dans l'expression des relations ou des actions entre ces acteurs. Il y a là une quantité d'équivalences que nous n'avons jamais réussi à épuiser entièrement, même dans un domaine limité; là encore, c'est une difficulté qu'il faut garder présente à l'esprit lorsque l'on veut comparer les problèmes d'analyse dans les sciences naturelles et dans les sciences sociales. Nous ne sommes certainement pas dans celles-ci en présence de situations aussi facilement mises en équation que dans celles-là. Cependant, ce genre de problème n'est peut-être pas particulier aux sciences sociales; on le retrouve dans certaines disciplines des sciences naturelles, et je pense, en l'occurrence, à la sémeiologie médicale. Si l'on voulait aujourd'hui constituer un langage de la pathologie, je pense qu'on serait bien en peine pour poser toutes les équivalences entre les différentes constellations de symptômes pris "n" à "n" et les phénomènes globaux baptisés du nom de maladies. Il existe en effet, dans un domaine de ce genre, une quantité de regroupements qui ont reçu un nom de baptême - souvent éphémère - mais aussi une quantité d'autres qui n'ont encore jamais été explicités par le langage. C'est, en fait, cette même situation que nous rencontrons le plus souvent dans les sciences sociales.

Supposons cependant ce premier type de problème résolu; lorsque l'on dispose d'un vocabulaire descriptif bien défini, à tous les niveaux de spécificité, il reste à exprimer les rapports variables que tous les éléments de ce vocabulaire peuvent entretenir les uns avec les autres, dans les divers types de phénomènes ou de "propositions" envisagés. Ici, je m'écarte un peu du langage de Leroy: j'entends par "relation", je crois, ce qu'il appelle lui-même "action", c'est-à-dire un changement d'état d'un certain "objet" provoqué par l'intervention d'un certain "sujet", l'objet pouvant être une abstraction aussi bien qu'un être réel.

Leroy a montré ce matin, par ses diagrammes, la façon dont on pouvait séparer ces deux pôles d'une relation. Mais une question vient à l'esprit devant de tels diagrammes : c'est le sort qui leur est réservé lorsqu'on les atomise eux-mêmes pour les introduire sous une forme linéaire dans une mémoire quelconque, autrement dit lorsqu'on sépare ce que le diagramme a réuni par certaines flèches judicieusement placées. A ce moment, toutes les relations, qu'elles désignent l'appartenance, l'existence, l'identité d'une part, ou une action proprement dite de l'autre, toutes ces relations ou les flèches qui leur correspondent, deviennent en quelque sorte indépendantes. On peut alors, au moment de la recherche, les rattacher à n'importe quel support et obtenir de la sorte toute une série d'interférences ou d'ambiguïtés, puisqu'on ne sait plus à quels éléments de l'analyse se rapporte chaque relation. En d'autres termes, les combinaisons réelles sont égarées parmi une quantité d'autres possibles, et l'on a perdu la valeur sémantique de l'ordre des éléments dans le diagramme. Je pense que Leroy nous expliquera cet après-midi comment il envisage de résoudre cette difficulté, et je la laisse momentanément de côté pour exposer tout de suite une deuxième catégorie de problèmes moins souvent débattus : c'est ce que j'appellerais l'existence des "sauts sémantiques" dans l'analyse d'un ensemble de notions contigües.

Dans le cas le plus simple, pour chaque terme pris isolément, le saut sémantique n'est pas autre chose que la synonymie. On peut construire soit à priori, soit à posteriori, le champ sémantique de chaque terme, et retrouver ainsi toutes les variations littérales sur un thème conceptuel donné, selon le contexte. Telle est au fond la fonction des "Thesaurus"; je pense que certains d'entre vous ont déjà rencontré ce genre d'ouvrage, ou leur description dans la littérature. Les Thesaurus sont des dictionnaires de synonymes, mais au sens élargi; ce sont plutôt des tables d'associations d'idées groupées par mots dont Roget, au 19ème siècle, a donné le spécimen le plus célèbre, en anglais, et dont Wartburg a fourni plus récemment à Vienne une version systématique pour la langue française. La méthode du Thésaurus est utilisée aujourd'hui par l'école anglaise de Cambridge dans la résolution des problèmes de traduction automatique. Mais elle n'est pas suffisante pour l'analyse documentaire parce que le passage d'une signification explicite à tout un ensemble de significations implicites ne s'effectue pas seulement au niveau de chaque mot pris un à un. Le problème le plus grave et le plus difficile, c'est celui des "sauts" sémantiques provoqués par la conjonction de deux, trois, ou plusieurs termes voisins. J'en donnerai un exemple concret par un texte très simple emprunté à un code juridique de l'ancien Orient. "Si le palais s'écroule provoquant la mort du roi, on emmurera l'architecte qui a construit le palais".

C'est donc une relation entre un certain accident considéré comme un délit, et une sentence juridique. L'analyse de cette phrase consiste naturellement à retenir tout d'abord les termes explicites, à savoir le palais et son écroulement, le roi et sa mort, l'architecte et sa condamnation. Mais une analyse qui ne citerait que ces seuls termes serait extrêmement pauvre. En particulier, elle ne permettrait vraisemblablement pas de retrouver ce texte chaque fois que l'on envisage par

exemple "Les conséquences juridiques des accidents dus à l'incompétence professionnelle". Dans cette proposition figurent en effet des notions qui font défaut dans le texte et qui n'ont aucune raison d'apparaître dans le champ sémantique d'aucun des termes de l'énoncé. Celle d'"accident",à la rigueur, pourrait se trouver dans le champ du verbe "s'écrouler". Mais la notion d'"accident mortel", qui doit être considérée dans les problèmes juridiques comme une notion globale insécable, est ici donnée par une conjonction de termes: "(le palais) s'écroule (provoquant la) mort (du roi)".

De la même manière le lien entre les deux propositions consécutives - la première qui constitue l'accident, la cause, et la seconde qui en est la conséquence juridique - montre que cet emmurage de l'architecte est une condamnation à mort, une condamnation juridique. Cela pourrait très bien ne pas être le cas : on peut fort bien imaginer que l'on emmure des gens sinon pour le plaisir, du moins selon son bon plaisir, c'est-à-dire dans un cadre tout autre que juridique. Cette notion "condamnation à mort" ne se trouve dans le champ sémantique d'aucun des termes de l'énoncé pris un par un, mais uniquement au niveau de la conjonction entre la première proposition et la seconde. Le fait de dire de la condamnation que c'est une condamnation à mort est une induction amplifiante, puisque rien ne nous dit que l'architecte soit mort; mais comme nous nous trouvons dans un contexte juridique, que l'on sait que la sentence est infligée à la suite d'un accident qui a été lui-même mortel, il est fort probable que le but de l'emmurage est de provoquer à son tour la mort du responsable de l'accident. Là encore, une méthode purement mécanique d'analyse sémantique, par des procédés tels que le Thesaurus, ne donnerait jamais ce genre d'extrapolation pourtant indispensable.

Voilà donc un premier "saut" de la phrase "nucléaire" à un premier "anneau" sémantique. Mais il y a d'autres notions impliquées dans l'énoncé, en particulier cette de "talion". Toute personne un peu familière avec la procédure orientale, la procédure juridique bien entendu, reconnaîtra dans cette proposition un cas d'application du talion, c'est-à-dire en somme le principe de l'équivalence du délit et de la peine. A l'accident mortel provoqué vraisemblablement par une faute professionnelle succède la mort du responsable professionnel en question. Mais l'inclusion du terme "talion" dans l'analyse marque un deuxième saut sémantique, non plus depuis la phrase originelle, mais déjà depuis une première interprétation de la phrase originelle. De même, une notion importante manque encore dans l'analyse : c'est celle de "responsabilité professionnelle". La raison de la condamnation de l'architecte n'est pas du tout celle qui conduit à condamner les meurtriers en général pour un crime prémédité. L'architecte est coupable en tant que responsable de la construction du palais : si ce palais s'écroule, c'est l'architecte que l'on condamne. Autrement dit, de la conjonction de plusieurs termes, dont certains sont des termes dérivés, et non pas donnés explicitement, on peut inférer qu'on se trouve devant un cas de "responsabilité" juridique, et même, plus spécifiquement, de "responsabilité professionnelle" à cause de la présence de l'architecte

Ces "sauts sémantiques", à partir d'une proposition nucléaire quelconque, ne peuvent pas être effectués, du moins actuellement, par des procédés purement déductifs, entièrement mécanisables. Cette impossibilité n'est nullement d'ordre philosophique; elle tient seulement au fait que nous connaissons encore fort mal les principes qui régissent ces générations successives de concepts, au fur et à mesure que l'on étend le champ des associations verbales. C'est la raison pour laquelle les méthodes fondées sur les dictionnaires de synonymes ou les dictionnaires d'associations suggérées par les mots, ne sauraient encore résoudre tous les problèmes de l'analyse documentaire.

Vous voyez donc que le problème des relations présente en fait deux aspects : un premier aspect, en quelque sorte mécanique, qui consiste à trouver des "trucs" pour conserver la valeur sémantique de l'ordre des mots, lorsque l'on passe d'un langage naturel à un système linguistique dont les termes sont en principe commutatifs; et un deuxième aspect beaucoup plus important à mes yeux, à savoir la détection des principes qui régissent les "sauts sémantiques" lorsque l'on passe d'une proposition explicite à l'ensemble des composantes implicites.

Je ne dirai qu'un mot assez court sur ce dernier aspect du problème, pour passer ensuite à cette étude des "trucs" dont j'ai parlé. Vous voyez tout d'abord que la situation dans laquelle nous nous trouvons rend fragile, en tout cas pour le moment, les espoirs que l'on pourrait avoir d'automatiser complètement l'analyse documentaire. Quand je dis "automatiser l'analyse" je pense naturellement non pas à l'automatisation des sélections documentaires, mais bien au passage automatique d'un texte formulé dans le langage naturel à sa version codifiée dans les termes d'un langage artificiel. La raison pour laquelle cette entreprise me paraîtrait pour le moment suspecte, au moins dans les sciences humaines, c'est que nous ne connaissons absolument pas les règles qui servent de guide pour effectuer toutes les transformations, tous les sauts dont j'ai parlé, depuis un énoncé quelconque jusqu'à l'ensemble de ses harmoniques sémantiques. Nous pouvons naturellement, lorsque l'on prend chaque énoncé isolément, essayer de donner une sorte de modèle des règles en question. Mais ces règles ne vaudront que pour l'exemple particulier, et il faudrait au fond traiter un par un les innombrables types d'énoncés que peuvent constituer toutes les combinaisons deux à deux, trois à trois, n à n, de différents mots judicieusement choisis dans les langages naturels. Cela ne me paraît pas du tout utopique, et je ne veux pas dire que cela ne soit pas faisable, ni même souhaitable, au contraire. Il existe d'ailleurs une discipline qui n'est pas encore fondée, mais qui n'attend pour naître que des spécialistes : c'est ce qu'on pourrait appeler la sémantique structurale. Elle entreprendrait au fond sur les sémantèmes - c'est-à-dire non plus sur les formes littérales, comme vous l'avez entendu ces derniers jours, mais sur les significations que recouvrent ces formes - elle entreprendrait donc sur les sémantèmes des recherches de modèles, des recherches de structures, selon les mêmes méthodes que la phonologie ou que la morphologie structurale ou, d'une façon générale, la linguistique structurale, mais au niveau des concepts

et non des formes. Malheureusement, pour jouer ce jeu, il faudrait au fond supposer le problème résolu, c'est-à-dire disposer déjà d'un code conceptuel qui permette de traduire de façon économique, mais sans perte d'information, le contenu sémantique des phrases du langage naturel. Si nous possédons cet inventaire, nous pourrions en effet envisager de rechercher, avec le concours de certaines machines, les correspondances entre les différentes formulations littérales et leur contenu sémantique, autrement dit de retrouver, comme je l'ai fait ici, les rapports entre les notions dégagées du texte "responsabilité", "talion", etc.. - et chacune des combinaisons, deux à deux, trois à trois, n à n, des différents termes littéraux. Nous en sommes aujourd'hui très loin; mais déjà quelques personnes ont en vue cette sémantique structurale. Si cette science se constitue, je crois qu'elle sera d'un très grand apport pour la documentation en général et surtout pour la documentation automatique, dont je ne vois pas du tout comment elle pourrait exister si l'on ne disposait pas, au préalable, de ces connaissances sur les transformations sémantiques.

Je passe donc sur cet aspect-là de nos problèmes, puisqu'il est encore dans les limbes, pour m'attacher au second, plus classique, à savoir l'exposé des différentes méthodes qui sont actuellement utilisées pour conserver la structure relationnelle d'un texte exprimé sous une forme rigoureusement analytique. Autrement dit, dans le cas des diagrammes que vous a montrés Leroy, comment, lorsqu'on "atomise" ce diagramme, peut-on conserver les interrelations entre tous ses éléments pris deux à deux, trois à trois, etc..

Dans les sciences humaines, le problème se pose exactement comme ailleurs; je prendrai l'exemple de la phrase précédente. Si l'on se borne à enregistrer sous une forme quelconque chacun des termes principaux de la phrase de base - le "palais", le "roi", l'"architecte" et les verbes "tuer", "écrouler", "emmurer" - il est évident que l'on risque d'obtenir, au moment de la recherche, certaines combinaisons fausses : "l'architecte tue le roi", "le roi tue l'architecte" etc.. Pour lever ces ambiguïtés, une solution évidente, d'ailleurs évoquée par Leroy ce matin, est d'adjoindre à chaque terme un indice, en quelque sorte une flexion, marquant la place logique de chaque terme dans l'énoncé. Parmi ces indices, les plus nécessaires sont le Sujet et l'Objet, j'entends par là le rôle logique de "Sujet" ou d'"Objet" que peut jouer chaque être dans un énoncé. Il existe d'autres cas, dont le nombre et la nature varient selon le domaine envisagé; mais cette procédure, que nous avons utilisée en fait dans certains secteurs de l'archéologie, n'est pas économique du tout. A la limite, on est obligé, en effet, d'adjoindre à chaque terme autant de flexions, c'est-à-dire de multiplier le nombre des termes autant de fois qu'il y a de cas logiques possibles pour chacun d'eux. Dans l'analyse iconographique, par exemple, c'est-à-dire l'analyse des scènes figurées sur des documents variés - peintures, sculptures, gravures - le nombre de "cas", c'est-à-dire le nombre de positions logiques que peuvent occuper les éléments du vocabulaire descriptif est de sept. Un "arbre"

au cas "Sujet", par exemple, ce peut être le thème unique d'une représentation comme par exemple sur certaines monnaies orientales où l'on trouve pour tout motif un arbre couvrant tout le champ. L'arbre au cas "Objet", ce sera par exemple l'arbre abattu par un héros ou encore l'arbre adoré par certains prêtres etc.. L'arbre au "Locatif", c'est celui dans lequel s'est cachée Europe, par exemple, pour échapper aux poursuites de Jupiter. Un autre cas l'"Instrumental" c'est, par exemple, la branche d'arbre maniée par un héros pour terrasser un fauve. La branche au "Qualificatif" enfin, c'est par exemple celle que porte en guise d'emblème un roi, une divinité agraire, etc..

Voilà donc 5 cas auxquels il faudrait envisager la plupart des éléments de l'analyse iconographique. Les conséquences de cette multiplication sont de deux ordres: elle entraîne, du point de vue strictement matériel, une augmentation considérable du nombre de termes dans l'analyse; et du point de vue opératoire, l'obligation, lorsque l'on recherche certains éléments indépendamment de leur fonction logique, d'examiner chaque fois plusieurs termes correspondant à ces différentes fonctions. C'est la raison pour laquelle nous ne nous sommes pas arrêtés à cette solution après l'avoir pratiquée pendant quelque temps. Celle que nous préférons, aujourd'hui, consiste à décliner - car il s'agit bien de déclinaisons - non pas chaque terme du vocabulaire analytique, mais seulement les grandes catégories sémantiques auxquelles ils appartiennent. Je m'explique. Imaginons un système analytique comportant quelques centaines d'éléments descriptifs, signalant toutes sortes d'êtres particuliers animés ou inanimés. Pour faciliter les recherches, il est presque toujours nécessaire de regrouper ces éléments en certaines catégories ontologiques - par exemple, les chiens, les chevaux, les lions, dans la catégorie des "animaux"; les cruches, les jarres, les amphores, dans la catégorie des "récipients" etc.. Une idée vient alors à l'esprit, c'est d'appliquer les flexions logiques non pas à chacun des termes du lexique, mais uniquement aux quelques grandes catégories récapitulatives du genre de celles que je viens d'indiquer. L'avantage de cette procédure n'est pas seulement d'aboutir à une très grande économie d'expression, mais aussi, et surtout, de donner accès à trois niveaux de recherche: un niveau purement lexical d'abord où l'on trouve d'un seul coup, avec les termes non déclinés du vocabulaire spécifique, toutes les occurrences d'un élément particulier; un second niveau ensuite, celui des termes génériques, déclinés, qui permet d'envisager les nombreux rapports qu'entretiennent les uns avec les autres les être appartenant à certaines catégories générales - par exemple, "les confrontations entre les monstres et les animaux domestiques" dans l'iconographie orientale indépendamment de la nature particulière des uns et des autres. Un troisième niveau enfin est fourni par la conjonction des deux précédents; il permet d'ajouter autant de spécifications lexicales que l'on veut (premier niveau) à une relation logique d'ordre général (deuxième niveau).

L'économie et la simplicité de cette démarche sont grandes; mais j'introduis tout de suite une réserve. Lorsque nous utilisons conjointement un terme lexical invariant et la désignation de la catégorie

à laquelle il appartient, il n'y a en fait qu'une probabilité pour que le terme en question soit bien au cas indiqué pour cette catégorie. Il arrive en effet que l'on ait sur une même image, ou dans un même texte, deux êtres appartenant à la même catégorie, mais occupant chacun des positions logiques différentes. C'est un compromis entre une sorte de langage idéal, mais probablement incommode à manier, qui rendrait compte de toutes les informations contenues dans une image ou dans un texte et d'autre part un code plus condensé, que l'on a pu simplifier dans la mesure où les propriétés naturelles du domaine envisagé suffisent à éliminer, sinon toutes les ambiguïtés, du moins la plupart d'entre elles.

J'ai voulu donner un exposé général de ces méthodes avant d'aller plus loin dans le détail, comme nous le ferons cet après-midi au cours d'une discussion de quelques cas concrets.

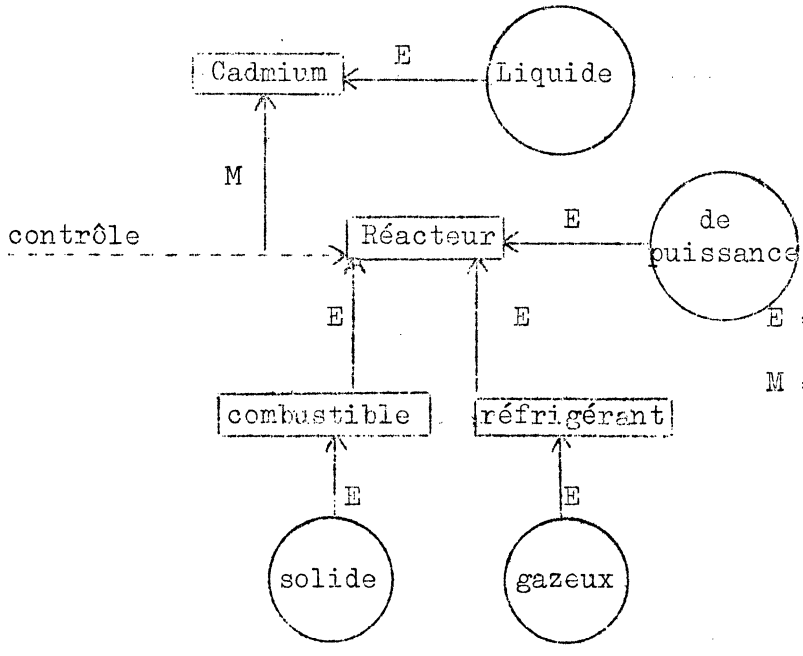
Nous verrons alors comment se posent ces problèmes d'analyse, et quels sont les aspects particuliers des solutions générales que j'ai indiquées ce matin.

DISCUSSION SUR LES SYSTEMES DOCUMENTAIRES

M. Gardin — Première question, que j'ai d'ailleurs déjà posée indirectement à M. Leroy, et que je lui pose maintenant directement : quelle est la méthode envisagée pour conserver les structures linguistiques formelles, lorsque l'on passe d'un diagramme envisagé sous sa forme synthétique, tel qu'il a été exposé ce matin, à une forme atomistique, en vue d'un enregistrement sur machine en particulier ?

M. Leroy — Je pense qu'une bonne façon de répondre à cette question consiste à traiter un exemple, tel qu'il l'avait déjà été à l'occasion de la Conférence de Francfort en juin 1959 (ADIA).

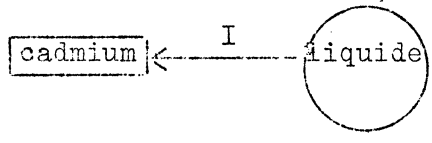
Supposons donc que nous ayons à transformer le diagramme suivant, simplifié pour les besoins de l'explication :



(Contrôle au moyen de cadmium liquide d'un réacteur à combustible solide et réfrigérant gazeux).

---

(1) A l'heure actuelle nous distinguons le cas particulier des relations d'état (I) des relations E .

exemple : 

Le fait que la machine qui était à notre disposition était une calculatrice IBM 650 nous avait conduits à adopter une disposition "à trois colonnes".

Dans la 1ère colonne nous avions en premier lieu des objets.

Exemple :

Réacteur
Cadmium
Combustible
Réfrigérant

Dans la 3è colonne nous mettions les propriétés ou objets liés aux objets de la 1ère colonne par la relation E. (1)

Réacteur		de puissance
Réacteur		Combustible
Réacteur		Réfrigérant
Combustible		solide
Réfrigérant		gazeux
Cadmium		liquide

De cette manière, nous avions déjà indiqué par exemple que le réacteur dont il était question possédait un combustible (2ème ligne) et que ce combustible était solide (4è ligne)

La deuxième colonne était celle des relations en contact avec les objets de gauche et contenait donc le symbole se rapportant à cette relation suivi d'un chiffre indiquant le sens de la flèche.

---

(1) Il avait fallu, bien entendu, prendre la précaution d'indiquer la qualité des éléments de la 3è colonne. En effet, le mot solide par exemple peut être considéré à la fois comme objet ou comme propriété si on ne fait pas intervenir un signe distinctif adéquat. Nous nous contenterons ici de commencer les mots-objets par une lettre majuscule.

Ainsi la disposition :

Réacteur	Contrôle 2	
Cadmium	M 2	
Contrôle	M 1	

indiquait qu'il s'agissait du contrôle d'un réacteur au moyen de cadmium.

(remarque : L'adjonction du mot contrôle - qui n'est pas un objet - dans la colonne de gauche était nécessaire pour indiquer le point de départ de la relation M).

La superposition des différents tableaux donne la carte que nous perforions (en remplaçant chacun des mots par des nombres) pour le document analysé et qui portait le numéro de celui-ci et même le numéro de la phrase-clé, car nous avions très souvent plusieurs phrases-clés par document (environ 5 en moyenne); en effet, les documents que nous avions analysés étaient des rapports de l'U S A E C traitant souvent de plusieurs sujets et nécessitant donc la mise en évidence d'au moins une phrase-clé par sujet.

Réacteur	Contrôle 2	de puissance
Réacteur	Contrôle 2	Combustible
Réacteur	Contrôle 2	Réfrigérant
Combustible		solide
Réfrigérant		gazeux
Cadmium	M 2	liquide
Contrôle	M 1	

La programmation avait été réalisée de façon à obtenir un classement par mot-clé.

Ainsi, si nous supposons que la carte ci-dessus correspondait à la phrase-clé n° 1, nous avions la disposition finale :

Carte	N° DOC.		N° phrase-clé	Autres indications
Réacteur	001	-	1	- Contrôle 2 - de puissance
	001	-	1	- Contrôle 2 - Combustible
	001	-	1	- Contrôle 2 - Réfrigérant
Combustible	001	-	1	- solide
Réfrigérant	001	-	1	- gazeux

De la sorte, lorsqu'une question sur le contrôle d'un réacteur à combustible solide et réfrigérant était posée, il fallait faire le diagramme correspondant et le transformer suivant le procédé que nous venons d'exposer en laissant évidemment vide la place du n° du document. La sélection s'opérait alors par un simple procédé de comparaison des lignes de la carte-question et des cartes-mémoires et par un contrôle de l'identité de l'ensemble : "N° Doc. N° phrase-clé" pour toutes les dispositions mises ainsi en évidence.

Je ne pense pas qu'il soit utile de rentrer ici davantage dans les détails; j'aurai l'occasion de montrer lors de mon prochain exposé comment cette méthode rentre dans le cadre d'une méthode plus générale. Mais il me semble que cela permet de voir déjà comment on peut transformer un diagramme de manière à ce qu'il soit utilisable sur machine.

M. Gardin — Je remercie Leroy de l'explication; elle m'intéresse doublement, d'abord parce qu'elle répond à la question que j'ai soulevée, et ensuite parce qu'elle présente une analogie de forme avec les solutions que nous avons nous-mêmes employées pour résoudre un problème identique dans le cas de codes intéressant des domaines complètement différents, qu'il s'agisse de textes ou d'images. Prenons par exemple les textes; je m'en tiendrai cette après-midi aux textes pour que nous ne soyons pas trop loin de nos domaines respectifs : textes, dans les sciences naturelles, et textes également dans les sciences humaines. Une première constatation est qu'on a là au fond un exemple d'analyse par propositions d'un énoncé synthétique, afin d'éviter les interférences entre les éléments appartenant à des propositions différentes, sans cependant perdre la valeur sémantique de l'ensemble, toutes ces phrases-clé successives faisant partie d'un même univers, à savoir celui du document considéré. En d'autres

termes, l'analyse procède là sur deux plans, d'une part l'univers global du document - le numéro du document constituant en somme l'unité générale de référence - et, d'autre part, à l'intérieur de cet univers, les différentes propositions successives qui définissent le contenu du document, chaque proposition recevant un numéro propre de façon qu'on ne puisse pas associer les éléments d'analyse relevant d'une proposition à ceux qui relèvent d'une autre. C'est au fond la définition d'une "syntaxe", mais réduite à une simple relation. En d'autres termes, pour reprendre l'exemple de Leroy, dans l'expression "contrôle de réacteur de puissance", il y a d'une part une composante "contrôle de réacteur", une composante "réacteur de puissance", et la conjonction de ces deux composantes dans une même unité, montrant que l'on a probablement affaire au "contrôle des réacteurs de puissance". Je dis "probablement" parce que, même si le risque d'ambiguïté est faible dans les sciences naturelles, c'est-à-dire dans l'analyse d'articles intéressant les sciences naturelles, tel n'est malheureusement pas le cas dans les sciences humaines. La raison est celle que j'ai indiquée ce matin, à savoir que les propositions étant mal définies, et l'ensemble des propositions constituant également une unité mal définie, on est obligé d'envisager des "discours" beaucoup plus grands, c'est-à-dire qu'au lieu d'avoir une moyenne de 5 phrases-clés par document, nous en aurions dans les sciences humaines beaucoup plus. L'imperfection de ces sciences, du point de vue de leur langage, c'est-à-dire l'imprécision des concepts qu'elles utilisent, fait qu'on doit envisager la signification possible de conjonctions plus nombreuses à l'intérieur d'unités plus grandes que celles auxquelles s'arrête en général l'analyse dans les sciences naturelles. C'est pourquoi le problème de l'expression des relations entre ces différentes propositions est vraisemblablement plus complexe dans la littérature des sciences sociales; mais la méthode des diagrammes reste certainement une manière d'introduire de l'ordre dans ces relations.

A la demande des participants, M. Gardin expose ensuite la nature et le fonctionnement d'un code particulier servant à l'analyse des ornements géométriques (1), puis la discussion reprend à propos de cet exposé.

M. Leroy — Je voudrais que J.C. Gardin nous dise en quelques mots s'il voit des applications possibles d'une telle méthode d'analyse dans d'autres domaines que celui des ornements. En particulier, je pense que dans les textes, on doit analyser les textes écrits eux-mêmes, mais aussi on peut avoir à "analyser" certaines formes géométriques,

(1) On pourra se reporter par exemple à : "On the coding of geometrical shapes and other representations, with reference to archaeological documents" de J.C. Gardin, Proceedings of the International Conference on Scientific Information, National Academy of Sciences vol II p. 889 - 901.

certains schémas; la méthode n'est-elle pas applicable dans ce cas ?

M. Gardin — Cette méthode est probablement applicable dans une quantité d'autres cas; dans mon esprit, elle n'est nullement particulière à l'ornementation abstraite. Une première application m'a été suggérée par M. Luhn de la compagnie IBM à New York, qui se demandait si on ne pourrait pas utiliser un système de ce genre pour l'indexation des "Trade-Marks", des marques de patente, dans l'industrie. En effet, il est très difficile de se fonder sur la signification de tels symboles pour les différencier les uns des autres. Leur allure formelle, géométrique ou figurative serait vraisemblablement plus caractéristique.

M. Leroy — Je remercie Gardin de sa réponse; y-a-t-il d'autres questions ?

M. Gutmann — Je voudrais demander à M. Gardin s'il peut mettre clairement en évidence la question suivante : pouvez-vous montrer nettement la différence qui existe entre votre méthode de classification des ornements et une classification classique hiérarchisée telle que la classification décimale.

M. Gardin — La différence majeure, à mes yeux, entre ces systèmes d'analyse et ceux qui régissent la Classification décimale universelle est de deux ordres. D'abord d'ordre pratique; une différence, à mes yeux fondamentale, est que nous n'avons pas du tout affaire ici à des classifications, mais uniquement à des analyses, à des systèmes analytiques, qui peuvent conduire ou ne pas conduire à des classifications. Une classification est un ordre figé une fois pour toutes; je ne veux pas dire que tous ses états soient absolument figés et immuables, mais, du point de vue de la structure et de l'utilisation, il existe des règles qui permettent certaines combinaisons et qui en écartent d'autres. C'est le cas de la Classification décimale universelle où, lorsqu'on forge un néologisme par exemple, on est lié par les règles de combinaison des différentes notions admises dans la classification; tandis que dans le cas de systèmes analytiques, tels que ceux dont nous parlons depuis ce matin, on peut absolument combiner n'importe quoi avec n'importe quoi. Il n'y a aucune règle d'exclusion lorsqu'on envisage la signification possible d'une combinaison particulière de notions existant dans le vocabulaire.

La raison de cette différence est d'ailleurs d'ordre matériel; c'est que, dès que l'on crée une "molécule" particulière, disons une combinaison particulière de ces éléments atomistiques d'un système d'analyse, et qu'on veut la promouvoir au rang de

"rubrique", pour parler en documentaliste, il faut lui donner une place matérielle, soit dans un fichier, soit sur les rayons d'une bibliothèque. Dès lors, on ne peut naturellement pas envisager toutes les combinaisons deux à deux, trois à trois, "n" à "n" de tous les éléments du système analytique qui a servi de base à la classification en question. On ne peut retenir que certaines combinaisons privilégiées, auxquelles on donne une cote. La cote est analytique bien sûr, c'est-à-dire qu'elle n'est pas arbitraire; elle donne une idée des différentes composantes du sujet auquel correspond le compartiment du fichier ou l'étagère de la bibliothèque. Mais enfin, à cause de cette place matérielle qu'occupent toutes les combinaisons de la classification en question, on ne peut être que peu généreux dans la formation de ces concepts combinatoires. Tandis que dans un système analytique tel que celui dont je parle, rien n'empêche de pousser l'analyse aussi loin qu'on le veut, c'est-à-dire d'exprimer les notions ou des phénomènes extrêmement particuliers par certaines combinaisons éphémères de notions préexistantes, sans donner à ces combinaisons aucune place matérielle dans le code, ni moins encore dans la bibliothèque ou le fichier.

En d'autres termes, un système analytique du type de ceux dont nous parlons n'est pas du tout une classification; c'est la somme virtuelle d'autant de classifications que l'on voudra, et elles se comptent par milliards, correspondant chacune à une série de combinaisons particulières envisagées au moment d'une recherche, mais qui rentrent en quelque sorte dans l'anarchie dès que la recherche est achevée. On peut introduire ici une observation supplémentaire. On entend souvent parler de "machines à penser" qui permettent de faire certaines découvertes - je parle uniquement au niveau documentaire; je ne parle pas du tout des machines heuristiques, parfaitement fondées celles-là. Cette conception des machines contient une part de faux et une part de vérité; une part de faux, dans la mesure où l'on ne retrouve jamais plus à la "sortie" que ce qu'on a mis à l'"entrée", et une part de vrai, pour autant que l'on découvre parfois la valeur sémantique d'une combinaison particulière dont tous les éléments ont été enregistrés au moment de l'analyse, mais sans que soit perçue la signification de telle ou telle association. Je prends un cas concret, que j'emprunterai à cette analyse du Coran dont j'ai parlé ce matin, c'est-à-dire à l'inventaire des concepts et combinaisons de concepts contenus dans cet ouvrage. Nous avons, lorsque l'analyse a été achevée, passé en revue, à titre expérimental, certains thèmes de la philosophie chrétienne, en particulier les thèmes de la scolastique. Notre idée était qu'il existait une hétérogénéité totale entre ce texte, que nous croyions connaître parfaitement, et les textes de la philosophie médiévale chrétienne. Or,en formulant certains thèmes tels que "le salut par la foi ou les oeuvres", ou bien, pour prendre des exemples familiers, "tendre la joue droite après la joue gauche", etc... on s'aperçoit que certains passages du Coran contiennent, je ne dirais pas le thème explicite envisagé, mais

au moins tous les éléments conjoints de ce thème, et qu'il y a une présomption pour que ce dernier ait été envisagé par l'auteur, consciemment ou non. Ce n'est pas une présomption, parce qu'encore une fois nous sommes alors devant une relation purement de conjonction, la causalité n'étant pas explicite; mais cette seule conjonction, lorsqu'elle est fréquente, invite à penser qu'il peut exister quelques rapports entre certains thèmes courants dans la littérature chrétienne des premiers siècles et, d'autre part, les données d'un texte apparemment aussi éloigné de celle-ci que le Coran.

Je suis arrivé à cette di gression en partant d'une question relative aux classifications; ce n'est pas un hasard. Dès que l'on essaie d'expliciter la différence entre les classifications traditionnelles et ces systèmes analytiques, on est bien obligé de se poser la question des mérites particuliers qu'offrent les seconds par rapport aux premiers. L'un de ces mérites est de permettre à des spécialistes de formuler certaines hypothèses équivalentes à des combinaisons originales de notions préexistantes, dans un domaine donné et d'observer leur valeur dans ce domaine.

M. De Benedetti — La méthode dont vous nous avez exposé les lignes principales n'est valable que dans l'optique d'un travail de codification effectué par des hommes.

Mais alors comment peut-on envisager de passer à l'automatisation de cette analyse, puisqu'en fin de compte, c'est celà qu'il faudrait absolument réaliser.

M. Gardin — Je crois que la réponse à cette question était déjà contenue, en partie en tous cas, dans l'exposé de ce matin, où j'ai montré qu'il n'y avait actuellement aucune règle connue qui permette de passer de la forme littérale d'un texte à la totalité de son contenu sémantique. Je pense que vous songez à l'analyse automatique, selon des voies par exemple statistiques, comme M. Luhn les a envisagées. Je crois que c'est une démarche en effet intéressante, qui donne des résultats, mais il reste qu'elle ne traite absolument pas du problème que posent les "sauts sémantiques" dont j'ai parlé ce matin, à savoir le passage d'un mot non pas seulement à tous ses synonymes ou à toutes ses acceptions différentes lorsqu'on le considère isolément, mais à l'ensemble des significations sur lesquelles débouche ce mot lorsqu'il est employé en conjonction avec d'autres. J'ai cité ce matin, en guise d'exemple, le problème "talion". Si vous voulez programmer une machine pour qu'elle repère dans les textes anciens tous les cas de "talion" - à savoir l'imposition d'une peine équivalente au délit commis - les instructions seront les suivantes : parcourir tous les textes, observer tous les délits commis, les sentences qui ont suivi, et chaque fois qu'il y a identité entre le délit et la sentence, imprimer "talion". Voilà une règle que l'on peut envisager, en

effet, pour l'analyse automatique si l'on veut qu'elle effectue, et il faut qu'elle l'effectue, ce "saut" dont j'ai parlé.

Le malheur est qu'il y a
je cite ce chiffre au hasard - certainement au moins quelques centaines de milliers de sous-programmes de ce genre pour l'ensemble des termes, et que nous connaissons extrêmement mal les règles sur lesquelles se fonde actuellement l'intuition pour pallier l'ignorance où nous sommes de ces règles. Ce que je maintiens fermement, l'expérience ayant montré l'importance de cette restriction, c'est qu'une analyse qui se situerait exclusivement au niveau littéral, même si elle tient compte du halo sémantique de chaque terme, ne résout absolument pas ce problème des significations nées de la conjonction particulière de 2 ou 3 termes. La notion de "talion" n'est pas du tout contenue dans le halo de la "condamnation à mort", ni dans celui du "crime capital", et pourtant elle résulte de la conjonction de deux notions de ce type, comme elle résulte de la conjonction du fait de couper la jambe à quelqu'un et d'avoir la jambe coupée, et ainsi de suite. J'ai cité aussi le concept de "responsabilité"; celui des "servitudes", en droit, est un autre exemple du genre, où il est absolument impossible, en tous cas aujourd'hui, de formuler un sous-programme, permettant de passer des formes littérales aux composantes sémantiques du premier, du second, du troisième degré etc..., qui sont souvent cependant les principales.

Votre deuxième remarque était, je crois : Si on ne peut pas automatiser les analyses, les méthodes dont je parle ne sont pas applicables en documentation. C'est possible, si l'on a de la documentation une vue très générale, c'est-à-dire si l'on veut envisager de répondre simplement de façon approximative, dans des domaines très larges, à des questions elles-mêmes générales. Mais, dès que l'on se place dans un domaine spécifique, et qu'on veut être en mesure de répondre à des questions fines, je ne crois pas qu'il y ait d'autre solution : on ne peut pas obtenir la précision à partir de l'imprécision. C'est ce qu'on essaie pourtant de faire, du moins est-ce l'hypothèse implicite derrière ces espoirs d'analyse automatique, dans beaucoup de domaines. Nous n'avons pas encore le bagage conceptuel suffisant pour formuler ces programmes d'analyse automatique. Or, il faut bien d'abord les formuler, à moins que l'on imagine un auto-programme de l'analyse automatique; mais je n'y crois guère, et j'espère que M. de Picciotto ou Mme Poyen me contrediront si j'ai tort.

M. Leroy — Je crois qu'après ces quelques mots, l'exposé de Y. Lecerf, qui s'intitule justement : "Analyse automatique", se présente très mal. En vérité, vous savez déjà à quoi vous en tenir parce que vous avez fait des travaux pratiques de linguistique; vous avez vu qu'il s'agissait d'effectuer en premier lieu une analyse syntaxique. Je me permettrais tout de même de dire que nous allons faire des travaux pratiques sur l'"Etablissement de diagrammes" qui ont réellement en vue le problème de l'analyse automatique. Vous verrez qu'à partir des résultats obtenus par M. Lecerf sur machine, donc à partir d'analyses syntaxiques, il est possible de construire des diagrammes. Je ne dis pas que la solution soit effectivement trouvée dès maintenant; au contraire, il faudra encore beaucoup de recherches . Mais on peut déjà construire "automatiquement" des éléments de diagramme et, à l'aide du diagramme général, dont j'aurai encore l'occasion de parler et qui pourra, du moins je l'espère, être construit automatiquement à partir d'un certain stade, la machine pourra tenir compte des problèmes sémantiques que J.C. Gardin vient d'évoquer.

JOURNEE D'ANALYSE

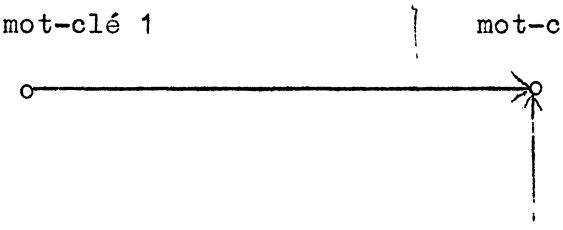
TRAVAUX PRATIQUES SUR L'ETABLISSEMENT DES DIAGRAMMES

M.DETANT Y. LECERF A. LEROY

PREMIERE PARTIE: ETABLISSEMENT DES DIAGRAMMES D'APRES LE SENS

Les diagrammes

Ils sont constitués essentiellement par des mots reliés par des flèches qui correspondent elles-mêmes à des mots jouant un rôle relationnel.



Ces flèches peuvent par exemple représenter des "actions" telles que :

Absorption
Décomposition
Préparation
Corrosion
Estimation

Ces actions jouent le rôle des verbes du langage réduit. Les mots-clés reliés par les flèches sont en premier lieu des noms d'objets ou d'entités qui correspondent aux substantifs.

Ex. Uranium)
Eau) objets
Neutron)
Electricité) entité.

Certains mots jouent le rôle d'adjectifs et d'adverbes; ils expriment respectivement les propriétés qui se rapportent aux objets et les "qualificatifs" qui précisent les actions.

Ex. modérateur organique
(objet) (propriété)

production en série
(action) (qualificatif)

Enfin le rôle des prépositions et de certaines conjonctions est tenu par ce qu'on appelle des relations (autres que les actions) auxquelles peuvent s'adjoindre également des qualificatifs.

Ex.: un ciel très bleu
(objet) (qualificatif)(propriété)

Les grandes catégories de mots qui apparaissent se rapportent donc :

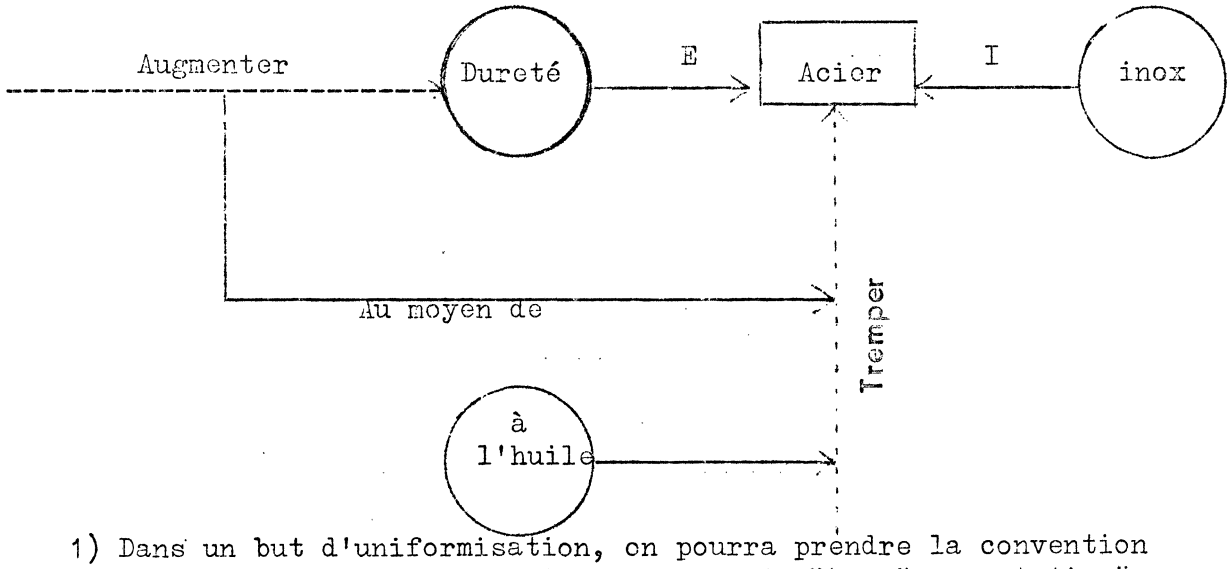
a) aux objets (ou entités)
b) aux actions
c) aux relations
d) aux propriétés et qualificatifs se rapportant aux trois catégories précédentes.

Exemple de diagramme

Il nous est dès lors facile de représenter les diagrammes caractéristiques. Pour plus de commodité, nous fixerons les conventions suivantes :

- les objets seront inscrits dans des rectangles
- les propriétés et qualificatifs dans des cercles
- les actions seront représentées par des flèches en traits pointillés au dessus desquels on inscrit le verbe correspondant
- les autres relations, comme par exemple celles qui interviennent entre des objets et leurs propriétés, seront représentées par des flèches en trait plein.

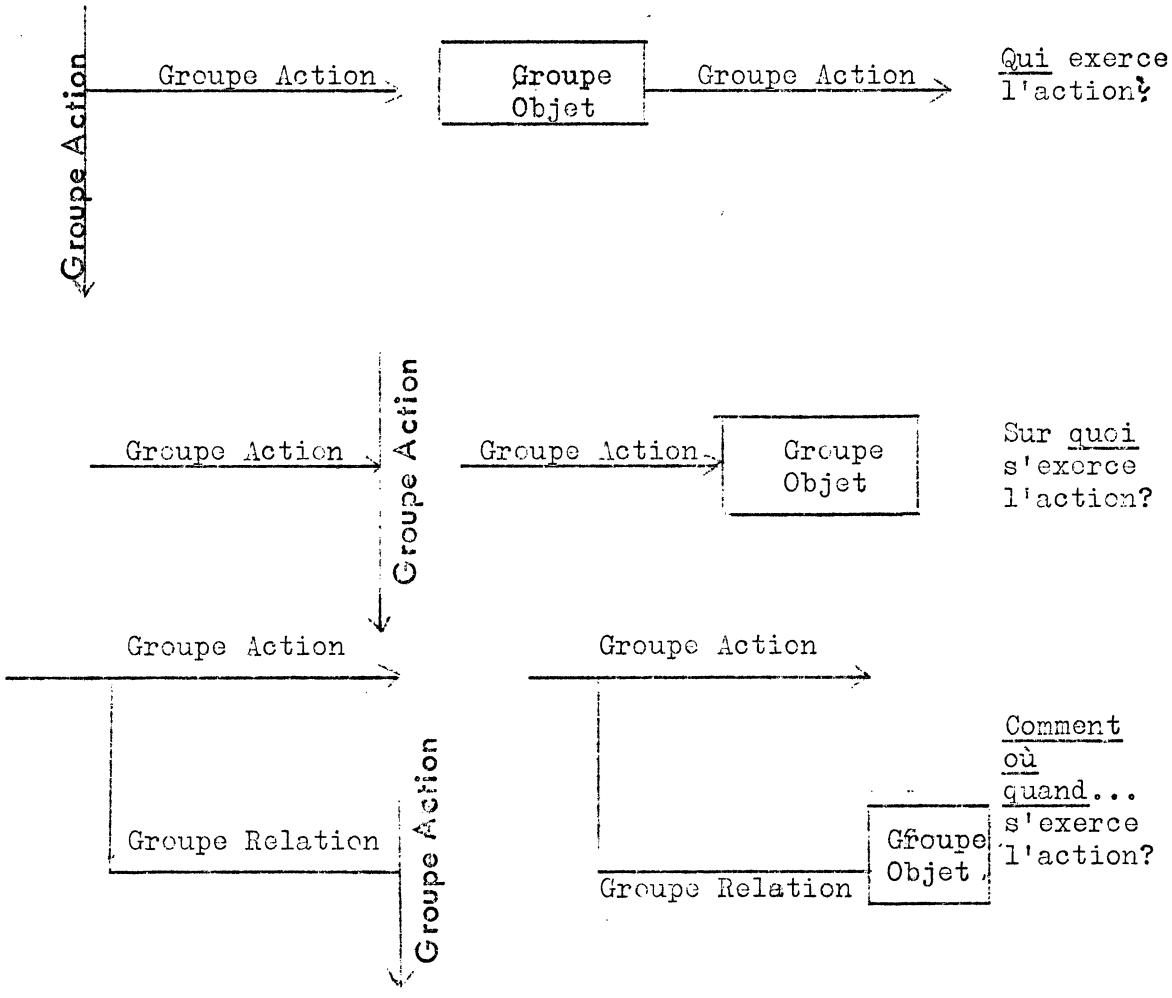
Ex : Augmentation(1) de la dureté de l'acier inoxydable au moyen d'une trempe à l'huile



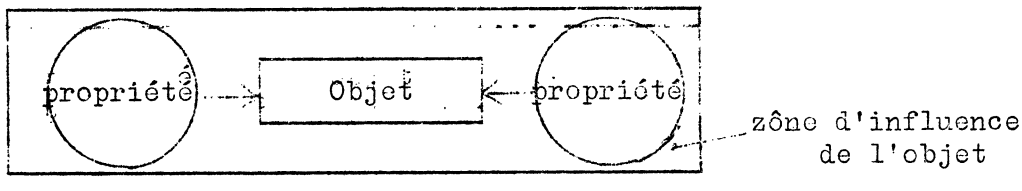
1) Dans un but d'uniformisation, on pourra prendre la convention de mettre "augmenter" pour "action d'augmenter" ou "augmentation".

Diagrammes élémentaires

Si l'on convient de considérer l'ensemble formé par un objet et ses propriétés comme un "groupe objet", l'ensemble formé par une action et les qualificatifs correspondants comme un "groupe action" et l'ensemble formé par une relation et ses "qualificatifs" comme un "groupe relation", les diagrammes élémentaires possibles sont les suivants :



Remarque : La disposition "en groupe"fait apparaître la notion de zône d'influence des mots développés au cours des travaux pratiques de linguistique.



Groupes "objet" et "action"

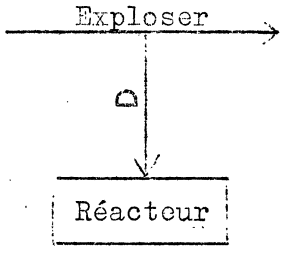
Pour respecter la pensée de l'auteur on doit se situer au même niveau descriptif que lui, c'est-à-dire utiliser ses expressions ou des expressions synonymes ne faisant pas intervenir de mots d'un autre niveau. Si, par exemple, on traite des qualités de prix d'un modérateur il n'y a pas lieu de faire appel à des notions de ralentissement de neutrons sur le diagramme correspondant.

On a donc intérêt à utiliser les notions du texte et à ne transformer que celles pour lesquelles on a décidé d'utiliser des synonymes.

Groupe relation

Il concerne les rapports de temps, de lieu, de circonstance, de but, de cause etc... Les relations correspondent très souvent aux prépositions et à certaines conjonctions.

ex:	à	relation	A
	dans	-	D
	vers	-	V
	sur	-	SU
	sous	-	SO
	après	-	AP
	avant	-	AV
	depuis	-	DE
	jusqu'à	-	JU
	au moyen de	-	M
	pour	-	P
	etc.....		

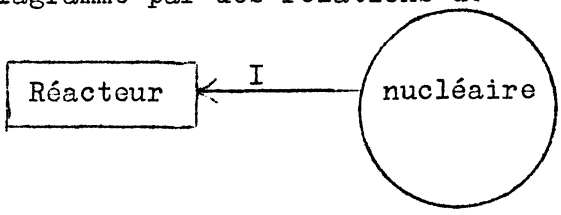


Il n'y a bien sûr pas lieu d'utiliser toutes les prépositions, et il faut au contraire ne garder que celles qui sont fondamentales, qui ont seules une raison d'exister d'un point de vue logique. Il est nécessaire de toujours s'interroger sur la véritable nature

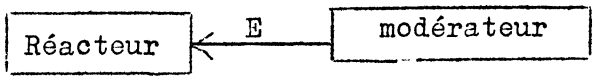
de la relation. (ex.: La préposition "à" peut correspondre dans certaines conditions - je vais à Paris - à la relation "V").

Remarque : Certains verbes du langage ordinaire - être, avoir, posséder etc - se traduisent sur le diagramme par des relations de type I et E.

exemple : Relation d'identité (I)



Relation d'appartenance (E)



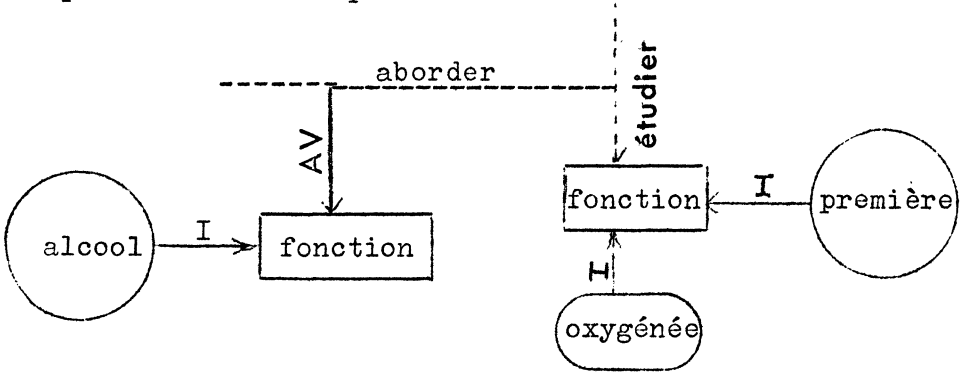
(La flèche est dirigée vers l'objet le plus général)

TRAVAIL PRATIQUE :

Mettre sous forme de diagrammes les phrases suivantes en inventant si nécessaire d'autres relations que celles que nous avons données.

- Le bleu, le vert et le rouge sont trois couleurs fondamentales dont le mélange fournit pratiquement l'ensemble des couleurs possibles.
- Avec la fonction alcool, nous abordons l'étude de la première fonction oxygénée.
- Les alcools résultent de la substitution de groupements oxhydriles OH monovalents aux hydrogènes des carbures.
- Si dans la formule d'un alcane on remplace un seul atome d'hydrogène par un oxhydrile, on obtient la formule d'un monoalcool.

SOLUTION - Ex. pour la deuxième phrase :



Seuls les articles et le pronom personnel "nous" n'apparaissent pas sur le diagramme, car ils n'ont pratiquement aucune importance.

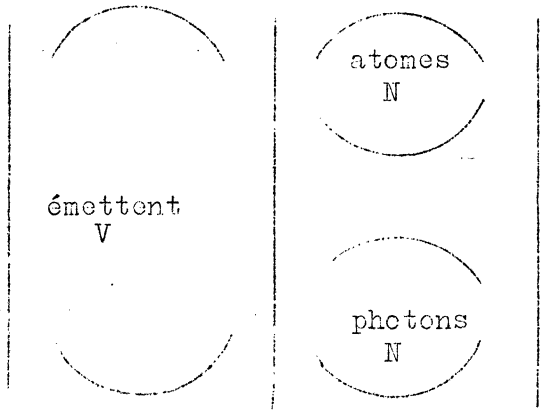
DEUXIEME PARTIE: ETABLISSEMENT DES DIAGRAMMES D'APRES LA FORME

SIMILITUDE DU LANGAGE ORDINAIRE ET DU LANGAGE DES DIAGRAMMES

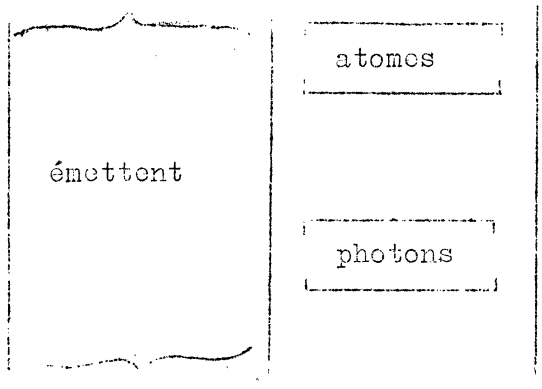
Dans les pages précédentes il a été souligné que les objets correspondaient très souvent aux substantifs, les propriétés aux adjectifs, les relations aux prépositions etc... Il en résulte que la considération de la seule nature d'un mot et de sa place par rapport aux autres fournit déjà une aide pour passer du langage ordinaire à celui des diagrammes.

Les travaux pratiques de linguistique avaient préparé le terrain à une telle transformation. En effet, le tableau obtenu "automatiquement" dans la cinquième partie mettait en évidence la structure polydimensionelle du langage et la possibilité de schématiser cette structure par des diagrammes plans.

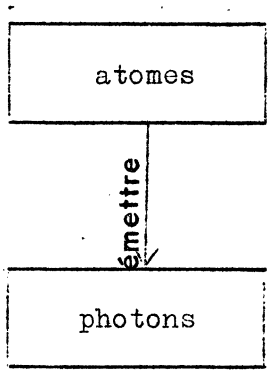
Prenons l'exemple simple du tableau de la sixième partie en effaçant la dernière colonne qui n'apparaît pas sur le diagramme.



Il peut se représenter d'une autre manière



Or le sujet traité "les atomes émettent des photons" correspond au diagramme suivant :



La similitude de structure apparaît nettement.

OUTILS NECESSAIRES POUR LE PASSAGE DU LANGAGE ORDINAIRE A CELUI DES DIAGRAMMES

Pour réaliser ce passage il faut posséder :

- un certain nombre de consignes

- un dictionnaire

Nous donnons ci-après un échantillon très simplifié de ce que pourraient être de telles consignes et un tel dictionnaire.

CONSIGNE GENERALE

Lire, un par un, les mots du tableau par colonnes. Pour cela on lit à la suite tous les mots d'une même colonne avant de passer à la suivante. Les colonnes sont ainsi explorées une par une en commençant par celle de gauche.

A chaque mot lu, on applique la sous routine 1 puis la sous routine 2.

Sous routine 1

(Porter sur le diagramme un dessin correspondant à un mot du tableau par colonnes)

- Le dictionnaire indique la famille sémantique d'un mot (d'où la façon de faire le dessin cherché) et l'inscription à porter sur le dessin.
- Le mot donné est-il souligné ? Si oui, c'est que le dessin correspondant a déjà été porté dans le diagramme à une autre occasion. Si non, on le souligne et on fait le dessin en question.

Sous routine 2

(Recherche des opérandes prioritaires)

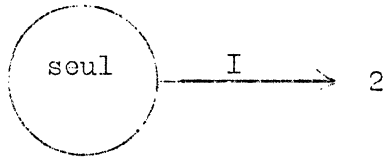
- Le dictionnaire indique la famille sémantique du mot d'où la liste de ses opérandes et la façon de trouver ceux-ci. (Le dictionnaire donne aussi parfois des consignes particulières).
- On cherche les opérandes c'est-à-dire les mots seuls aptes à occuper des places marquées sur le dessin. S'ils existent dans le tableau par colonnes, on leur applique la sous routine 1 puis la sous routine 2 s'il y a lieu.

LISTE DES FAMILLES SEMANTIQUES

1) PROPRIETE

Dessin

Une flèche et un cercle contenant l'inscription mentionnée dans le dictionnaire.

Exemple : 

Place marquée

La pointe de la flèche. (marquée 2).

Opérande 2

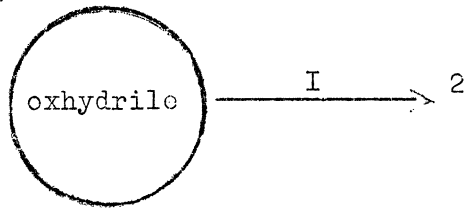
On le trouve en explorant dans le tableau par colonnes les colonnes précédentes en regard du mot considéré. On prendra comme opérande 2 le premier mot qui ne soit pas associé à un cercle (c'est-à-dire qui ne soit pas une propriété ou un objet - apposition).

2) OBJET:

Voir si le mot en regard dans la colonne précédente est associé à un rectangle.

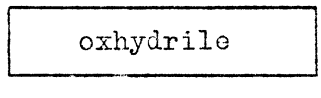
si oui

le nom joue le rôle d'apposition et sera traîté comme une propriété.

Exemple : 

si non

Il s'agit d'un objet proprement dit, on le note dans un rectangle; il ne comporte pas de place marquée

Exemple : 

3 ACTION

Dessin

Flèche, à côté de laquelle on porte l'inscription mentionnée dans le dictionnaire.

Exemple : 1 ——obtenir——→ 2

Sauf indication du contraire dans le dictionnaire, les opérandes 1 et 2 sont les deux premiers substantifs de la colonne suivante. On déclenche alors une exploration permettant de voir s'il y a ou non inversion.

Cas particuliers

Il peut arriver que l'action nécessite pour être décrite, l'emploi d'un schéma composé.

Exemple : 1 ——remplacer——→ 2 pour le mot substitution

M ↓ 3

le dictionnaire l'indique.

4 RELATIONS

a) Voir si la relation s'introduit en opérande 1 ou en opérande 2 d'une action. Si oui, elle n'est pas notée, c'est le mot en regard dans la colonne suivante qui est directement porté en opérande 1 ou en opérande 2.

b) Si non, la relation est notée à l'aide d'une flèche.

Exemple : 1 ——M——→ 2

Sauf indication du contraire :

opérande 1 = mot en regard colonne précédente

opérande 2 = mot en regard colonne suivante.

5 FAMILLE NT

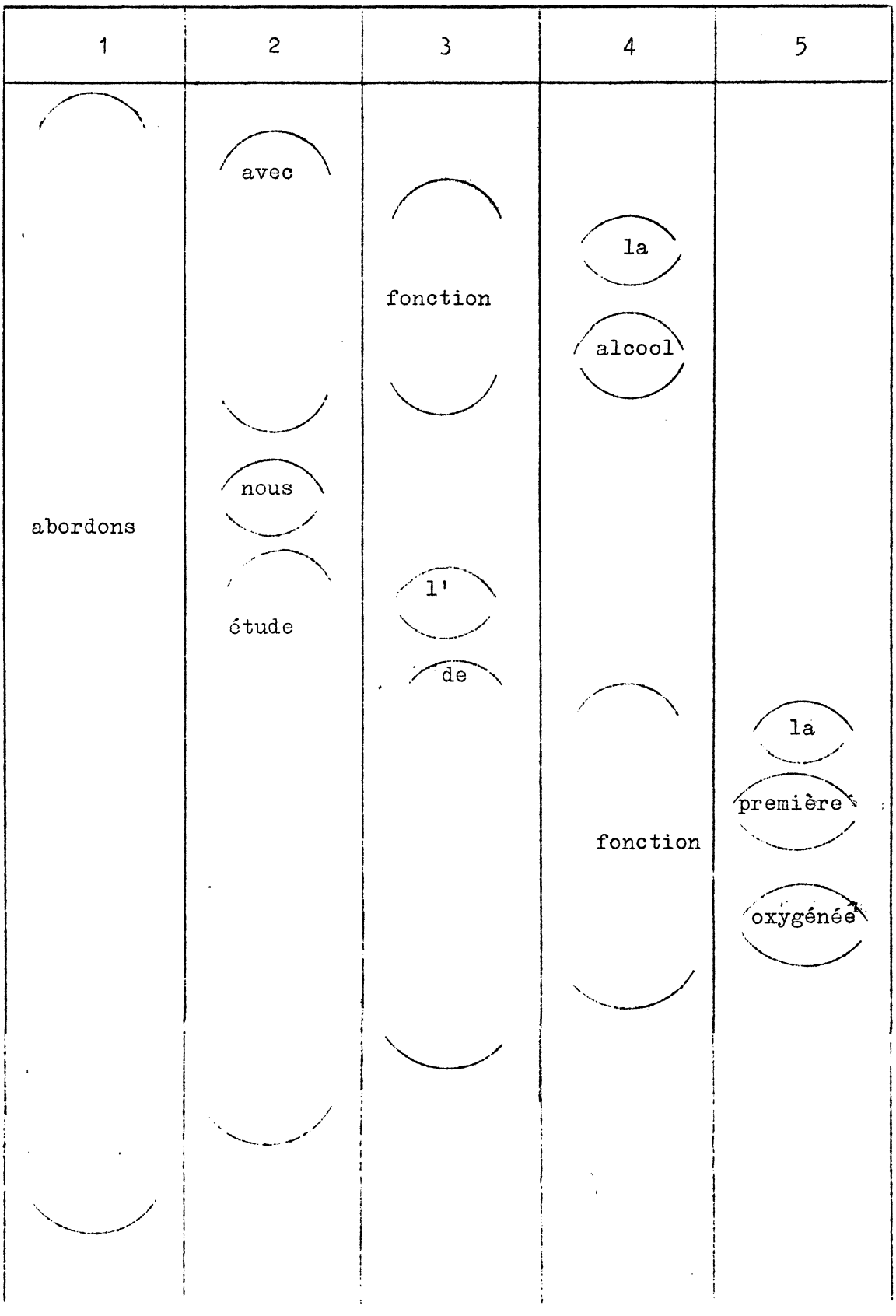
Mots non transcrits sur le diagramme.

TRAVAIL PRATIQUE :

Un ordinateur électronique a dressé les tableaux correspondant aux trois dernières phrases de l'exercice précédent. A partir de la consigne générale et avec l'aide des extraits du dictionnaire donnés ci-après construire les diagrammes correspondants.

PREMIERE PHRASE

1. Tableau



PREMIERE PHRASE

2. Extrait du dictionnaire

Tableau par Colonnes

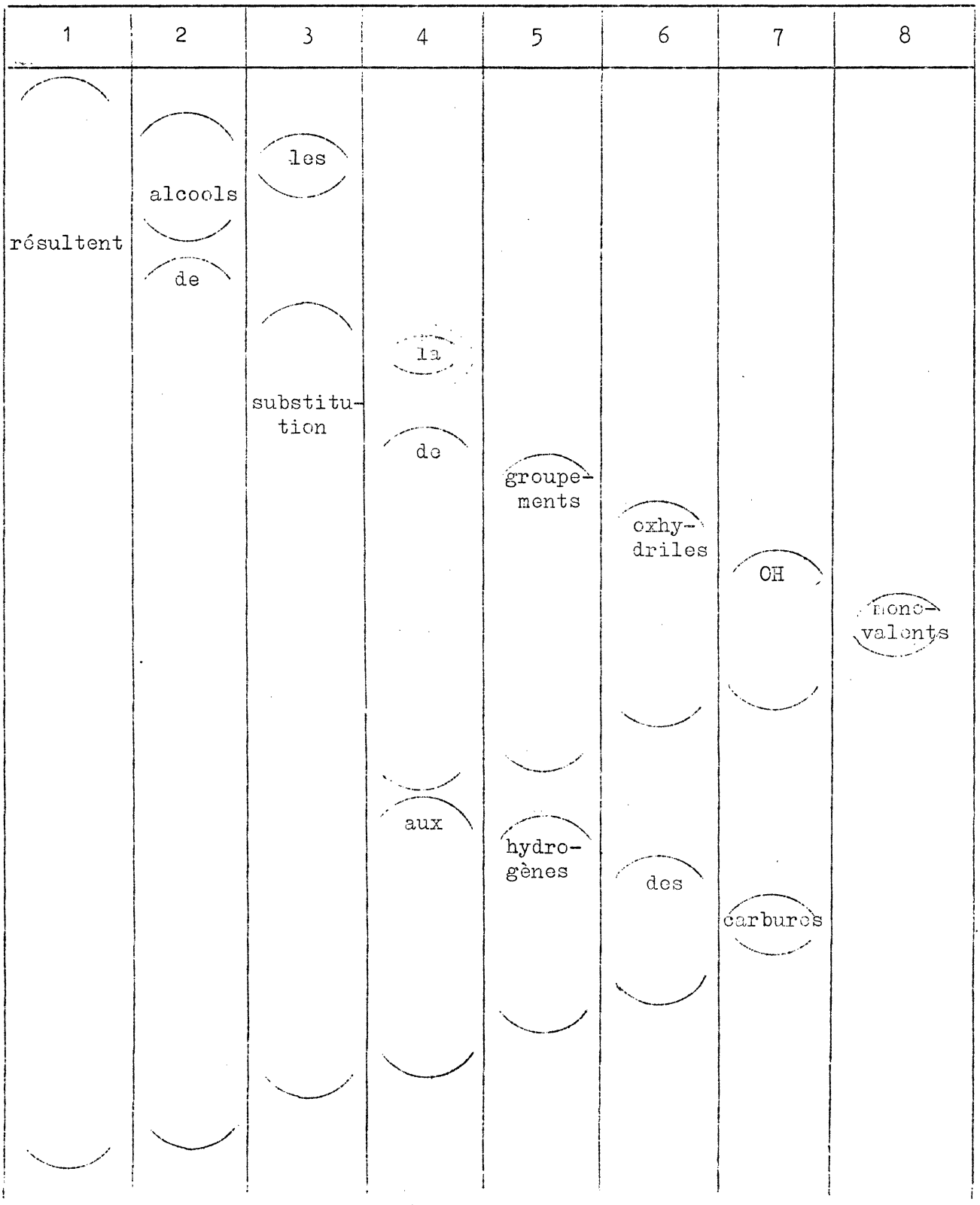
Mot	famille grammaticale
abordons	verbe
avec	préposition
nous	substantif
étude	substantif
fonction	substantif
l'	article
de	préposition
la	article
alcool	substantif
fonction	substantif
la	article
première	adjectif
oxygénée	adjectif

Diagramme

inscription à porter sur le dessin	famille sémantique du dessin	consignes particulières
aborder	action	
AV	relation	
	NT	
étudier	action	op. 1 = par op. 2 = de en colonne suivante
fonction	objet	
	NT	
E	relation	
	NT	
alcool	objet	
fonction	objet	
	NT	
première	propriété	
oxygénée	propriété	

DEUXIEME PHRASE

1. Tableau



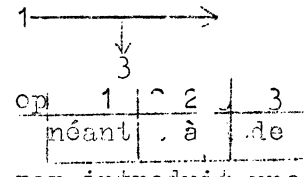
DEUXIEME PHRASE

2. Extrait du dictionnaire

Tableau par colonnes

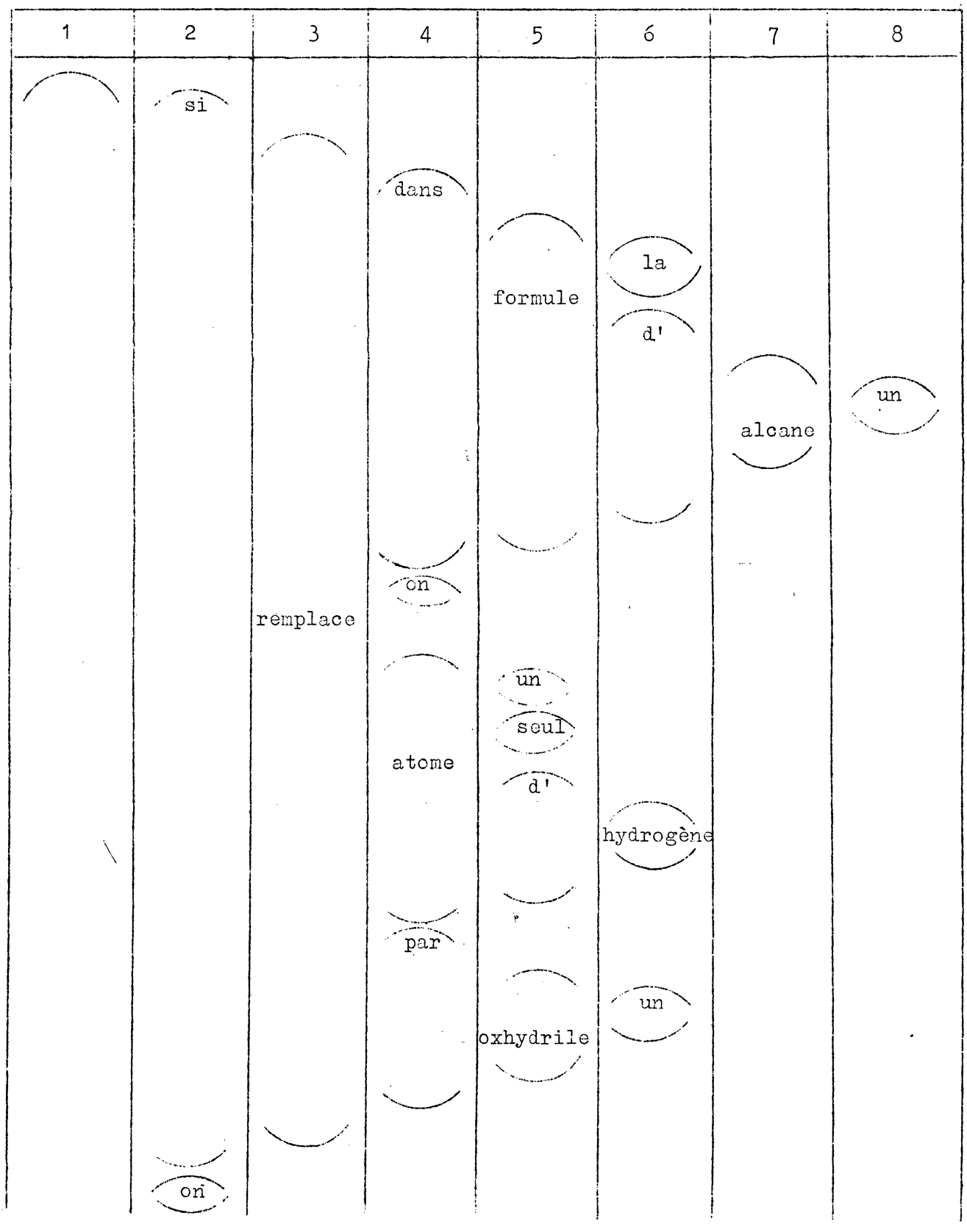
Mot	famille grammaticale
résultent	verbe
alcools	substantif
de	préposition
les	article
substitution	substantif
la	article
de	préposition
aux	préposition
groupements	substantif
hydrogènes	substantif
oxhydriles	substantif
des	préposition
OH	substantif
carbures	substantif
monovalents	adjectif

Diagramme

inscription à porter sur le dessin	famille sémantique du dessin	consignes particulières
R	relation	op.2=1°substantif colonne suiv. op.1=de en colonne suivante
alcools	objet	
E	relation	
	NT	
remplacer	action	 par introduit une ambiguité
	NT	
E	relation	
à	relation	
groupements	objet	
hydrogènes	objet	
oxhydriles	objet	
E	relation	
OH	objet	
carbures	objet	
monovalents	propriété	

TROISIEME PHRASE

1. Tableau



TROISIEME PHRASE

1. Tableau

1	2	3	4	5	6	7	8
obtient	formule	la					
		d'					
			monoal- cool	un			

TROISIEME PHRASE

2. Extrait du dictionnaire

Tableau par colonnes

Mot	famille grammaticale
Obtient	verbe
si	conjonction. de subordination
on	substantif
formule	substantif
remplace	verbe
la	article
d'	préposition
dans	préposition
on	substantif
atome	substantif
par	préposition
monoalcool	substantif
formule	substantif
un	article
seul	adjectif
d'	préposition
oxhydrile	substantif
un	article
la	article
d'	préposition
hydrogène	substantif
un	article
alcane	substantif
un	article

Diagramme

inscription à porter sur le dessin	famille sémantique du dessin	consignes particulières
obtenir	action	
Q	relation	
	NT	
formule	objet	
remplacer	action	
	NT	
E	relation	
D	relation	
	NT	
atome	objet	
M	relation	
monoalcool	objet	
formule	objet	
	NT	
seul	propriété	
E	relation	
oxhydrile	objet	
	NT	
	NT	
E	relation	
hydrogène	objet	
	NT	
alcane	objet	
	NT	

SOLUTION : Ex. pour la phrase : Avec la fonction alcool, nous abordons l'étude de la première fonction oxygénée.

On lit le mot de la colonne de gauche : abordons

Application de la sous-routine 1 pour le mot : abordons

Le dictionnaire indique la famille sémantique = action

d'où le dessin : 1 --aborder--> 2

Application de la sous-routine 2 pour le mot : abordons

Les opérandes 1 et 2 sont les deux premiers substantifs de la colonne suivante = nous, étude

Application de la sous-routine 1 pour le mot : nous

Le dictionnaire indique la famille sémantique = NT

Le mot "nous" ne doit pas être porté sur le dessin.

Application de la sous-routine 1 pour le mot : étude

Le dictionnaire indique la famille sémantique = action

d'où le dessin =

1 --étudier--> 2 et aborder, 2 étudier 1

Application de la sous-routine 2 pour le mot : étude

Le dictionnaire donne des consignes particulières : opérande 1 = par
opérande 2 = de

Seul le mot de existe (dans la colonne suivante)

Application de la sous-routine 1 pour le mot : de

Le dictionnaire indique la famille sémantique = relation

Le mot de s'introduit en opérande 2 de l'action étudier; il n'est donc pas noté, c'est le mot en regard dans la colonne suivante "fonction" qui est directement porté en opérande 2.

Application de la sous-routine 1 pour le mot : "fonction"

Le dictionnaire indique la famille sémantique = objet

Le mot en regard "de" dans la colonne précédente n'étant pas associé à un rectangle, on note l'objet dans un rectangle qui ne comporte pas de place marquée :

[fonction] et aborder, étudier, [fonction]

etc...

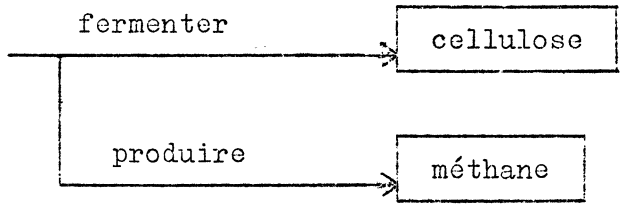
En continuant d'appliquer les consignes, on arrive au diagramme indiqué dans la solution aux travaux pratiques de la première partie.

TROISIEME PARTIE: SUPERPOSITION DES DIAGRAMMES

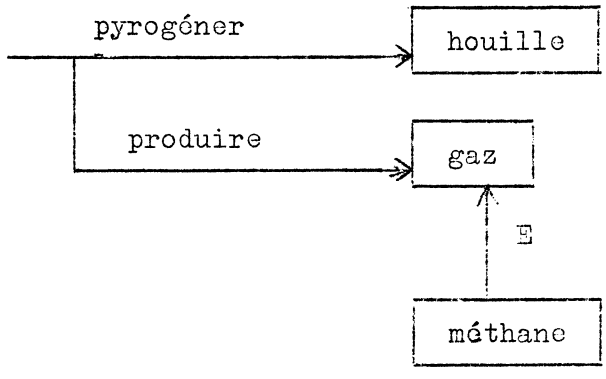
LIAISONS ENTRE DIAGRAMMES

Jusqu'ici nous nous sommes efforcés d'obtenir des diagrammes correspondant à des phrases isolées. Mais il n'y a logiquement pas lieu de garder cette subdivision; un texte est normalement un tout qui ne possède de ponctuation que pour les commodités de la lecture. Il faut donc chercher les liaisons possibles entre phrases et "superposer" les diagrammes isolés.

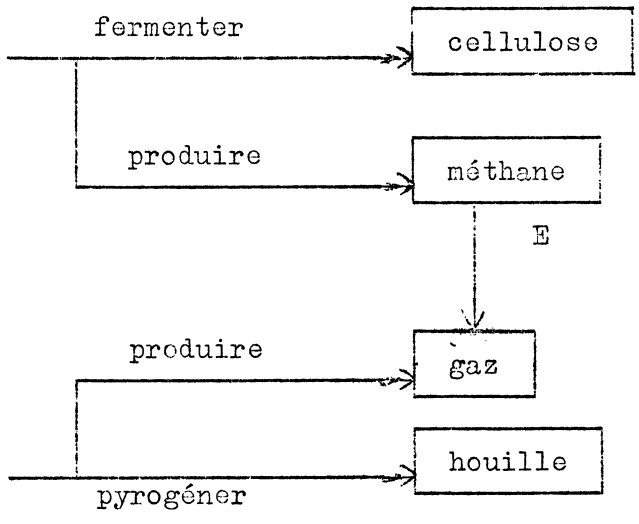
Exemple : a) La fermentation de la cellulose fournit du méthane



b) La pyrogénation de la houille produit un gaz qui contient du méthane.



Il semble logique de "superposer" les 2 diagrammes.



Mais il faut bien prendre garde de ne pas créer ainsi une information fausse - C'est ainsi que la superposition ci-dessus n'est valable que si elle permet d'exprimer la phrase :

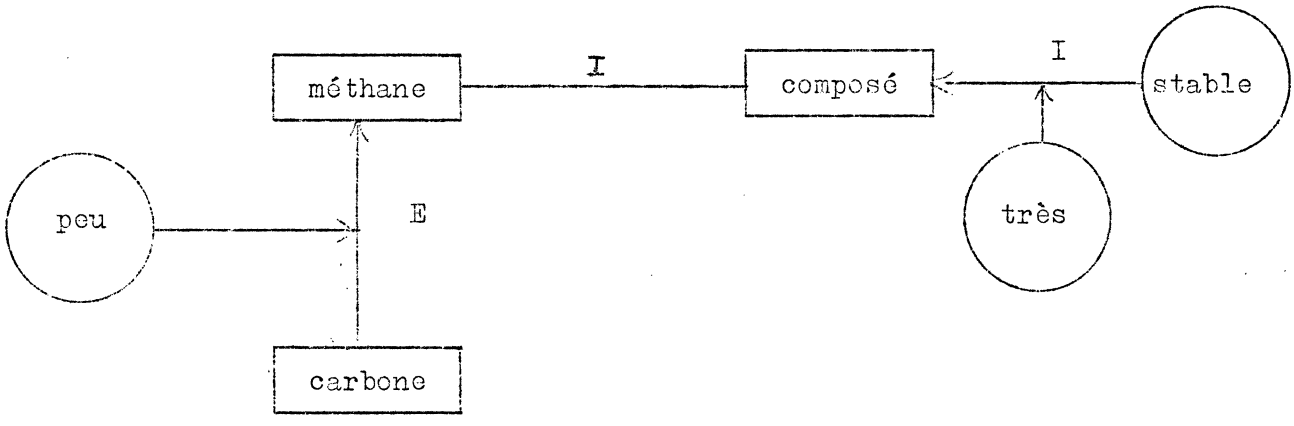
> La pyrogénation de la houille fournit un gaz qui contient du méthane qui peut également être produit par la fermentation de la cellulose.

et non pas la phrase :

> La pyrogénation de la houille fournit un gaz qui contient du méthane qui est produit par la fermentation de la cellulose.

Il y a donc lieu d'ajouter certains symboles pour indiquer comment le diagramme doit être lu. Bien souvent les seuls numéros des documents ou des parties de documents correspondant au diagramme en question permettront de lever l'ambiguité.
Un autre cas de superposition se présente lorsqu'un "objet" dans une phrase est désigné dans d'autres phrases par un pronom.

> Exemple : Le méthane est un composé très stable. Il contient peu de carbone....



> Enfin il peut se faire que la liaison soit réalisée par le fait de positions hiérarchiques ordinaires.

TRAVAIL PRATIQUE

Réaliser la superposition, s'il y a lieu, des diagrammes obtenus au cours de l'exercice précédent.

ANALYSE AUTOMATIQUE

(Programmes de Conflits)

Y. LECERF

INTRODUCTION

Entre le présent texte, et l'exposé fait au séminaire de février 1960, il y a toute la distance qui doit normalement séparer, de la présentation orale d'une expérience sur ordinateur, le compte-rendu écrit de celle-ci. Le compte-rendu écrit se doit d'être beaucoup plus complet, et cela sur au moins deux points :

A.- Premier point : rappel d'un contexte de problèmes scientifiques, contexte qui annonce et rend nécessaire le montage de dispositifs expérimentaux.

L'expérience de février 1960 a permis d'obtenir des renseignements sur l'efficacité et la rentabilité de certains procédés de calcul d'adresses. Elle n'est que la première dans une série d'essais sur ordinateur, pour l'étude et la mise au point de méthodes générales de consultation de catalogues de règles.

Nous pensons en effet que les problèmes de l'analyse automatique doivent être attaqués simultanément, d'une part sur le front de la grammaire et de la sémantique, mais d'autre part aussi, sur le front des techniques de calcul d'adresses, de raisonnement automatique concernant des adresses, de consultation et d'exploitation automatique des catalogues en général.

Aussi, nous commencerons par montrer qu'il ne sert à rien d'introduire dans la mémoire d'un ordinateur de vastes catalogues de règles grammaticales et sémantiques, si ces catalogues ne sont pas consultables, exploitables automatiquement; que ce caractère d'exploitabilité ne dépend pas seulement du catalogue, mais aussi de l'état d'avancement de la technique de consultation automatique des catalogues en général, qui est une technique de calcul d'adresses; qu'un catalogue non exploitable dans un certain état de cette technique peut le devenir si celle-ci fait des progrès.

Nous rappellerons qu'il est nécessaire, pour l'analyse automatique, de consulter non pas des catalogues grammaticaux et sémantiques pauvres, mais bien des catalogues riches, très riches en information.

Si cette condition n'est pas remplie, les opérations se trouveront bloquées dès le début et au niveau du texte d'entrée, dont il sera impossible de lever les ambiguïtés. Se contenter d'une grammaire sémantique pauvre, c'est se condamner aux contresens les plus grossiers

La grammaire et la sémantique, tout court, sans considération d'exploitabilité automatique (et cette restriction est importante), ne sont plus des sciences à inventer. Elles existent déjà.

Tout le monde sait que l'article s'accorde avec le nom, l'épithète avec le substantif, et ainsi de suite. Tout le monde sait que seule une chose susceptible de mouvement peut bouger; que seule une personne, ou une créature assimilée à une personne, peut rire ou pleurer; que si une personne est dite travailler dans une pièce, il s'agit d'une pièce d'habitation et non d'une pièce de monnaie , en vertu de cette règle sémantique qu'un volume ne peut pas être inclus dans un volume plus petit.

Cette dernière règle touche déjà à la biologie, la physique, (non compressibilité d'un être humain en train de travailler) et la géométrie (inclusion de volumes). Aussi il apparaît que les meilleurs recueils de lois permanentes valables sur le plan des choses signifiées, que ces recueils très riches ne coûtent rien: la société en a déjà payé le prix, par des siècles d'observations scientifiques. Les meilleurs traités de sémantique, sans considération d'exploitabilité automatique, (et cette restriction est importante,) ce sont les traités usuels de mécanique, de physique, de logique, les encyclopédies de chimie, d'histoire naturelle, de géographie, et ainsi de suite, sans parler des traités existants de sémantique, qui donnent des règles de bon sens courant.

Mais ces traités immensément riches, beaucoup plus riches encore que nous n'avions souhaité, ne sont pas, dans l'état actuel des choses et à notre connaissance, exploitables automatiquement pour lever les ambiguïtés du langage ordinaire.

De même, l'énorme masse d'informations contenues dans les traités de grammaire, ouvrages dont certains comportent des milliers de pages, ne sont pas, sous leur forme brute actuelle, exploitables pour lever les ambiguïtés du langage ordinaire.

Si nous en venons maintenant aux catalogues réputés consultables, exploitables par des ordinateurs, force est de constater que l'on n'en a jusqu'ici publié que fort peu dans le monde; que ceux qui ont été publiés ne contiennent pas beaucoup de règles; que même ces quelques règles sont parfois en contradiction avec les faits, faute d'être accompagnées de listes suffisantes d'exceptions.

Le rassemblement de catalogues de règles grammaticales et sémantiques exploitables en machine, c'est-à-dire consultables automatiquement, exige un travail minutieux. La progression est lente. Chaque règle de ces catalogues revient finalement assez cher, surtout si l'on réfléchit au nombre de règles qu'il faudra rassembler.

La construction de ces catalogues se révèle si lente et si difficile, que plusieurs équipes ont remis à plus tard le rassemblement de règles sémantiques, se spécialisant provisoirement dans la

grammaire. Encore y a-t-il des discussions sur le point de savoir si certaines manières de noter les adresses de mots liés par les règles ne conduisent pas à des catalogues quasi infinis. On imagine facilement les conséquences qui en résulteraient en ce qui concerne les coûts, d'une part de la construction d'un tel catalogue, d'autre part de la mémoire machine destinée à le contenir.

Au total, les particularités imposées par les conditions d'exploitabilité automatique, et tout particulièrement cette contrainte de devoir donner, directement ou indirectement, les adresses des mots liés par chaque règle, coûtent fort cher, et d'autant plus cher que les mécanismes de consultation automatique sont moins perfectionnés.

D'où l'intérêt de réduire le prix des catalogues en améliorant les techniques de consultation automatique et de calcul d'adresses.

B.- Second point : explication du choix de tel mécanisme plutôt que de tel autre.

Nous nous efforcerons de poser d'une manière générale les problèmes de consultation de catalogues, puis de montrer progressivement comment des considérations de coût conduisent à choisir tel et tel procédé. L'exposé y gagnera en généralité.

Dans une annexe, nous rassemblerons les résultats acquis au cours des chapitres, et définirons les conditions de l'expérience sur ordinateur.

On trouvera plus loin, sous la signature d'Eric Morlet, l'organigramme du programme.

Cette expérience n'aurait pas pu être réalisée dans les délais prescrits, sans la bienveillance de M. P. Gillis, professeur à l'Université libre de Bruxelles, qui a bien voulu mettre à notre disposition pour les essais l'ordinateur IBM 650 de son laboratoire de calcul électronique et nous permettre de faire appel au programmeur de ce laboratoire, Eric Morlet, qui a assuré avec la plus grande célérité la rédaction et la mise au point du programme.

PREMIERE PARTIE

CALCUL D'ADRESSES ET CONSULTATION DE CATALOGUES DE REGLES EN ANALYSE AUTOMATIQUE

I, 1 PERSPECTIVES DE L'ANALYSE DOCUMENTAIRE AUTOMATIQUE.

Par analyse documentaire, on désigne habituellement les opérations qui permettent de passer d'un document rédigé en langage ordinaire, et riche en informations, à un texte en langage documentaire, texte qui est généralement bien moins long et bien plus pauvre en informations que le document initial.

Dans la mesure où elle implique l'élimination des renseignements jugés les moins importants, l'analyse documentaire s'apparente au résumé. Dans la mesure où elle implique un changement de forme, un changement de langage, elle s'apparente à la traduction.

L'importance des opérations du type "résumé", comparée à celle des opérations du type "traduction", dépend beaucoup des circonstances et du système documentaire utilisé. Or il peut se révéler que les unes soient plus difficiles à automatiser que les autres, d'où la nécessité de faire la part de chacune d'entre elles. On peut imaginer certains cas limites où le résumé est tout dans l'analyse, et d'autres cas où il n'est rien.

Donnons un exemple. Les centres de documentation qui existent actuellement dans le monde publient de nombreux recueils d'abstracts. On peut considérer que ces abstracts constituent en eux-mêmes des documents. On peut se proposer de transformer ces abstracts en diagrammes de Braffort et Leroy, sans les condenser plus, puis de stocker en vue de la sélection, les diagrammes ainsi obtenus. Dans ce cas très particulier, l'analyse documentaire serait essentiellement une traduction de langage ordinaire en langage documentaire.

Même lorsque la partie résumé joue un rôle important, on peut trouver avantage à effectuer successivement les opérations de traduction et de résumé, ne serait-ce que pour cette raison que le langage ordinaire est une notation très ambiguë, si bien qu'avant de résumer un texte il y a intérêt à en expliciter d'abord, de façon univoque, la signification.

Si l'on tient compte des efforts actuellement accomplis dans le monde en vue de promouvoir la traduction automatique, on peut espérer que le passage de langues ordinaires vers des langages documentaires, sans résumé, pourra se faire automatiquement et avec une rentabilité convenable dans des délais non trop éloignés. Un premier pas, et non le moindre, serait alors accompli sur la voie de l'analyse documentaire automatique.

Cet exposé ne traitera que de la traduction automatique vers des langages documentaires, à l'exclusion de toute opération de résumé.

Encore nous limiterons-nous aux problèmes relatifs à la levée des ambiguïtés des textes à analyser.

I, 2 AMBIGUITES DU LANGAGE ORDINAIRE : HOMONYMIES, FAUSSES RELATIONS ENTRE LES TERMES DE L'ENONCE.

A/ Homonymies.-

L'existence d'homonymies, et de ces homonymies du langage écrit que l'on appelle homographies, est bien connue de tout le monde. Mais on ne se représente pas toujours la complication qui en résulte pour l'analyse automatique.

Prenons l'exemple d'une petite phrase: "Pose la table !" et supposons que nous disposons seulement d'un dictionnaire automatique :

pose	la	table
verbe	pronom	verbe
nom	article	nom
	nom	

Le tableau ci-dessus ne donne qu'une idée très grossière de la multiplicité des acceptions associables aux trois formes "pose" "la" et "table". Pour trouver l'interprétation correcte de la phrase, il faut faire un choix entre les possibilités ainsi suggérées :

V, Pr, V	V, Art, N	N, Pr, V	N, Art, N
V, Pr, N	V, N, V	N, Pr, N	N, N, V
V, Art, V	V, N, N	N, Art, V	N, N, N

qui sont au nombre de : 2 x 3 x 2 = 12

Prenons maintenant l'exemple d'une phrase un peu plus longue :

Le	manoeuvre	pose	le	tube	sur	la	table
Art	V	V	Art	V	Adj	Art	V
Pr	N	N	Pr	N	Prép	N	N
						Pr	

A ce tableau il correspond :

2 x 2 x 2 x 2 x 2 x 2 x 3 x 2 = 384 possibilités.

Autre exemple: la phrase "Dans cette dure lutte contre la montre, l'illustre Coppi ferme la marche" contient sept mots ayant forme de verbe, alors qu'un seul y remplit cette fonction.

Les méthodes selon lesquelles "on doit commencer par déterminer quels mots sont des verbes" rencontreront peut-être, avec une telle phrase, quelques difficultés d'application.

Dans	cette	dure	lutte	contre	la	montre
Prép.	Dém.	Adj	N	Prép.	Art	N
		N	V	N	N	V
		V		V	Pr	

l'	illustre	Coppi	ferme	la	marche
Art	Adj	N	Adj	Art	N
Pr	V		Adv	Pr	V
			N	N	
			V		

Total des possibilités :

1 x 1 x 3 x 2 x 3 x 3 x 2 x 2 x 2 x 1 x 4 x 3 x 2

= 10.368

L'augmentation rapide, généralement exponentielle en fonction du nombre de mots, du nombre de possibilités associables à une

phrase, ne doit pas effrayer outre mesure. D'une part en effet les ordinateurs calculent vite. D'autre part, le temps nécessaire pour effectuer un choix parmi n possibilités n'est pas forcément proportionnel à n. Si les meilleures conditions sont réalisées, ce temps peut même varier comme Log n, c'est-à-dire rester à peu près proportionnel à la longueur des phrases.

Cependant, si ces conditions de choix idéales ne sont pas réalisées, le temps de résolution des homographies pourra augmenter beaucoup plus vite que la longueur des phrases. La difficulté du traitement des homonymies ne doit pas être sous-estimée. Les tableaux dessinés plus haut n'indiquaient d'ailleurs qu'une faible partie des possibilités à envisager pour chaque mot.

Si dans ces tableaux nous avions fait mention de la possibilité, qui existe toujours, d'employer substantivement n'importe quel mot lorsqu'on veut le désigner en tant que mot, le nombre de combinaisons grammaticales serait apparu comme bien plus grand.

D'autre part, lorsque le mot "table", par exemple, ou bien le mot "pose", ou "tube", ou "manoeuvre", ou "contre", ou "montre", ou "marche", ou "dure", ou "lutte", etc. ont, dans une phrase, fonction de verbe, il reste encore à préciser s'il s'agit de l'impératif présent (2ème personne du singulier), de l'indicatif (1ère ou 3ème personne) ou du subjonctif (1ère ou 3ème personne). Ce n'est donc pas une hypothèse grammaticale "verbe", mais bien cinq, qu'il faudrait inscrire en dessous de ces mots dans les tableaux.

Enfin, les alternatives grammaticales proposées par ces tableaux ne peuvent constituer qu'un tri préalable, un dégrossissage. Chacune d'elles recouvre en général plusieurs significations, et le but de l'analyse n'est-il pas de départager les significations ?

Employé substantivement, le mot "pose" peut désigner soit l'action de poser, soit une attitude du corps, soit un temps d'exposition en photographie, etc. Employé substantivement, le mot "table" peut représenter soit un meuble, soit un tableau de renseignements disposés méthodiquement (table de logarithmes, etc), soit, dans certains cas, une assemblée de personnes ("Toute la table éclate de rire" - "La table ronde s'est réunie..."), etc.

Pour obtenir des notations non ambiguës, il faudrait par exemple numéroter ces significations, le plus difficile étant naturellement de les départager automatiquement.

De même, qu'il s'agisse de documentation ou de traduction, il importe de séparer, parmi les substantifs qui s'écrivent "montre" : "montre 1" qui indique l'heure; "montre 2" qui est une vitrine; "montre 3" qui est l'action de faire étalage d'un talent, d'un sentiment. Parmi les substantifs féminins "marche", on distinguera "marche 1", partie d'un escalier, "marche 2", action de mettre un pied devant l'autre, "marche 3", morceau de musique, etc.

Certaines phrases restent ambiguës même pour un analyste humain. Exemples: "La fille de cet homme illustre le livre". "La première marche et la seconde valse".

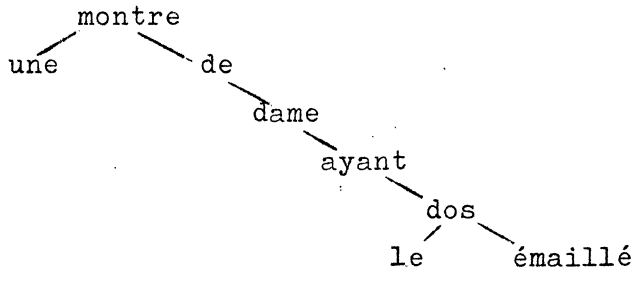
B/ Fausses relations entre les termes de l'énoncé.

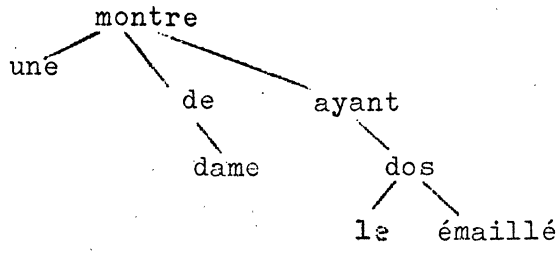
Une seconde source d'ambiguïté du langage ordinaire se manifeste lors de l'interprétation des rapports entre mots ou entre groupes de mots. Charles Bally (1), à qui nous avons emprunté le sous-titre de ce paragraphe, donne des exemples de phrases-pièges, typiques de ce genre d'ambiguïté: "Les fils de fonctionnaires morts à la guerre" (on ne sait pas qui, des fonctionnaires ou de leurs fils, est mort à la guerre). Autre phrase: "J'ai vu la fille du fermier qui nous vend des légumes" (on ne sait pas qui, du fermier ou de sa fille, est le vendeur).

Il n'est pas question de demander à une machine de faire un choix dans les exemples ci-dessus, puisque l'esprit humain lui-même hésite. Mais pour certaines autres phrases du même type, l'analyste humain lève facilement l'ambiguïté, grâce parfois à des indices grammaticaux ("Les filles de fonctionnaires morts à la guerre") ou sémantiques ("une montre de dame dont le dos est émaillé"). Il sait aussi reconnaître ce rapport particulier entre certains mots qui fait d'eux un tout indissociable, une locution (exemple: "prendre ses jambes à son cou"). Il sait, d'une manière générale, rétablir les liens entre les mots.

Les relations entre mots constituent un composant essentiel de la signification de toute phrase. Le fait que ces relations ne soient, dans le langage naturel, pas notées explicitement par écrit, introduit des ambiguïtés en nombre considérable; car, du point de vue d'une machine qui doit utiliser certaines informations, il n'y a rien de plus ambigu que l'absence totale de ces informations dans les données. Il ne s'agit plus seulement ici de quelques phrases-pièges telles que celles citées plus haut, mais bien de toutes les phrases, puisque les relations entre mots n'apparaîtront jamais autrement que comme le résultat d'un calcul.

Une fois acquis ce résultat, il faudra le noter, ne serait-ce qu'à l'intérieur de la machine. Les diagrammes de Braffort et Leroy (2) s'y prêtent fort bien. On peut aussi se servir, comme le professeur Ceccato (3), de corrélogrammes, ou, comme Chomsky (4) de diagrammes de dérivation, ou, comme Tesnière (5) de stemmas. Toutes ces notations ont une inspiration graphique et seront décrites plus loin. Donnons simplement ici celle que nous utilisons, et qui est très proche de celle de Harper et Hays (6)





En conclusion, il n'y a pas d'analyse documentaire concevable (ni d'ailleurs de traduction) sans que l'on commence par lever les ambiguïtés du texte d'entrée.

I, 3 RESOLUTION DES AMBIGUITES PAR CONSULTATION DE CATALOGUES DE REGLES NORMATIVES.

L'analyste humain, pour résoudre les ambiguïtés relatives tant aux mots qu'aux rapports qui les lient, fait appel aux règles normatives que les usagers d'une langue naturelle sont censés respecter : règles grammaticales, règles sémantiques.

Voyons si une machine peut utiliser de telles règles.

Les premières imposent des contraintes relativement strictes. Donnons des exemples d'ambiguïté départagés par ce moyen; dans chacune des phrases suivantes :

"Ces pages sont beaux"

"Ces pages sont belles",

le mot "pages" prend une signification différente (1. jeune noble. 2. feuille de papier). C'est la règle grammaticale d'accord en genre de l'attribut avec le sujet et qui permet de choisir chaque fois.

Dans les phrases suivantes :

"Les filles de fonctionnaires morts pendant la guerre"

"Les filles de fonctionnaires mortes pendant la guerre"

c'est la règle d'accord du genre de l'épithète qui permet chaque fois de lever l'ambiguïté, et de savoir qui, des fonctionnaires ou de leurs filles, est mort pendant la guerre.

Les règles grammaticales sont quelquefois enfreintes par les auteurs. Bien que ces cas soient assez rares, on peut les prévoir, en associant à chaque règle une "pénalité", d'autant plus grande que la règle est plus stricte. Si aucune solution correcte n'est trouvée, on choisira la solution incorrecte qui contredit le moins de règles, c'est à-dire la moins pénalisée.

Les règles sémantiques interviennent également dans la résolution des ambiguïtés du langage ordinaire. Si la phrase "une montre de dame ayant le dos émaillé" s'interprète sans difficulté, c'est parce

que nous savons que le dos des dames n'est généralement pas émaillé.

Les règles sémantiques n'ont rien d'absolu. On peut construire des phrases qui les violent systématiquement. Aussi, pour ces règles, le système des pénalités apparaît comme encore plus nécessaire que dans le cas des règles grammaticales.

Une machine ne peut pas utiliser des informations si on ne les lui a pas, préalablement, données. Si l'on veut lever automatiquement les ambiguïtés du langage naturel, il faut introduire en mémoire machine des renseignements sur les normes de ce langage, par exemple sous la forme d'un catalogue de règles. Si ce catalogue est incomplet, la qualité du programme s'en trouvera d'ores et déja partiellement compromise, puisque la machine n'aura aucun moyen de rétablir les renseignements manquants; les phrases pour l'interprétation desquelles ces renseignements auraient été nécessaires apparaîtront comme ambiguës.

Ainsi, rien qu'en observant le "catalogue de règles", ou ce qui en tient lieu, on peut déjà fixer une borne supérieure à l'efficacité d'un programme.

Ceci dit, il ne suffit pas d'introduire un catalogue de ce genre en mémoire machine. Il faut encore en organiser la "consultation" automatique. Cela implique que le catalogue soit "consultable", et c'est là que s'introduit la principale difficulté. Dans un ordinateur en effet, il ne peut y avoir d'information sans adresse, et inversement, pour consulter une information, il faut connaître son adresse ou bien pouvoir la retrouver au moyen d'un calcul.

Une phrase étant donnée, on devra, pour en lever les ambiguïtés résoudre les deux problèmes suivants :

1° Trouver dans le catalogue ou dans ce qui en tient lieu, les règles grammaticales et sémantiques utiles pour le cas de cette phrase; pour cela il faudra calculer les adresses de ces règles.

2° Ces règles une fois trouvées, il faudra les exploiter. Les règles grammaticales, comme les règles sémantiques, imposent certains rapports entre les catégories auxquelles appartiennent les éléments du langage. (Par exemple : si tel mot est féminin, tel autre mot l'est aussi. Si tel et tel objet sont liés par une relation d'inclusion, le contenant est plus grand que le contenu, etc.) Pour vérifier que les prescriptions en question sont satisfaites, il faut trouver les mots qu'elles concernent; on devra donc calculer les adresses de ces mots dans la phrase donnée.

I, 4 GRAMMAIRE, SEMANTIQUE, ET CALCUL D'ADRESSES.

Ainsi, pour lever les ambiguïtés de phrases du langage ordinaire, il faut satisfaire simultanément ces deux conditions qui sont :

1° L'introduction d'informations en quantité suffisante dans la machine, sous forme par exemple d'un catalogue de règles normatives.

2° La mise au point de mécanismes susceptibles d'assurer la consultation automatique de ce catalogue, avec tout ce que cela implique de difficultés de calcul d'adresses.

En vérité, il n'est pas possible de dire que l'une de ces fonctions est plus difficile à assurer que l'autre. L'histoire des recherches en traduction automatique montre en effet que si l'on chasse la difficulté de l'une, on la retrouve dans l'autre.

Ou bien en effet, on impose à l'avance un mécanisme de consultation simple, et alors, le catalogue devient très compliqué, très volumineux, très différent des corpus de règles que l'on trouve dans les traités ordinaires de grammaire ou de sémantique, et la construction de ce catalogue devient alors coûteuse, et coûteuse également la mémoire machine qui doit le contenir.

Ou bien au contraire, on impose à l'avance l'idée d'un catalogue bon marché; il est alors relativement facile de rassembler des règles normatives de forme quelconque, ne serait-ce qu'en tirant parti des nombreux traités déjà existants. Il est possible de se faire une idée de la quantité d'information ainsi réunie, en effectuant des sondages, et de voir si elle est suffisante pour un usage déterminé, par exemple pour la levée des ambiguïtés d'un certain type de textes avec un certain pourcentage d'erreurs. Mais alors, c'est dans le programme de consultation que la difficulté fait sa réapparition: les catalogues bon marché sont en effet rédigés, comme les grammaires ordinaires, comme les traités de sémantique ordinaires, sans indication d'adresses.

La règle d'accord grammatical du sujet et du verbe, par exemple, y sera énoncée sans que l'on précise quelle est l'adresse du sujet relativement au verbe, combien de mots peuvent les séparer dans l'ordre linéaire d'énonciation des phrases, quelles sont ces configurations de mots qu'il est ainsi permis d'intercaler entre sujet et verbe. Ces renseignements, que les catalogues bon marché ne contiennent pas explicitement, devront finalement être quand même rétablis, à partir des informations de ces catalogues, par un calcul d'adresses, de sorte que le programme de consultation deviendra compliqué, coûteux en temps machine, et difficile à construire.

Il est clair qu'il s'agit toujours d'une même difficulté aux multiples visages : si on la chasse du catalogue, elle reparaît dans le programme de consultation; si on la chasse de ce dernier, elle reparaît dans le catalogue, et l'on peut imaginer autant que l'on en veut,

des états intermédiaires, où la difficulté se partage entre les deux fonctions complémentaires. Ceux qui ont choisi une politique de catalogue cher à consultation simple diront que la traduction automatique pose essentiellement des problèmes de catalogue, des problèmes linguistiques. Ceux qui ont choisi une politique de catalogue bon marché à consultation difficile, diront que la traduction automatique pose surtout des problèmes de calcul d'adresses.

I, 5 FRACTIONNEMENT DES PROBLEMES EN SOUS-PROBLEMES - NOTION DE FILTRE

Le travail de traitement des ambiguïtés de phrases du langage ordinaire peut être réparti entre plusieurs automates (ou entre plusieurs sous-programmes pour une même machine) assurant des fonctions complémentaires, de façon à fractionner le problème d'ensemble en plusieurs sous-problèmes plus simples. On peut faire cette répartition comme suit :

1° un "dictionnaire automatique" examine isolément chaque mot du texte d'entrée et suggère, pour chacun d'eux, diverses interprétations.

exemple simplifié :

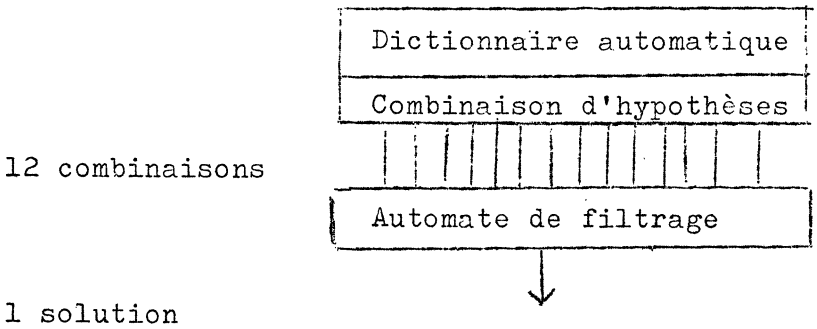
Pose	la	table
V	Pr	V
N	Art	N
	N	

2° un "automate à combiner les hypothèses" assemble de toutes les manières possibles les interprétations proposées pour chaque mot.

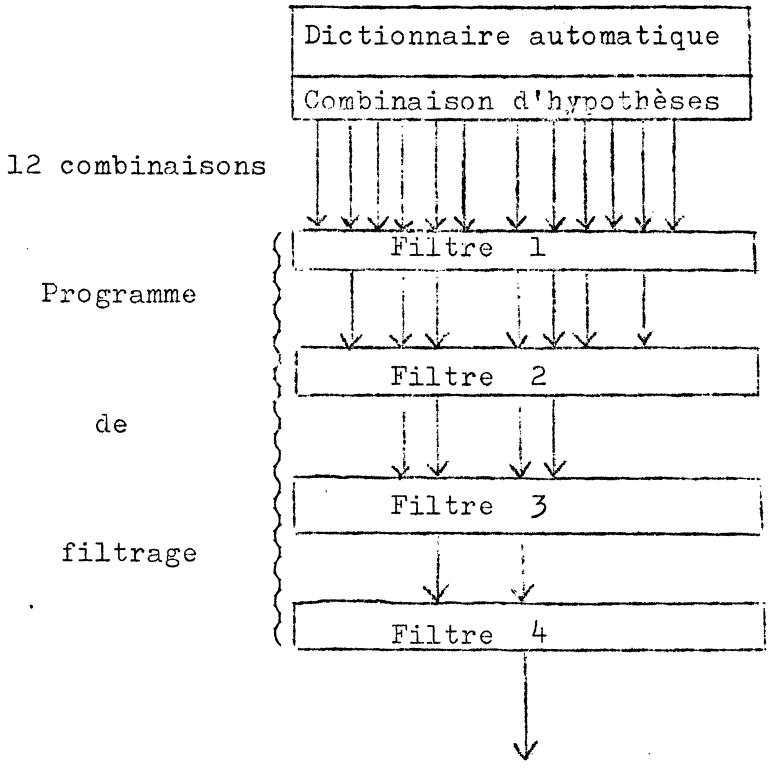
Pour la phrase ci-dessus, il y aurait 2 x 3 x 2 = 12 combinaisons proposées : V Pr V, V Pr N, V Art V, V Art N, etc.

3° un "automate de filtrage" reçoit ces combinaisons, examine chacune d'elles en fonction de catalogues de règles grammaticales et sémantiques. Si une combinaison viole l'une au moins de ces règles, elle est rejetée, sinon, elle est acceptée.

Cette répartition des fonctions se schématise comme suit :



Mais il est clair que la troisième fonction, celle du filtrage, peut elle-même se répartir entre plusieurs "filtres" partiels, disposés par exemple en série :



Chacun de ces filtres fonctionnera par consultation de règles normatives, ou par un processus équivalent, mais la méthode de consultation pourra varier d'un filtre à l'autre, pour que l'on obtienne une meilleure adaptation des moyens. Si en effet tel ou tel procédé de consultation se révèle favorable dans le cas de certaines règles, cela n'implique pas qu'il soit optimal pour l'ensemble du catalogue. Le contraire est même assez probable. Ainsi le découpage en filtres permet une grande souplesse dans le choix des méthodes, et l'utilisation simultanée de plusieurs procédés, chacun intervenant dans le traitement des problèmes pour lesquels il est le mieux adapté.

La terminologie des "filtres" doit surtout être considérée comme un langage documentaire commode pour la description de certains procédés de consultation de règles. En pratique, en effet, il conviendra de mélanger autant que possible les filtres avec l'automate à combiner les hypothèses.

Montrons la chose sur un exemple. Soit un filtre très élémentaire, qui ferait intervenir uniquement un catalogue de séquences binaires interdites, telles les séquences Article - Verbe personnel, ou bien Préposition - Verbe personnel, séquences que l'on ne rencontre jamais en français.

Considérons la phrase :

Le	guide	lance	une	boule	de	neige
Art Pr	V N	N V	Art	N V	Prép	N V

Sur les 2 x 2 x 2 x 1 x 2 x 1 x 2 = 32 combinaisons d'hypothèses que l'on peut former à partir du tableau ci-dessus, 26 contiennent l'une au moins des séquences interdites Article - Verbe personnel, ou Préposition - Verbe personnel. Pour cette phrase, le filtre binaire grossier a donc un rendement d'environ 75%, ce qui n'est pas si mal.

Comparons maintenant les deux montages :

a/ filtre binaire distinct de l'automate à combiner les hypothèses;

b/ filtre binaire incorporé dans l'automate à combiner les hypothèses.

Dans le premier montage, les 26 solutions contenant des séquences interdites sont explicitement formulées, puis éliminées.

Dans le second montage, ces 26 solutions ne sont même pas formulées.

On peut imaginer par exemple que la combinaison d'hypothèses se fait de gauche à droite. Chaque fois qu'un mot est ajouté, on vérifie qu'il ne forme pas, avec son prédécesseur, une séquence interdite. Exemple :

Art
Art V (alarme)
Art N
Art N N
Art N N Art
Art N N Art V (alarme)
Art N N Art N etc.

La première alarme élimine 8 solutions virtuelles, la seconde en élimine 2. En outre, on récupère la séquence assemblée avant l'alarme, pour former d'autres combinaisons.

Il s'agit d'un phénomène absolument général : en mélangeant les filtres avec l'automate à combiner les hypothèses, on arrive à supprimer une grande partie de ces opérations inutiles, qui consistent à proposer des solutions fausses pour devoir ensuite les barrer.

Le principe reste le même, mais on élimine les solutions fausses par paquets, et à l'état de solutions virtuelles, sans avoir dû les formuler explicitement.

Comme le nombre de solutions fausses à éliminer est en général extrêmement grand, le gain de temps, qui résulte de leur traitement et de leur préélimination à l'état virtuel, est considérable. Aussi les sous-programmes du type "filtre" n'existeront-ils pratiquement jamais à l'état pur, sauf pour des tests préliminaires destinés à donner une idée du pouvoir d'élimination de telle ou telle méthode.

Dans la pratique en effet, un filtre sera presque toujours mélangé avec l'automate à combiner les hypothèses, et, par voie de conséquence, mélangé à l'ensemble des autres filtres.

Cependant, il n'est pas commode de décrire et d'étudier des programmes où toutes les fonctions sont mélangées. De même qu'en biologie, lorsque deux réseaux, tels ceux du système nerveux et du système circulatoire, sont étroitement imbriqués, on a coutume de les décrire d'abord séparément, de même, en analyse automatique, on gagne souvent beaucoup à raisonner en termes de filtres. C'est ainsi par exemple que le pouvoir d'élimination, c'est-à-dire la faculté de rejeter les hypothèses fausses suggérées par un dictionnaire, ne change pas lorsque l'on passe d'un schéma de filtres mélangés au schéma constitué par les mêmes filtres supposés isolés et disposés en série. Pour l'étude des pouvoirs d'élimination, le second schéma est plus commode que le premier, d'où l'intérêt de la notion de filtre.

SECONDE PARTIE

UTILISATION DE GRAPHES DE REPERAGE, EN VUE DE REDUIRE LE PRIX DE REVIENT DES CATALOGUES

II, 1 GRAPHES LINGUISTIQUES ET GRAPHES DE REPERAGE.

Il ne peut y avoir, dans la mémoire d'un ordinateur, d'information sans adresse. Une fois emmagasinée, une information ne peut être utilisée que si l'on connaît cette adresse, ou si l'on peut la retrouver indirectement au moyen d'un calcul. Lorsque les informations à exploiter sont nombreuses, on évite autant que possible de dresser une liste explicite de leurs adresses. Il est plus élégant de recourir au second procédé, celui du calcul d'adresses, si c'est possible.

Ainsi, le traitement en machine d'êtres linguistiques fait intervenir en réalité :

1° Des êtres linguistiques proprement dits (mots, règles de grammaire etc.);

2° Leurs adresses;

3° Les repères (explicites ou implicites) en fonction desquels on calcule ces adresses.

Il faut se garder de confondre, dans quelque mesure que ce soit, ces trois sortes d'êtres, qui relèvent de statuts totalement différents.

Les êtres linguistiques sont à considérer comme des données de fait. Il n'est au pouvoir de personne d'empêcher la langue allemande ou la langue russe de comporter des déclinaisons: ce sont là des vérités expérimentales.

Au contraire, les êtres de seconde et de troisième catégorie, à savoir les adresses et les repères sont de purs auxiliaires de calcul. De leur existence fugitive, qui se déroule toute entière à l'intérieur de la machine, l'usager de la traduction automatique ou de la documentation automatique n'a pas directement connaissance: cet usager voit le texte d'entrée, le texte de sortie, et peu lui importent les opérations intermédiaires.

Aussi, les êtres de seconde et troisième catégorie, adresses et repères de calcul, sont-ils soumis dans une très large mesure à l'arbitraire de l'ingénieur, qui les choisira de façon à minimiser le coût des opérations. Ces êtres ne se jugent pas en termes de vrai ou de faux, comme les êtres linguistiques proprement dits, mais en termes de cher ou de bon marché.

Pour en revenir aux graphes qui jouent un rôle en linguistique et en traduction automatique, il importe de faire très nettement la distinction entre :

1° Les graphes linguistiques proprement dits, d'une part, qui sont des êtres linguistiques. Leurs traits représentent des liens grammaticaux ou sémantiques. Les extrémités de ces traits indiquent quels mots, ou quels groupes de mots, sont liés par le rapport grammatical ou sémantique en question.

2° Les graphes de repérage, d'autre part, qui sont des êtres de troisième catégorie, des repères de calcul. Leurs traits représentent des liens d'adressage et rien de plus. Lorsque deux mots, par exemple, sont liés par un tel trait, cela signifie que chacun d'eux est repéré directement par rapport à l'autre, ou si l'on veut que l'adresse de l'un est calculable directement à partir de l'adresse de l'autre lorsque l'on utilise le graphe comme repère. On notera que la relation: "avoir une adresse par rapport à ..." est transitive. Deux mots repérés par rapport à un même troisième sont repérés entre eux, mais pour calculer l'adresse de l'un à partir de celle de l'autre, il faut calculer d'abord celle du mot intermédiaire.

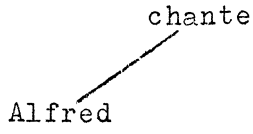
Ainsi, graphes linguistiques et graphes de repérage sont des êtres complètement distincts. Malheureusement, il arrive très souvent qu'un graphe d'une sorte et un de l'autre aient le même tracé, et des apparences identiques. Le risque de confusion est alors grand. On sait que la géométrie analytique connaît des situations analogues, lorsque par exemple un couple d'axes de coordonnées coïncide avec les asymptotes d'une hyperbole équilatère. La même apparence (couple de droites perpendiculaires) recouvre alors à la fois un être géométrique (couple d'asymptotes) et un groupe de repérage (couple d'axes de coordonnées), mais on se garde bien de les confondre pour autant.

A/ GRAPHES LINGUISTIQUES

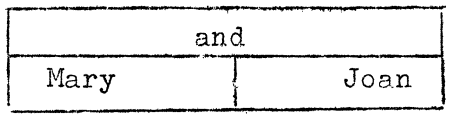
La forme la plus couramment utilisée pour la description des rapports grammaticaux ou sémantiques, est celle du langage ordinaire. Elle se prête en effet à toutes les nuances. Elle permet d'énumérer des exceptions, de signaler que les auteurs sont en désaccord sur tel ou tel point. Les faits linguistiques constituent une réalité si riche et si complexe, que l'on n'en peut jamais rendre compte de façon simple et schématique, à moins dêtre incomplet.

Certains linguistes ont cependant jugé que certains de ces rapports jouaient un rôle particulièrement important, et les ont représentés graphiquement. C'est leur droit.

Tesnière (5), par exemple, utilise un trait ascendant pour représenter le lien sujet-verbe :



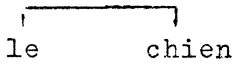
Le professeur Ceccato (3) représente comme suit les rapports dits de correlation :



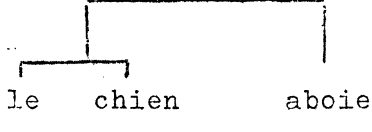
Sans dessiner de graphes, certains linguistes insistent tellement sur l'importance spéciale de certains rapports, que le lecteur en vient naturellement à une représentation graphique.

C'est le cas par exemple des rapports article nom, sujet prédicat, etc. en grammaire traditionnelle

article - nom :



sujet - prédicat :



etc.

De même, la procédure de découpage, et les trois fonctions de la glossématique (7) (16) suggèrent une représentation graphique, dont nous nous inspirerons par la suite :

Pourquoi ces linguistes considèrent-ils certains rapports comme spécialement importants ? Ceccato nous explique que les

correlations représentent les opérations de la pensée. Tesnière donne, pour ses stemmas, une justification analogue (5).

Les rapports grammaticaux ou sémantiques sont des faits expérimentaux. Les opérations de la pensée sont des faits expérimentaux (difficiles à atteindre il est vrai).

Les graphes linguistiques visent à représenter des faits expérimentaux. On peut les juger en termes de vrai ou de faux. On peut discuter des importances de ces faits expérimentaux.

On ne peut pas les déclarer arbitraires.

B/ GRAPHES DE REPERAGE

De même qu'il n'est jamais interdit de tracer, ou d'imaginer, un couple d'axes de coordonnées, et de décider de s'en servir pour calculer des adresses, de même il n'est jamais interdit de tracer ou d'imaginer des graphes, puis de s'en servir pour repérer conventionnellement des adresses d'êtres linguistiques.

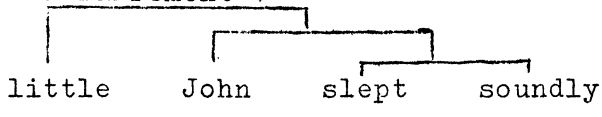
Ce sont là des décisions que l'on peut prendre arbitrairement, elles n'engagent que la personne qui les a prises.

En tant que graphe linguistique, le dessin :

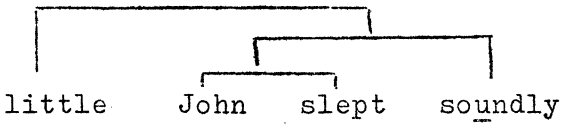
est une réalité rigide, il bénéficie de la garantie de nombreux linguistes, qui voient en lui l'expression de vérités expérimentales.

En tant que graphe de repérage, le même dessin :

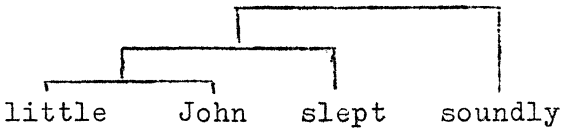
n'exprime rien d'autre qu'un choix arbitraire de l'ingénieur. Il joue le même rôle que les systèmes d'axes de coordonnées en géométrie : tout le monde est libre de placer des axes de coordonnées n'importe où; cela n'engage à rien, cela coûte le temps de les dessiner ou simplement de les imaginer, cela peut faire gagner beaucoup de temps de calcul. En tant que graphe de repérage, le dessin ci-dessus peut donc être modifié arbitrairement :



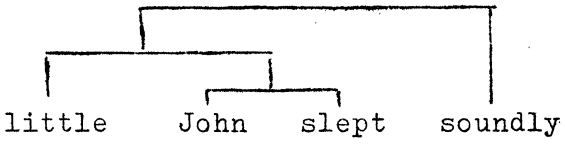
ou bien :



ou bien :



ou bien :



Il n'en coûte aucune entorse à la vérité, qui n'est pas en cause, mais simplement de l'argent au budget de l'ingénieur. Aussi ce dernier s'efforcera-t-il de minimiser le coût des calculs. Si le graphe de repérage le plus économique coïncide avec un graphe linguistique, tant mieux. Sinon, tant pis, en matière de repérage et à performances égales, l'économie passe d'abord.

De même que l'on peut préférer des axes de coordonnées rectangulaires, ou au contraire obliques, que l'on peut les placer n'importe où, les prendre mobiles, tournants, glissants, comme l'on veut, de même, on est libre de définir comme l'on veut la forme et les caractéristiques de graphes de repérage pour la linguistique.

On prendra cependant garde à ceci que :

a/ Pour exister vraiment, et mériter d'être utilisé comme repère, un graphe de repérage devra être complètement défini, de manière que soit fixée sa position relativement aux adresses linéaires d'énonciation des mots de n'importe quelle phrase.

b/ La définition du graphe de repérage est arbitraire avant que l'on fasse un choix à son sujet, mais non après. On est lié par les décisions prises.

c/ De même que la commodité d'axes de coordonnées dépend beaucoup de la façon dont on les a choisis, de même, il y a des tracés plus ou moins avantageux pour les graphes de repérage. Un mauvais choix peut coûter cher en temps machine par exemple.

Le résultat du calcul restant le même, on pourra avoir à manipuler des adresses simples dans un cas, et compliquées dans l'autre; on pourra avoir plus ou moins de difficulté à évaluer des adresses en se repérant à partir du graphe choisi.

L'expérience montre que les graphes de repérage les plus économiques coïncident généralement avec des graphes linguistiques.

d/ Lorsqu'il est possible de traiter deux problèmes indépendamment l'un de l'autre, il y a souvent intérêt à faire intervenir deux types différents de graphes de repérage, de manière à disposer chaque fois de l'instrument le mieux adapté. Dans les programmes à "filtres" multiples, en particulier, chaque filtre peut avoir, au besoin, un système d'adressage qui lui soit propre.

Ces quatre remarques permettent de guider le choix de graphes de repérage.

II, 2 ROLE DES GRAPHES DE REPERAGE.

Les graphes de repérage apportent au calcul d'adresses une facilité essentielle : la possibilité de raisonner sur l'adresse d'un mot ou d'un groupe de mots sans connaître explicitement cette adresse.

En effet, les graphes de repérage jouent un rôle comparable à celui des systèmes de coordonnéesmobiles en géométrie. Ce sont des repères flottants.

Soit une phrase inconnue, une phrase que nous n'avons pas encore lue: il n'est pas possible de dire si le sommet de l'arborescence de repérage, en forme de graphe Harper-Hays, que l'on peut associer à cette phrase correspondra au premier mot dans l'ordre d'énonciation, ou bien au second, ou bien à n'importe quel autre. Cela ne veut pas dire que la correspondance soit quelconque, bien au contraire. La hiérarchie de l'arborescence est liée à l'ordre linéaire de la phrase par des relations très précises; mais comme celles-ci font intervenir certains renseignements (nature des mots, etc) qui ne seront connus qu'avec la donnée de la phrase, il n'est pas possible de donner explicitement la position de l'arborescence sans connaître la phrase.

Puisque les adresses sur l'arborescence sont liées aux adresses linéaires inconnues, raisonner sur les premières, c'est raisonner indirectement sur les secondes.

Si l'on veut bien se rappeler que c'est précisément sur cette faculté qu'il offre, de raisonner sur des quantités inconnues sans les expliciter, que s'est fondé l'immense succès du calcul algébrique, on aura une idée des services que peuvent rendre les graphes d'adressage en analyse automatique et en traduction automatique.

Les problèmes de phrase inconnue sont typiquement ceux que pose la mise au point de ces techniques. Il est certain que lorsqu'une phrase entre dans une machine pour être analysée ou traduite, cette phrase devient bien connue et bien déterminée; mais il est alors trop tard pour faire le programme ou le modifier. Ce programme doit en effet être construit avant, de sorte qu'au moment de sa construction, on ne sait pas quelles phrases seront traitées.

Le programme doit être conçu de façon à analyser correctement ces phrases futures et inconnues quelles qu'elles soient.

Pour cela, on doit équiper ce programme de catalogues grammaticaux et sémantiques, contenant les règles normatives qu'une phrase inconnue quelconque est censée devoir respecter.

Pour la levée des ambiguïtés dans ces phrases futures et inconnues, l'ordre des mots jouera certainement un rôle important. (Exemple : il porte la ferme; il ferme la porte). Aussi les catalogues devront-ils mentionner les règles qui expriment des relations entre les rangs des mots dans ces phrases inconnues.

Le procédé qui consiste à donner en fonction d'un graphe de repérage les adresses des mots liés par telle ou telle règle du catalogue répond à ce besoin. Ce procédé équivaut en effet à l'énoncé, chaque fois, d'une relation entre adresses linéaires inconnues.

II, 3 INFLUENCE SUR LE PRIX DES CATALOGUES.

Les graphes d'adressage permettent de réduire, par un perfectionnement des procédés de calcul d'adresses, la contrainte très lourde de l'énonciation d'adresses explicites dans les catalogues de règles grammaticales et sémantiques.

Une règle n'étant utilisable que dans la mesure où l'on connaît les adresses linéaires, dans l'ordre d'énonciation de la phrase, des mots ou autres éléments entre lesquels elle établit des liens, il faut satisfaire cette exigence, et énoncer les adresses de ces éléments en même temps que l'on énonce la règle.

A/ LES CATALOGUES ADRESSES LINEAIREMENT

Certains auteurs indiquent explicitement ces adresses dans leurs catalogues. Il en résulte, certes, une grande facilité de consultation, mais aussi un alourdissement et une complication si considérables de ces catalogues, que le travail de compilation de règles devient extrêmement difficile.

Parmi les tenants de cette tendance, caractérisée par une "consultation" facile de catalogues compliqués, on peut citer l'école de Z. Harris à l'Université de Pennsylvanie. Les règles normatives décrites et rassemblées, par exemple, par Aravind K. Joshi (8) ou H. Hiz (9) mentionnent chaque fois les adresses linéaires des mots entre lesquels elles introduisent des relations.

Ces règles s'énoncent par exemple comme suit: "Si le mot de rang x-3 appartient à la catégorie c_3, le mot de rang x-2 à la catégorie c_2, le mot de rang x-1 à la catégorie c_1, alors le mot de rang x n'appartient sûrement pas à la catégorie c_0". Toutes les adresses étant données, la consultation se fait très simplement.

Par contre, la construction d'un catalogue de ce genre est extrêmement longue et difficile, et les règles qu'il rassemble sont bien différentes de celles que l'on peut trouver dans des grammaires ordinaires.

Il est facile de montrer que dans ces notations, l'expression de la règle sujet verbe comporterait un nombre quasi infini de lignes. Entre le sujet et le verbe d'une phrase, on peut en effet intercaler des configurations d'allure cyclique: "L'homme qui a vu l'homme qui a vu l'homme etc.... est venu". Ces configurations peuvent même combiner des cycles différents.

Aussi, l'un des grands problèmes, pour les chercheurs de cette tendance, est d'arriver à replier le catalogue sur lui-même afin de lui donner des dimensions convenables et pouvoir le construire; d'où un certain nombre d'études, parmi lesquelles celles de R. S. Solomonoff (10), visant à permettre l'identification automatique de cycles. Mais c'est de l'Université de Pennsylvanie, avec Z. Harris (11) (12), puis N. Chomsky (4), qu'est partie l'idée extrêmement remarquable de faire appel à des "transformations linguistiques" pour réduire les dimensions des catalogues de règles.

L'un des objectifs principaux de la linguistique mathémathique aux U.S.A. demeure la simplification de ces catalogues à adresses explicites, et la réduction de leur volume à des proportions acceptables. Il est certain qu'une fois construits, de tels catalogues seront très faciles à consulter. Aussi, lorsque l'on choisit une politique d'adressage linéaire explicite, les difficultés se concentrent presque toutes dans les opérations de compilation du catalogue.

B/ LES CATALOGUES ADRESSES SELON DES GRAPHES

A l'opposé de cette première tendance, on voit se dessiner, depuis plusieurs années et chez un nombre croissant d'auteurs dans le monde, l'amorce d'une seconde politique, qui viserait à simplifier les catalogues en reportant la difficulté sur le programme de consultation.

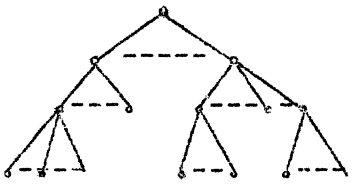
Leurs catalogues sont adressés selon des graphes, et non pas directement en adresses de l'ordre linéaire. Il en résulte un allègement certain de ces catalogues. Par contre, pour utiliser les règles qui y sont contenues, il faut déterminer la correspondance entre les adresses sur le graphe et les adresses linéaires, c'est-à-dire faire un calcul.

Parmi ces auteurs, le professeur Ceccato (3) occupe certainement une place à part. C'est à lui et à son école (13) (14), que revient le très grand mérite d'avoir utilisé les rectangles de corrélation comme graphes d'adressage indirect, en partant de l'idée que la pensée procéderait de cette façon.

Grâce au réseau correlationnel, des règles grammaticales et sémantiques peuvent être énoncées sans indication explicite des

adresses linéaires des éléments qu'elles lient.

Ci-dessous, un exemple de graphe utilisé par l'équipe de traduction automatique du C.N.R.S. -Paris (15)



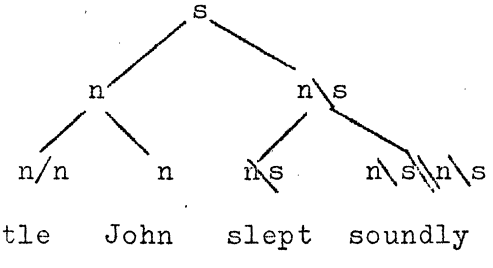
Harper, Hays (6) et l'équipe de la Rand Corporation présentent leurs arborescences comme étant des graphes linguistiques, mais ils s'en servent visiblement comme de graphes de repérage, et se guident sur des considérations d'économie ou de commodité pour en fixer la forme.

Enfin, Bar Hillel (17) donne un exemple de procédure d'analyse où l'on utiliserait, comme repère d'adressage indirect, le diagramme de dérivation bien connu :

Ce n'est pas ce graphe qui est nouveau, c'est l'idée de s'en servir comme repère de calcul d'adresses.

Grâce à des adresses indirectes, représentées par certaines positions le long de ce groupe, Bar Hillel peut écrire des règles d'assemblages sans indiquer explicitement des adresses linéaires

$$n/n + n \longrightarrow n$$
$$n\backslash s + n\backslash s\backslash\backslash n\backslash s \longrightarrow n\backslash s$$
$$n + n\backslash s \longrightarrow s$$



Cette méthode, inspirée en partie par Lambeck (18), a été reprise par Richard S. Glantz (19) dans le rapport NBS 6856.

Le procédé d'assemblage indiqué par Bar Hillel est un peu incorrect. (Il ne satisfait pas aux règles générales d'assemblage qui seront énoncées dans la troisème partie). Bar Hillel déclare (12) d'ailleurs rencontrer certaines difficultés.

II, 4 ELEMENTS DU COUT DE L'ENONCIATION D'UNE REGLE DANS UN CATALOGUE

Donnons quelques indications sur le coût de l'énonciation d'une règle dans un catalogue, en relation avec le système d'adressage de ce catalogue. Nous considérerons uniquement le cas de règles liant des mots (on verra plus loin que le cas des règles liant des groupes de mots peut, par certains procédés, être ramené à celui-ci).

a/ L'adresse relative d'un mot par rapport à un autre doit être donnée lorsque l'on commande de vérifier que leurs natures satisfont à certaines relations ou règles d'accord. La grammaire exige par exemple que l'on s'assure de l'accord en genre et en nombre d'un sujet et de son présumé attribut. La sémantique suggère certains autres contrôles. Si un "chien" est déclaré être vivipare, ou non-ovipare, il s'agit vraisemblablement de l'animal-chien, et non pas de l'objet chien de fusil. Pour ordonner ces contrôles, il faut indiquer l'adresse de l'attribut par rapport au sujet ou vice versa.

b/ L'adresse relative de deux mots est bien définie par la description du chemin qui conduit de l'un à l'autre (nombre de mots intermédiaires, nature des mots intermédiaires, nature ascendante ou descendante des arêtes s'il s'agit d'une arborescence).

Ceci tient au fait que la relation d'adressage est transitive :

Si la position de A est connue par rapport à celle de B,

etque celle de B est connue par rapport à celle de C,

alors, la position de A est connue par rapport à celle de C.

On peut étendre ce raisonnement au cas d'un nombre quelconque d'intermédiaires.

c/ L'adresse relative de deux mots est d'autant moins coûteuse, qu'elle s'énonce d'un petit nombre de façons différentes. Un exemple d'adresse très chère, est celle de l'attribut par rapport à son sujet, lorsque l'on utilise comme repère l'ordre linéaire d'énonciation des mots dans le discours. Le nombre et les natures des mots qui séparent le sujet de son attribut sont en effet extrêmement variables d'une phrase à l'autre :

"Cette femme est contente"

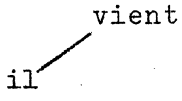
"Cette femme, c'est dans sa nature, est toujours contente"

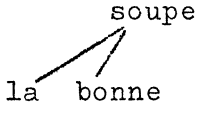
"Cette femme, c'est dans sa nature et nous n'y pouvons rien changer, n'est jamais, jamais, contente".

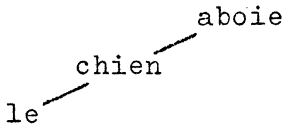
L'adresse est d'autant plus coûteuse à donner, qu'il faut pour cela énoncer un plus grand nombre de configurations intermédiaires possibles. La règle sujet attribut est donc très coûteuse en repère linéaire.

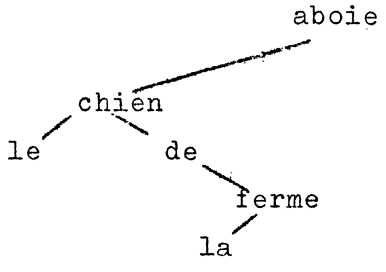
d/ Etudions quelques exemples. L'adresse la plus simple et la moins coûteuse que l'on puisse imaginer, est celle qui s'énonce, toujours de la même manière et sans mot intermédiaire, selon un lien direct.

Or la plupart des couples de mots concernés par les règles grammaticales du type: accord en genre, accord en nombre, accord en personne, etc., sont, dans les arborescences de Tesnière ou dans celles de Harper et Hays, liés directement;







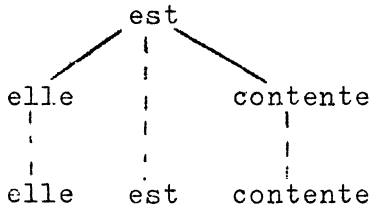


Pour certains couples toutefois, l'on doit renoncer à un lien direct; il sera quand même possible d'énoncer les adresses relatives de ces éléments liés, grâce à la transitivité de la relation d'adressage.

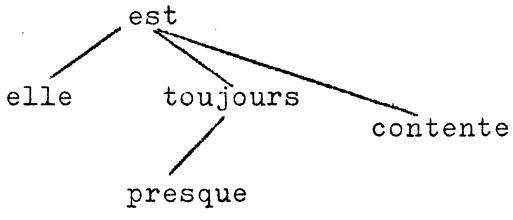
C'est d'ailleurs ce que nous faisions plus haut en repère linéaire pour la règle sujet attribut et il en résultait un coût élevé.

Dans le repère de Tesnière (5) comme dans celui de Harper et Hays (6), il n'y a pas de lien graphique direct entre le sujet et l'attribut. Voyons les répercussions de cette situation sur le coût d'énonciation de la règle sujet attribut.

Dessinons une arborescence :



Le chemin de "elle" à "contente" comporte toujours, le long de l'arborescence, un seul mot intermédiaire.



Ce mot intermédiaire est toujours un verbe d'état, supérieur hiérarchique commun de "elle" et "contente".

Au total, pour donner l'adresse de l'attribut par rapport au sujet, il suffit de décrire une seule configuration, ce qui est très peu coûteux.

On remarquera que l'absence de lien direct entre le sujet et l'attribut dans ce type de graphe n'entraîne pas d'augmentation fâcheuse du coût d'énonciation de la règle sujet attribut.

e/ Les éléments d'évaluation qui viennent d'être donnés concernent uniquement le coût d'énonciation d'une règle dans un catalogue.

Certes, la complexité de la forme sous laquelle on énonce les règles peut avoir des répercussions sur le coût des opérations de consultation; mais ce dernier dépend aussi d'autres facteurs, qui seront étudiés plus loin.

TROISIEME PARTIE

LES OPERATIONS D'ASSEMBLAGE
RENTABILITE D'UNE FAMILLE DE GRAPHES DE REPERAGE

III, 1 LE PROBLEME DE L'ASSEMBLAGE D'UN GRAPHE DE REPERAGE CORRESPONDANT A UNE PHRASE DONNEE.

"Assembler" le graphe de repérage correspondant à une phrase donnée, c'est calculer la position de ce graphe en fonction du repère de l'ordre linéaire d'énonciation de cette phrase.

On se rappelle que les adresses des catalogues grammaticaux et sémantiques sont exprimées selon le graphe de repérage, et qu'il en résulte d'ailleurs une importante simplification de ces catalogues.

Il faut maintenant payer cet avantage par un travail supplémentaire de calcul. En effet, la phrase donnée, et toutes les hypothèses interprétatives que l'on peut formuler à son sujet, sont exprimées en fonction du repère linéaire.

Pour lever les ambiguïtés de la phrase, il faut confronter ces hypothèses avec les catalogues grammaticaux et sémantiques; il faut donc changer de repère.

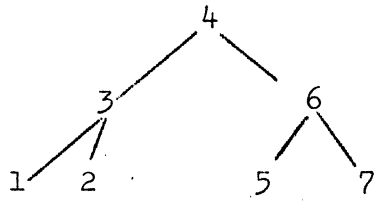
Si l'on dispose d'une machine puissante, ce supplément de calcul ne constitue pas une rançon bien lourde, en comparaison des avantages tirés de la simplification des catalogues. On voudra bien se souvenir, en effet, que les catalogues relèvent de méthodes de construction artisanales, du moins jusqu'à présent. Les calculs au contraire se font à la machine.

Il existe divers procédés de calcul d'assemblage. Le choix d'une méthode doit en effet être guidé, dans une large mesure, par l'observation des caractéristiques particulières de la famille de graphes utilisée, si bien que dans les ensembles d'analyse automatique à filtres multiples où l'on profite de la possibilité qu'il y a de faire varier, d'un filtre à l'autre, le type de graphe d'adressage utilisé, la méthode d'assemblage variera elle aussi d'un filtre à l'autre. Dans le choix d'une procédure d'assemblage pour un certain filtre, on doit d'autre part tenir compte des possibilités de l'ordinateur dont on dispose. Il reste finalement une certaine latitude de choix, car,

pour une même situation et de mêmes données, il y a souvent plusieurs manières équivalentes de faire un calcul. Nous allons essayer cependant de dégager certains aspects généraux des calculs d'assemblage.

A/ C'est l'existence de calculs d'assemblage qui caractérise les procédures d'analyse automatique à graphes de repérage flottants.

Lorsque Harper et Hays assemblent, trait à trait, leurs graphes en fonction de phrases données (6) l'opération se traduit par l'inscription de numéros aux sommets de ces graphes :



Et "assembler" un graphe en effet, n'est rien d'autre que fixer sa position en fonction de l'ordre linéaire d'énonciation des mots.

De même, bien qu'utilisant une procédure différente et des graphes différents, le professeur Ceccato assemble, construit, ses réseaux corrélationnels, en inscrivant, dans les rectangles de corrélation, des numéros d'ordre linéaire d'énonciation de mots dans les phrases données.

Si enfin, nous avons classé parmi les méthodes à graphes flottants la procédure de Bar Hillel(17), procédure dont cet auteur lui-même dénonce le caractère imparfait, c'est en raison du fait qu'elle comporte des mécanismes d'assemblage.

En réalité, il faudrait modifier un peu cette procédure pour en faire vraiment un méthode à graphes de repérage flottants et lever du même coup les objections de Bar Hillel.

B/ Seconde caractéristique générale des calculs d'assemblage : leur déroulement est contrôlé surtout par l'observation des résultats. De ce fait, les procédures d'assemblage sont toujours plus ou moins des méthodes de tâtonnement automatique . On peut cacher un peu cet aspect et accélérer les opérations, d'une part en conduisant plusieurs essais de tâtonnement à la fois, grâce à l'accumulation d'hypothèses dans un réservoir (Ceccato) et, d'autre part, en faisant intervenir l'observation de résultats partiels dans le contrôle du calcul. Ces procédés permettent d'économiser du temps, par prévision de l'échec de certains tâtonnements non encore effectués réellement, ce qui rend possible leur élimination à l'état virtuel. Le schéma du tâtonnement n'en subsiste pas moins.

Pourquoi en est-il nécessairement ainsi ? En raison du principe même des techniquesde graphes de repérage flottants. Tout le

bénéfice de ces techniques réside dans la simplification des catalogues grammaticaux et sémantiques qui intervient lorsque ces catalogues sont adressés, non plus en fonction de l'ordre linéaire, mais en fonction du graphe de repérage.

Mais de ce fait, on ne peut consulter ces catalogues que dans la mesure où l'on peut exprimer les adresses relatives des mots de la phrase en termes de rapports d'adressage dans le graphe de repérage, c'est à dire dans la mesure où l'on a, au moins partiellement, construit ce graphe.

Ainsi, les informations grammaticales et sémantiques ne peuvent guider le calcul d'assemblage que par l'intermédiaire des résultats, partiels ou non, de ce calcul.

Comme ces informations sont indispensables au contrôle du calcul, il y a tâtonnement.

Les méthodes de tâtonnement ne sont pas forcément lentes. Leur convergence peut se révéler très rapide.

C/ C'est un caractère commun à toutes les procédures de tâtonnement, que de comporter les deux fonctions essentielles suivantes : d'une part, une fonction de formulation d'hypothèses, ou si l'on veut, de mise en oeuvre des essais; d'autre part, une fonction d'appréciation du résultat des essais, avec éventuellement choix de l'ordre des essais suivants.

Le moins que l'on puisse exiger de l'organe à effectuer des essais, c'est qu'il satisfasse aux conditions que voici :

- d'une part, aptitude virtuelle à faire n'importe quel essai. Cela signifie que, une famille de graphes de repérage étant choisie et bien définie, l'organe chargé de proposer des assemblages devra être virtuellement apte à construire n'importe quel graphe associable à une phrase, c'est-à-dire n'importe quel graphe de la famille;

- d'autre part, aptitude à proposer autant de solutions que l'organe de contrôle veut bien en accepter. Au niveau des résultats cela impose de continuer à faire des essais même lorsqu'une solution a été acceptée. Au niveau des résultats partiels, cela impose de continuer à essayer tous les chemins, même lorsque l'un d'eux a été accepté en tant que résultat partiel.

D/ Les opérations d'assemblage ne peuvent être menées à bout que si les hypothèses faites sur les natures des mots de la phrase donnée sont correctes, ou du moins, si ces hypothèses sont en accord avec l'ensemble des règles contenues dans les catalogues grammaticaux et sémantiques. La procédure d'assemblage a donc la valeur d'un filtre, qui répondrait oui aux hypothèses correctes en acceptant de les assembler, et non aux hypothèses incorrectes, en refusant d'en faire l'assemblage.

Rappelons que la notion de filtre, et le rôle des filtres, ont été définis en I,4.

E/ En tenant compte de ce qui précède, on peut décrire au moins une procédure générale d'assemblage, valable pour n'importe quelle famille de graphes de repérage.

Il s'agit de la méthode triviale, qui consiste à construire un par un tous les graphes de la famille associables à une phrase de m mots, si m est le nombre de mots de la phrase donnée. Soit l'un de ces graphes. Puisqu'il est "construit", les positions relatives des mots de la phrase sont exprimables selon ce graphe ; on peut donc consulter les catalogues grammaticaux et sémantiques et vérifier si ces rapports d'adressage selon le graphe enfreignent ou non des règles normatives. L'assemblage proposé est alors accepté ou refusé, et ainsi de suite. Si les hypothèses faites sur les natures des mots sont correctes, et que la phrase est d'autre part correcte, c'est-à-dire interprétable en fonction des règles grammaticales et sémantiques contenues dans les catalogues dont on dispose, il existe au moins une solution, c'est à dire au moins un graphe qui ne conduise pas à des violations de règles de ces catalogues. Comme tous les graphes possibles sont proposés, celui-là l'est aussi, et il est accepté.

Cette procédure est sans doute la plus longue, la plus lourde et la plus simple que l'on puisse imaginer. Il n'est pas question de l'utiliser telle quelle pratiquement. Mais en tant que méthode générale, elle a une grosse importance théorique. Elle est en effet parfaitement automatisable, et applicable en droit au cas de n'importe quelle famille de graphes de repérage.

L'existence d'une telle méthode justifie en effet l'emploi de graphes de repérage en général. Elle montre d'autre part qu'entre deux tracés graphiques, le choix est arbitraire, puisqu'il existe une méthode générale indépendante du tracé, si bien que seules des considérations de prix peuvent guider ce choix. On suppose naturellement que les deux tracés se prêtent l'un et l'autre à un repérage non ambigu, et permettent d'exprimer les règles des catalogues.

En pratique il faut évidemment utiliser des procédures plus rapides, mais il est bon de savoir que le problème de l'assemblage peut toujours être résolu.

III, 2 LES METHODES PAR ASSEMBLAGE PROGRESSIF.

Les méthodes par assemblage progressif ne sont que des variantes accélérées de la procédure générale théorique décrite au chapitre précédent, variantes dans lesquelles on vérifie les graphes au fur et à mesure de leur construction.

Supposons que cette construction commence par l'établissement d'un lien direct entre le $n^{ième}$ et le $p^{ième}$ mot de la phrase donnée, puis que la consultation du catalogue montre une incompatibilité entre certaines règles normatives et l'existence de ce trait dans le graphe. Le trait est alors barré et l'on essaie d'autres possibilités.

Or, en supprimant le trait, on élimine d'un seul coup tous les graphes qui l'auraient contenu.

On rejette ainsi des solutions fausses, non plus une à une, mais paquets par paquets. Ces graphes se trouvent, de plus, éliminés à l'état virtuel, c'est-à-dire sans que l'on ait dû les construire complètement. Il en résulte une grosse économie d'opérations et de temps.

Quelques remarques :

a/ on doit prendre garde à ne pas perdre de temps en proposant le même graphe plusieurs fois. Il existe en effet plusieurs manières différentes d'assembler un même graphe morceau par morceau. Aussi convient-il de compléter la définition de la famille de graphes de repérage utilisés en leur imposant, non seulement certaines caractéristiques de forme, mais encore un certain ordre d'assemblage, en liant par exemple cet ordre à la forme du graphe ou à un système d'orientation;

b/ le contrôle des graphes en cours de construction s'exercera alors à la fois sur leur forme et sur l'ordre des opérations qui ont conduit à les proposer. Si cet ordre n'est pas conforme à celui prescrit dans la définition de la famille, on rejettera le graphe même si sa forme est correcte;

c/ lorsque l'on met en place les mécanismes d'une méthode par assemblage progressif, on doit faire bien attention à ne pas introduire d'erreur systématique par élimination automatique de solutions permises. Lorsque le catalogue autorise la construction d'un certain trait du graphe, cela ne signifie pas qu'il est interdit de ne pas construire ce trait. Aussi les deux éventualités (trait construit, trait non construit) devront-elles être l'une et l'autre envisagées avec toutes leurs conséquences.

Les méthodes par assemblage progressif ont une grosse importance à plusieurs points de vue. D'une part, elles sont assez rapides pour être réellement utilisées; on pourra faire appel à elles pour construire une grande variété de filtres; on peut fonder sur elles des calculs de rentabilité. D'autre part, leur principe ne change pas lorsque l'on modifie, légèrement par exemple, le graphe d'adressage utilisé. La fonction "rentabilité d'une famille de graphes" reste ainsi définie lorsque l'on fait varier la forme et les caractéristiques de ces graphes, puisqu'il ne cesse jamais d'y avoir une méthode de consultation possible par assemblage progressif. Ces considérations de rentabilité permettront de guider le choix de graphes de repérage convenables.

III,3 REPERAGES D'ADRESSES DE MOTS A L'INTERIEUR DE GROUPE DE MOTS, INFLUENCE SUR LE COUT.

La question qui se pose au moment de chaque assemblage partiel, dans les méthodes à assemblage progressif, peut être formulée dans les termes suivants : étant données les deux fractions de graphe que l'on se propose de réunir par un trait, étant données les caractéristiques des mots et des groupes de mots qu'elles contiennent, les catalogues autorisent-ils l'assemblage de ces deux fractions de graphe ?

Avant de résoudre automatiquement cette question, il faut au moins la poser automatiquement, c'est-à-dire donner les caractéristiques des mots et groupes de mots composant les fractions à assembler.

Nous allons montrer qu'il faut se garder de donner à l'avance ces caractéristiques de façon explicite, sous peine de faire constamment dans la machine des transports énormes d'information . On verra alors qu'il est plus commode de donner ces renseignements sous forme implicite, en indiquant simplement les adresses où l'on peut les trouver.

Chacune de ces caractéristiques peut jouer un rôle un jour ou l'autre, mais dans le cas d'une opération d'assemblage déterminée, seules certaines d'entre elles sont pertinentes. Le catalogue commandera alors la formulation explicite de ces seules caractéristiques pertinentes pour l'opération considérée. Pour ces caractéristiques utiles seulement, il faudra aller chercher le détail des renseignements aux adresses indiquées.

A/ LE NOMBRE DE CATEGORIES DE MOTS EST GRAND

Pour caractériser la réactivité d'un mot dans une certaine acception, ou, si l'on veut, ses possibilités d'emploi dans cette acception, il faut faire intervenir un bon nombre de composantes.

Composantes grammaticales: s'agit-il d'un nom, d'un verbe, etc. A-t-il un genre, un nombre, une personne, un mode, un temps, etc.

Composantes sémantiques, plus nombreuses encore: si c'est un objet, quelle est sa forme, son volume, etc. Si c'est une action, peut-elle être le fait d'un être inanimé, etc.

Le professeur Ceccato utilise de grandes fiches pour consigner ces renseignements.

Il paraît souhaitable d'inscrire ces caractéristiques une seule fois en mémoire machine (ou au maximum deux, si le dictionnaire général est en mémoire lente que l'on utilise des dictionnaires réduits en mémoire rapide).

Pour cela, on ne transportera pas dans les calculs ces caractéristiques explicites, mais simplement leurs adresses.

B/ LE NOMBRE DE CATEGORIES DE GROUPES DE MOTS EST QUASI INFINI.

Soit un groupe de mots, qui sont pris chacun dans une certaine acception, le groupe étant lui même pris dans une acception bien déterminée. Voyons quels éléments caractérisent la réactivité d'un tel groupe de mots, ou, si l'on veut, les possibilités d'emploi de ce groupe dans l'acception considérée. Pour cela étudions quelques exemples :

On admet généralement que la séquence

"petit Pierre"

constitue une chaîne nominale et qu'il est permis de lui adjoindre à gauche un adjectif masculin singulier :

gentil + petit Pierre ⟶ gentil petit Pierre

ce que Bar Hillel écrirait: n/n + n ⟶ n

Mais il n'en serait pas de même si l'adjectif était par exemple au pluriel :

gentils + petit Pierre ⟶ alarme

Il est donc important de signaler que "petit Pierre" est une chaîne nominale du type "masculin singulier".

Bar Hillel suggère d'ailleurs (17) d'augmenter en conséquence le nombre de ces catégories :

gentils = n/n masculin pluriel

gentil = n/n masculin singulier

petit Pierre = n masculin singulier

n/n	+	n	⟶ n
(masc. sing.)		(masc. sing.)	(masc. sing.)

n/n	+	n	⟶ alarme
(masc. plur.)		(masc. sing.)	(c'est-à-dire interdiction d'assembler)

Rappelons qu'il est utile de vérifier que ces règles d'accord sont bien respectées; non pas que l'on mette en doute le style des auteurs des textes donnés à l'entrée d'une machine à traduire; ce qui est douteux, c'est l'interprétation donnée à chaque mot susceptible d'homonymies, c'est l'acception que l'on prête aux mots, et, partant, aux groupes de mots, dans la phrase à analyser.

La règle d'accord de l'adjectif donnée ci-dessus, peut permettre de lever une homographie telle que celle du mot "braves" dans la phrase "tu braves petit Pierre"

braves$_1$ = verbe 2ème personne singulier

= $n\backslash s /\!/ n$

braves$_2$ = adjectif

= n/n pluriel

On voit alors que "braves$_2$" ne peut s'assembler avec "petit Pierre" :

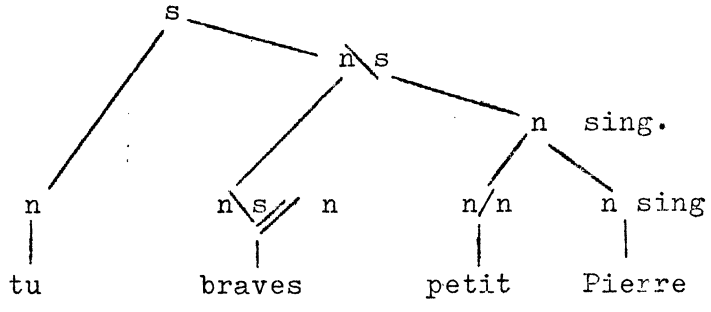
braves$_2$ + petit Pierre $\longrightarrow$ alarme

en effet

n/n + n $\longrightarrow$ alarme

plur. sing.

"braves$_2$" ne peut pas non plus s'assembler avec "tu". D'où résolution de l'homographie : dans la phrase "tu braves petit Pierre", braves est sûrement un verbe : l'analyse se fait finalement comme suit, en prenant par exemple la méthode de Bar Hillel :



Ainsi, il est important de bien caractériser la "réactivité" d'un groupe de mots, c'est-à-dire son aptitude à avoir des rapports avec d'autres éléments de la phrase. On montre facilement que, dans le cas très simple d'une chaîne nominale :

a/ toutes les caractéristiques du nom principal jouent un rôle dans la réactivité de la chaîne. Caractéristiques grammaticales (cf. ci-dessus) mais aussi caractéristiques sémantiques : Bar Hillel propose par exemple de distinguer des chaînes nominales d'objets et des chaînes nominales d'êtres animés. Dans une phrase comme "Le vieux chien aboie", ces distinctions permettront d'éliminer l'hypothèse chien$_1$ = objet inanimé (chien de fusil) ou du moins, de la déclarer improbable.

b/ bien qu'intervenant moins souvent, les caractéristiques des autres mots sont susceptibles d'influer sur la réactivité de la chaîne nominale.

C'est ainsi qu'une chaîne nominale sans article à gauche peut annexer à gauche un adjectif, peut annexer à gauche un article.

$$\text{le} + \text{cheval} \longrightarrow \text{le cheval}$$

$$\begin{bmatrix} n/n \\ art \end{bmatrix} + \begin{bmatrix} n \\ \text{sans art} \end{bmatrix} \longrightarrow \begin{bmatrix} n \\ art \end{bmatrix}$$

mais :

$$\text{le} + \text{le cheval} \longrightarrow \text{alarme}$$

$$n/n + n_{art} \longrightarrow \text{alarme}$$

ce qui oblige à distinguer les sous catégories $n_{\text{sans art}}$ d'une part, n_{art} d'autre part.

Si la première lettre à gauche dans le groupe de mot est une voyelle, le groupe aura de l'affinité pour "l". Sinon, il n'admettra que "le" ou "la".

ces caractéristiques augmentent déjà beaucoup le nombre de chaînes nominales. On montre facilement que les articles situés à droite du nom principal jouent aussi un rôle , exemple :

un panier d'osier de Pierre

un panier de l'osier de Pierre.

Si enfin on fait intervenir des critères sémantiques, on est bien obligé de constater que tous les mots de la chaîne risquent d'avoir une influence sur la réactivité de celle-ci; par suite, leur prise en considération peut être utile pour lever certaines homonymies difficiles.

Encore ne s'est-il agi que de chaînes nominales, c'est-à-dire d'un cas assez simple. Il faut faire des études analogues pour les chaînes plus complexes.

Le nombre d'espèces de chaînes devient alors colossal, même pour un nombre relativement faible de mots. Or, il ne s'agit que de distinctions pertinentes (ou plutôt susceptibles d'être pertinentes dans certains cas). Rendre ces distinctions impossibles, ce serait perdre systématiquement de l'information.

Dans ces conditions, le mode de traitement suivant s'impose :

- d'une part, on évitera comme pour les mots, de transporter les caractéristiques explicites des groupes de mots. On transportera simplement les adresses d'endroits où ces caractéristiques sont accessibles;

- mais d'autre part, on ne pourra pas faire un dictionnaire de groupes de mots, comme l'on avait fait un dictionnaire de mots. Les caractéristiques des groupes de mots ne pourront pas être trouvées dans une liste, elles devront être le résultat d'un calcul;

- l'adresse de l'endroit ou les caractéristiques d'un groupe de mots seront accessibles, c'est, dans de telles conditions, l'adresse d'attaque de la sous routine qui permet de les calculer explicitement;

- le dictionnaire de groupes de mots n'en existera pas moins, mais il sera réduit au strict minimum. Seules y figureront les caractéristiques qui ne peuvent être calculées sur la base des caractéristiques des mots composants. On trouvera en particulier dans ce dictionnaire les locutions et les idiotismes.

C/ REPERAGE DES MOTS A L'INTERIEUR DES GROUPES DE MOTS

Pour calculer certaines des caractéristiques d'un groupe de mots à partir de celles des mots constituants, il faut bien repérer les adresses de ces mots constituants.

Celà signifie qu'il faudra structurer les adresses des éléments composant ces groupes de mots, de façon que l'on puisse retrouver rapidement les renseignements correspondants.

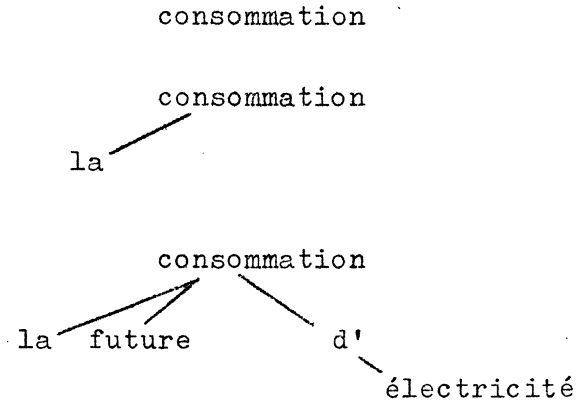
L'ordre linéaire des mots répond relativement mal à ce besoin. Il ne permet pas, en effet, d'attribuer des adresses fixes aux mots qui portent les renseignements importants. Le nom principal d'une chaîne nominale est-il le premier mot en partant de la gauche, le second, etc ? Cela dépend. Le sujet d'une proposition subordonnée est-il le premier substantif de cette proposition, le second, etc ? Celà dépend.

Au contraire, la structuration interne des adresses d'un groupe de mots au moyen d'un graphe de repérage permet d'attribuer des adresses fixes aux renseignements importants, et la plupart du temps, des adresses conditionelles relativement simples pour l'ensemble des renseignements utiles.

De cette structuration interne des adresses au moyen de graphes de repérage, dépendent les temps d'accès aux divers renseignements. Dans le choix d'un type de graphe de repérage, il faudra donc tenir compte des fréquences d'utilisation de ces divers renseignements. Ces fréquences dépendent du travail que doit exécuter le filtre considéré. Une fois de plus, il apparaît que l'on peut avoir intérêt à choisir des types de graphes de repérage différents pour des filtres différents.

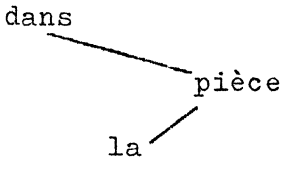
Donnons quelques exemples de structuration interne de groupes de mots selon les arborescences de Tesnière, ou Harper et Hays :

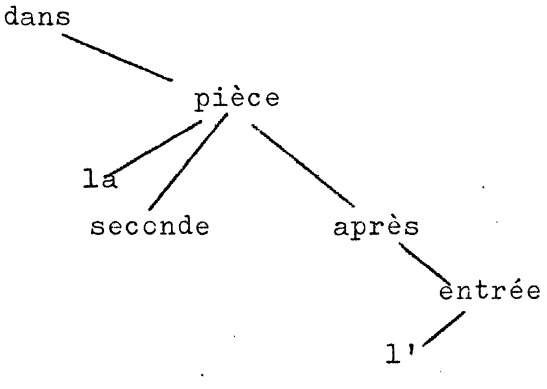
Dans une chaîne nominale, les caractéristiques qui conditionnent le plus souvent la réactivité de la chaîne sont liées à la nature du nom principal. Aussi est-il normal de mettre celui-ci en tête de l'arborescence :



Lorsque l'on veut aller chercher d'autres renseignements (existence d'un article à gauche, etc.) celà peut être un peu plus long, s'il faut passer par l'intermédiaire de l'adresse du mot de tête. Mais ces renseignements sont moins fréquemment utiles.

Des observations analogues pourraient être faites dans le cas de groupes de mots quelconques. Exemple :





La fonction du groupe de mots est principalement caractérisée par le mot "dans". La nature du substantif qui indique "dans quoi", doit aussi intervenir assez souvent et il est normal de mettre le mot "pièce" en seconde position et ainsi de suite.

Dans les travaux pratiques de la journée de linguistique, il a été fait allusion à ces problèmes de structuration des adresses dans les groupes de mots.

III, 4 COÛT GLOBAL DE CONSTRUCTION ET D'UTILISATION D'UN FILTRE. RENTABILITE D'EMPLOI D'UNE FAMILLE DE GROUPES DONNEE.

A/ COÛT DE CONSTRUCTION D'UN FILTRE

a/ Les catalogues :

L'élément de loin le plus coûteux, lors de la construction d'un filtre de quelque importance, est le travail de rassemblement du catalogue de règles grammaticales et sémantiques que le filtre aura pour objet d'exploiter. On a déjà donné une idée des difficultés que rencontre la compilation de catalogues consultables automatiquement. Le volume de ces catalogues coûtera d'autre part de la place en mémoire machine.

Si l'on fait appel aux techniques de graphes de repérage, c'est essentiellement dans le but de réduire les dépenses de catalogue.

En II, 4, nous avons déjà donné des indications sur la manière d'évaluer le coût d'inscription d'une règle dans un catalogue et ses variations en fonction de la nature des graphes de repérage utilisés.

b/ Le programme :

Nous le citons pour mémoire. Son coût de construction est généralement faible devant celui des catalogues.

B/ COÛT D'UTILISATION D'UN FILTRE

A l'emploi, un filtre coûte principalement du temps machine. Pour faire des évaluations précises, il faudrait tenir compte de la nature de la machine utilisée.

Cependant, les variations dues aux différences entre les machines, variations non négligeables, apparaissent dans certains cas comme assez faibles si on les compare à la multiplication astronomique de dépense qui peut résulter de l'emploi d'une procédure maladroite. Dans les problèmes combinatoires, et celui des homonymies en est un, le nombre d'opérations à faire peut croître dans des proportions fantastiques, si l'on n'y prend pas garde.

Bien que constituant un mode grossier d'évaluation des prix de revient, le décompte des opérations donne certainement des indications utiles.

Comme les programmes d'assemblage sont des programmes de tâtonnement, le nombre d'opérations de tâtonnement joue certainement une rôle grossièrement multiplicatif en ce qui concerne le temps machine, le second facteur étant la durée moyenne de chaque tâtonnement.

a/ Nombre de tâtonnements.

On se rappelle que les tâtonnements ont pour objet la formulation d'hypothèses sur les positions relatives du graphe de repérage et du repère linéaire, puis l'élimination des hypothèses impossibles. S'il y a N éventualités de positions relatives, dont une seule bonne, il faudra éliminer les N-1 autres éventualités.

Pour simplifier, aulieu de dire "positions de graphe" nous dirons "graphes" tout court. Il y a donc N possibilités de graphes et on désire éliminer ceux qui contredisent les règles du catalogue.

Dans la méthode générale théorique citée comme curiosité au III, 1, E, le nombre de tâtonnements est égal à N.

Dans la méthode par assemblage progressif du III,2, on peut espérer en s'y prenant bien, obtenir que le nombre de tâtonnements soit à peu près proportionnel à Log. N.

Donnons quelques ordres de grandeur.

Si les graphes utilisés sont des arborescences quelconques ayant pour sommets les mots de la phrase donnée, N est le nombre d'arborescences à m sommets. En voici quelques valeurs :

Nombre m de mots de la phrase donnée	N	$\log_{10} N$
10 mots	10^9	9
20 mots	5×10^{24}	24
30 mots	6×10^{42}	42
40 mots	3×10^{62}	62

On voit que la proportionnalité à log N ne constitue pas un mince avantage.

Si les graphes utilisés sont des arborescences projectives c'est-à-dire des arborescences hiérarchiques représentables par inclusion de parenthèses :

(((le) chien) aboie (souvent))

N est beaucoup moins élevé. Donnons un tableau comparatif:

N (parenthèses et projectivité)	N (arborescences quelconques)
$7\ 10^{3}$	10^{9}
$4\ 10^{13}$	$5\ 10^{24}$
$5\ 10^{21}$	$6\ 10^{42}$
$6\ 10^{29}$	$3\ 10^{62}$

Cette différence est très nettement à l'avantage des arborescences projectives et de la notation de parenthèses.

Les nombres des tableaux ci-dessus ont été calculés par Eric Morlet.

b/ Coût de chaque tâtonnement.

A chaque tâtonnement, il y a consultation du catalogue. Le temps d'accès à ce dernier constitue un élément important.

Pour savoir quelles règles consulter, il faut d'autre part énoncer les caractéristiques des groupes de mots assemblés par tâtonnement. On a examiné le coût de ces opérations en III, 3.

C/ CHAMP DE RENTABILITE D'UN FILTRE.

La première préoccupation que l'on doit avoir en entreprenant la construction d'un automate, ou sa simulation sur machine, c'est de définir sa fonction. Aussi rappellerons-nous (cf I, 3 et I,5) ce qu'il en est de celle des filtres. La fonction d'un filtre est de consulter automatiquement un certain catalogue ou sous-catalogue, bien déterminé, de règles grammaticales ou sémantiques, en vue de contribuer à lever les ambiguïtés de phrases du langage ordinaire, en éliminant les hypothèses qui contredisent une règle au moins du catalogue.

Comment caractériser un filtre dont on entreprend la construction ? Il serait absurde de fixer par avance de façon rigide et limitative le contenu du catalogue associé au futur filtre. En effet, une fois que l'on aura mis au point les mécanismes de consultation, on s'efforcera de les faire servir de la façon la plus large possible, c'est-à-dire d'ajouter des règles dans le catalogue., si bien que ce dernier doit être considéré comme un ensemble ouvert.

Aussi caractérisera-t-on un filtre, non pas par l'ensemble des règles que l'on pourrait à la rigueur entasser dans son catalogue sans changer le mécanisme de consultation, mais par un ensemble plus limité, celui des règles dont ces mécanismes peuvent assurer la consultation à un prix acceptable.

Lorsqu'une règle s'harmonise bien avec le mécanisme d'un filtre, elle s'écrit facilement et en peu de signes dans le catalogue de celui-ci. Lorsqu'il y a mauvaise adaptation, il faut de plus en plus de place pour exprimer la règle dans la notation du catalogue et le prix devient prohibitif, si bien que la possibilité d'inscrire cette règle demeure toute théorique. C'est donc par la définition d'un "champ de rentabilité maxima" que l'on peut le mieux préciser l'utilité attendue d'un filtre.

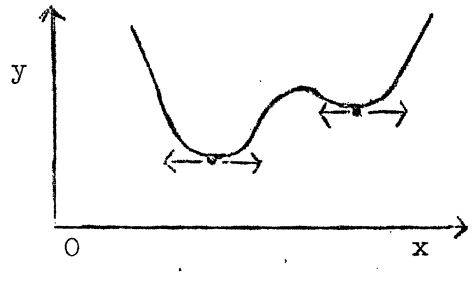
D/ RENTABILITE D'EMPLOI D'UNE FAMILLE DE GRAPHES DE REPERAGE.

Le choix d'une famille de graphes de repérage a des répercussions sur toutes les rubriques de coût énumérées plus haut. En Aa, par réduction du prix des catalogues, ce qui est d'ailleurs la raison essentielle que l'on a d'utiliser de tels graphes. En Ab (pour mémoire). En Ba, par l'intermédiaire du nombre N de graphes de la famille. En Bb, pour toutes sortes de raisons, et en particulier celles exposées au chapitre III, 3.

Comme les ordres de grandeur de ces diverses dépenses sont mal connus , on ne peut pas les ajouter . On peut difficilement apprécier si telle économie réalisée sur tel poste compense avantageusement une dépense corrélative effectuée sur un autre poste. Il faudrait pouvoir juger sur pièces et en fonction d'une machine.

D'où l'utilité d'expériences partielles sur ordinateur.

Cependant, on sait qu'une fonction varie peu au voisinage d'un extremum. Les graphes optimaux doivent donc rendre stationnaires le coût et la rentabilité :



Ce caractère stationnaire peut être éprouvé en faisant subir à la forme des graphes de petites modifications, et en examinant séparément les conséquences sur le coût des différents postes Aa, Ab, Ba et Bb.

Si nos évaluations sont exactes, la quasi totalité des graphes suggérés par des linguistes ont, chacun dans un certain domaine d'utilisation, un caractère stationnaire, ce qui est à l'honneur de la linguistique.

Il reste que seules des expériences partielles sur machines peuvent permettre d'évaluer convenablement les coûts et de choisir en connaissance de cause, pour chaque filtre, la famille de graphes de repérage qui convient le mieux.

E/ UTILISATION D'UN FILTRE LENT POUR LA CONSTRUCTION AUTOMATIQUE DE FILTRES PLUS RAPIDES

L'emploi de graphes de repérage permet une simplification des catalogues de règles. On peut alors construire ces catalogues plus rapidement et réaliser une économie très considérable de travail humain.

On sait qu'il faut payer en servitudes supplémentaires de calcul la rançon de ces avantages. L'augmentation du temps de consultation coûte des heures machine.

Jusqu'à quel point est-il possible de pousser la politique des catalogues bon marché à consultation relativement coûteuse ? S'il s'agissait de construire un programme définitif, destiné à tourner pendant des années, on ne pourrait accepter qu'il soit lent. La réduction du prix des catalogues se traduirait rapidement par une perte d'argent.

Mais le problème ne se pose pas du tout dans ces termes. Il est possible, à partir d'un filtre lent, de construire automatiquement les catalogues d'un filtre plus rapide, de sélectivité inférieure ou égale, catalogues volumineux et compliqués dont la construction coûterait, en travail humain, un prix exorbitant.

On peut donc accepter que les premiers programmes construits soient lents, puisqu'ils tourneront simplement le temps nécessaire pour la construction automatique de programmes équivalents mais plus rapides.

Cette manière de procéder s'apparenterait beaucoup à celle de l'apprentissage humain. Lorsqu'un enfant commence à écrire, ses premiers "programmes" sont du type "programmes de tâtonnements" ; ils sont contrôlés par l'observation des résultats des gestes. Bien que très lente et maladroite, cette procédure est la plus favorable à l'apprentissage, car elle minimise le volume du catalogue de renseignements à transmettre du professeur à l'élève. Le professeur n'a pas besoin de décrire des contractions complexes de muscles de la main, ni les commandes nerveuses qui les provoquent. Il décrit simplement le résultat : la forme des lettres.

Mais une fois que les gestes lents de construction de lettres sont acquis, une fois que le premier "programme" existe, il se transforme en d'autres "programmes" beaucoup plus rapides, faits de gestes réflexes, avec une part de tâtonnement très réduite.

Ainsi, on peut accepter qu'un premier programme, dit "d'apprentissage", soit relativement lent, si cet inconvénient permet, en échange, une réduction du prix des transferts d'information du professeur à l'élève, c'est à dire une réduction du prix des catalogues de règles. On peut accepter cela, à condition : d'une part que l'on ait le moyen de construire automatiquement, à partir de filtres lents, des filtres rapides (qui eux-mêmes permettront de construire automatiquement des filtres encore plus rapides); d'autre part, que la lenteur du filtre d'apprentissage reste dans des limites raisonnables, pour que le coût de son utilisation temporaire demeure acceptable.

QUATRIEME PARTIE

CONSTRUCTION D'UN FILTRE PARTICULIER A TITRE D'EXEMPLE

IV, 1 PROBLEME DU CHOIX D'UNE FAMILLE DE GRAPHES, EN VUE DE LA CONSTRUCTION D'UN FILTRE PARTICULIER

Toute la quatrième partie sera consacrée au détail de la construction d'un filtre particulier, l'idée étant de montrer par un exemple comment on peut utiliser les méthodes décrites jusqu'ici, et comment on peut essayer de minimiser les coûts.

On se servira bien entendu de graphes de repérage (dont la forme est donc à choisir), et de la méthode par assemblages progressifs décrite en III, 2.

Conformément à ce qui a été dit en III, 4, C, nous commençons par définir un "champ de rentabilité maxima". Rappelons qu'il s'agit d'un ensemble partiel de règles grammaticales ou sémantiques, comprenant celles que le filtre sera appelé à consulter le plus souvent, et qu'il doit être capable de consulter et d'exploiter au meilleur prix.

Il est normal de souhaiter que ce champ recouvre en priorité les règles dont l'usage est le plus courant, à savoir les règles grammaticales, plus un certain nombre de règles sémantiques choisies parmi celles qui interviennent le plus fréquemment dans le discours.

Cependant il est clair que les lois grammaticales sont de loin les premières du point de vue de la fréquence d'utilisation. Une règle telle que l'accord article-nom peut intervenir un très grand nombre de fois dans une phrase un peu longue. C'est donc pour les

règles grammaticales que l'on s'efforcera le plus de réduire les coûts de consultation.

Pour être précis, il faudrait même introduire parmi elles une pondération en importances, de manière à tenir compte de la fréquence moyenne avec laquelle on effectue la dépense de consultation de chaque règle; mais nous ne disposons pas, pour le moment, de statistiques détaillées sur ces fréquences. Pour commencer, d'ailleurs, et prendre les premières décisions, il suffit d'une appréciation grossière de ces fréquences moyennes.

Le problème qui se pose maintenant est celui du choix d'un type de graphes de repérage. Comme on l'a vu en III, 4, D, les caractéristiques de la famille de graphes de repérage utilisée dans un filtre à assemblage progressif ont une influence déterminante sur les performances de ce filtre, ainsi que sur sa rentabilité.

Comme on l'a exposé plus haut en II, 1, les graphes de repérage sont de purs instruments de calcul, sans prétention linguistique. Les grammairiens, sémanticiens et linguistes, (ou en tous cas ceux d'entre eux qui refusent de tenir compte des prix de revient) n'ont aucune qualité pour nous imposer ce choix. Mais en fait, l'on a le plus grand intérêt à tenir compte de leurs travaux et des observations qu'ils ont effectuées concernant les propriétés du langage.

C'est en effet une vérité constante, en matière de repérage, que les repères optimaux du point de vue de la rentabilité reflètent au moins une partie des propriétés de la chose repérée.

Le passé de la science du repèrage en matière extra-linguistique en fournit des exemples répétés. En géométrie: les meilleurs axes de coordonnées pour une hyperbole équilatère sont soit des droites coincident avec les asymptotes, soit des droites coïncident avec les axes de symétrie. Dans les deux cas, que l'on ait choisi l'un ou l'autre de ces modes de repérage optimaux, on est conduit à faire coincider le centre de coordonnées avec le centre de symétrie de l'hyperbole, si bien que le repère reflète au moins les propriétés de celle-ci.

De ce fait, que les repères optimaux reflètent généralement une partie des propriétés de la chose repérée, la géographie donne un autre exemple, par les procédés qu'elle utilise pour le repérage d'un point sur la surface terrestre. Pour ce repérage on se sert habituellement d'une surface abstraite en forme de sphère ou d'ellipsoïde. Cette surface porte les méridiens et les parallèles. Elle n'est pas, comme la surface terrestre, hérissée de collines, ni plissée en montagnes et vallées. Elle n'a donc pas vraiment la forme de la terre. Mais, comme surface de référence, elle présente un caractère optimal indiscutable. Il est simple de repérer à partir d'elle la hauteur des montagnes et la profondeur des océans.

Si l'on faisait un recensement des points pour lesquels la surface terrestre coïncide avec celle de la surface fictive de référence, on en trouverait sans doute relativement peu. Cependant,

personne ne s'aviserait de dire que la forme de la surface de référence est sans rapport avec celle de la terre. Bien au contraire, lorsqu'il faut décrire la seconde en quelques mots, on donne la forme de la première.

Aussi, pour en revenir au problème du choix d'une famille de graphes de repérage en vue de la construction d'un filtre d'analyse automatique du langage naturel, nous pensons que si nous avions toutes les données et tous les moyens nécessaires pour guider entièrement ce choix par des considérations de prix, de commodité d'utilisation concernant telle ou telle langue naturelle, d'économie de temps machine et de temps de compilation de catalogue, et que nous parvenions à un optimum de rentabilité, alors, il existerait certainement un rapport étroit entre les propriétés de la famille de graphes choisie et les propriétés de la ou des langues naturelles pour lesquelles cette famille se révèlerait optimale.

Comme nous n'avons, pour l'instant, pas fait suffisamment de statistiques et d'expériences sur ordinateur pour être à même de calculer des rentabilités en valeur absolue, comme nous avons seulement les moyens, décrits en III, 4, D, d'apprécier le cractère stationnaire de la rentabilité d'une famille de graphes donnée, et de constater, s'il y a lieu, qu'elle représente un optimum relatif, le mieux que nous puissions faire est de nous guider sur les propriétés du langage, c'est à dire sur les travaux des linguistes, pour proposer des formes de graphes. Les calculs de rentabilité interviendront ensuite pour accepter ces formes ou bien les refuser, et même éventuellement pour les modifier légèrement.

A défaut, en effet, de pouvoir mesurer la rentabilité en valeur absolue, nous sommes à même d'apprécier si elle augmente, reste stationnaire, ou diminue, lorsque l'on fait subir de petites modifications à la forme des graphes (voir en II, 4, D).

Ainsi, nous sommes conduits à étudier les formes proposées par les linguistes.

IV, 2 GRAPHES LINGUISTIQUES SUGGERES PAR LA SYNTAXE DE L'ECOLE DE COPENHAGUE.

Les linguistes de l'école de Copenhague n'ont pas, à notre connaissance, proposé de graphes. Mais un graphe n'est jamais que la représentation de certaines relations, et c'est la donnée de ces dernières qui compte. On peut alors remarquer que, par leur caractère abstrait et général, les trois"fonctions" de la glossématique définies par L. Hjelmslev (16) se prêtent bien à l'emploi de représentations graphiques schématiques. Il en est de même de la procédure de découpage, ou "division", proposée par cet auteur.

Dans son livre : "structure immanente de la langue française", Knud Togeby (7) a montré comment l'on devait appliquer, dans l'analyse d'une langue naturelle, les principes de l'école de Copenhague.

Prenons l'exemple d'une phrase simple: "Le vieux meunier siffle son chien" et découpons-la en fonction des indications données par Togeby. Nous empruntons à la partie "syntaxe" du livre cité plus haut (7) les relations suivantes, utiles pour cette phrase :

Proposition = sujet + prédicat

Prédicat = objet + verbe

Groupe nominal = article + membre nominal

Membre nominal = épithète + substantif.

De plus, Togeby signale :

- que l'épithète est subordonnée au substantif,
- que l'article est subordonné au membre nominal,
- que, selon certains auteurs du moins, l'objet est subordonné au verbe,
- que, dans une certaine perspective, on peut considérer le sujet comme subordonné au verbe.

Ces quelques lignes rendent assez mal compte de l'esprit de l'oeuvre de Togeby, et des jugements nuancés qu'elle contient. Knud Togeby a l'art et le mérite de bien savoir montrer combien toute opinion que l'on peut avoir en matière de "subordination" ou de "supériorité" hiérarchique entre éléments des phrases, est discutable.

En même temps qu'il indique son opinion et les raisons qui la justifient à ses yeux, il prend soin de citer en détail les auteurs d'opinion contraire, et de noter que ces auteurs ont parfois aussi de très bonnes raisons, pour aboutir cependant au résultat inverse du sien quant à la subordination de tel groupe à tel autre.

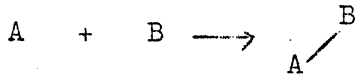
Lorsque les arguments pour ou contre se révèlent également forts (A subordonné à B, et simultanément B subordonné à A), le rapport n'est plus de simple subordination (ou de "sélection", dans la terminologie glossématique), mais de double et réciproque subordination ("solidarité" dans la même terminologie). Si aucun rapport de subordination ne peut être décelé entre deux termes résultant d'un découpage, on dit qu'il y a entre eux "combinaison".

Telles sont - pour autant que l'on puisse les évoquer en aussi peu de mots - les idées développées dans la syntaxe du livre: "Structure immanente de la Langue française" (7)

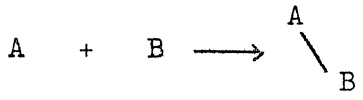
Voyons le parti que l'on peut en tirer lors du choix de graphes d'adressage.

Reprenons pour cela les quelques règles de découpage données plus haut à propos d'un exemple.

Effectuons maintenant l'opération inverse du découpage, en notant les relations d'assemblage comme suit :

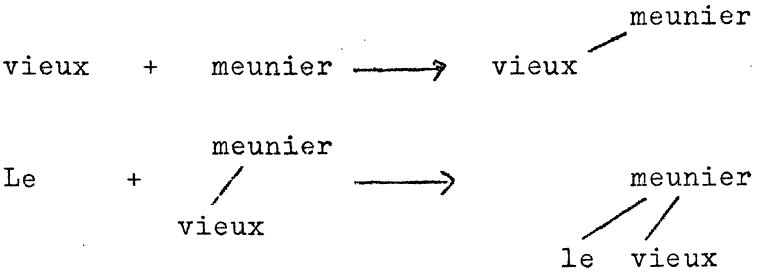


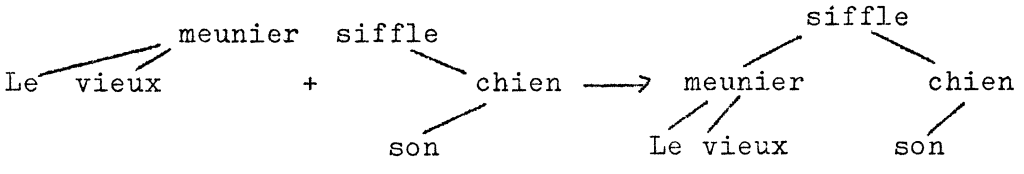
ou bien



selon que A est subordonné ou supérieur à B.

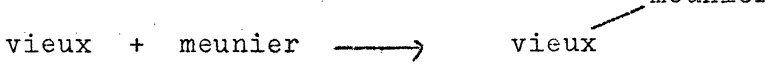
Il vient :





Ainsi, on peut reconstruire, à partir des données de Togeby, des graphes qui se révèlent très semblables à ceux de Harper et Hays ou de Tesnière. Il s'agit d'arborescences projectives, représentables en notation de parenthèses.

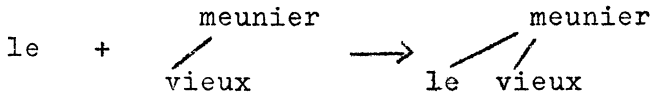
Ainsi :



peut s'écrire :

(vieux) + (meunier) ⟶ ((vieux) meunier)

De même :



s'écrit :

(le) + ((vieux)meunier) ⟶ ((le)(vieux)meunier)

et d'une manière générale chaque opération d'assemblage entre deux éléments A et B peut se représenter par un "conflit" du type :

(A) + (B) ⟶ ((A)B)

ou

(A) + (B) ⟶ (A(B))

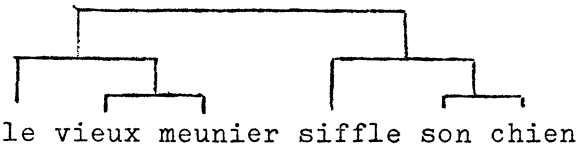
selon que l'un ou l'autre est déclaré subordonné.

Faisons maintenant abstraction de tout ce qui concerne les rapports hiérarchiques, et considérons simplement les opérations d'assemblage.

vieux	+	meunier	⟶	vieux meunier
le	+	vieux meunier	⟶	le vieux meunier
son	+	chien	⟶	son chien
siffle	+	son chien	⟶	siffle son chien

$$\left\{\begin{array}{c}\text{le vieux}\\ \text{meunier}\end{array}\right\} + \left\{\begin{array}{c}\text{siffle}\\ \text{son chien}\end{array}\right\} \longrightarrow \left\{\begin{array}{c}\text{le vieux meunier}\\ \text{siffle son chien}\end{array}\right\}$$

Ces opérations suggèrent le graphe :

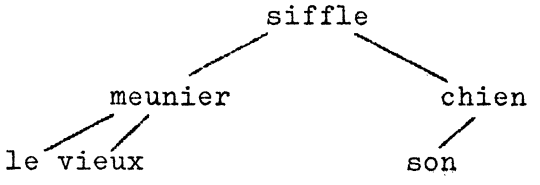


où chaque trait représente une opération. Les numéros de traits indiquent l'ordre des opérations.

Nous pouvons maintenant faire un bilan de tout ce qui vient d'être suggéré en matière de représentation graphique.

a/ En déformant un peu, comme il est normal lorsqu'on part de la linguistique pour aboutir à des graphes de repérage, la

pensée de Togeby, par négligence systématique des rapports autres que ceux de sélection, par schématisation et supression de certaines nuances, nous avons obtenu des arborescences voisines de celles de Tesnière, Harper et Hays, très propres à structurer l'intérieur des groupes de mots, ce qui constituait l'un des besoins definis au III, 3.



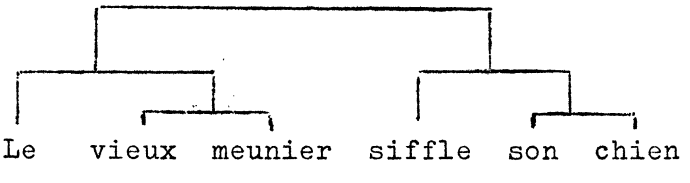
(Que Togeby nous pardonne, la vérité linguistique n'est pas en cause dans les problèmes de repérage, et nous avons signalé la chose au II, 1, en donnant la définition des graphes de repérage en général. Ces petits écarts par rapport à la pensée de Togeby, deviendront en machine de grosses économies).

Ces arborescences ont en outre l'avantage de se prêter à une représentation parenthétique linéaire,

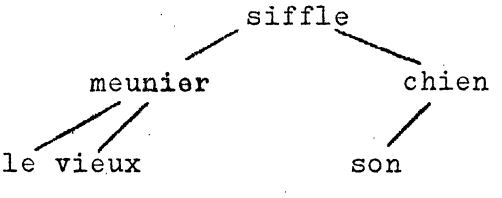
(((le) (vieux)meunier) siffle ((son) chien))

donc très favorable au traitement en machine. Sur ce point particulier, les formes suggérées par Togeby se révèlent nettement plus perfectionnées que celles de Tesnière.

b/ En refaisant selon une certaine succession (assemblage) des opérations que Togeby effectue selon une succession exactement inverse (découpage), nous obtenons de plus ce dont le III,2 indiquait la nécessité et qui manquait au graphe de Tesnière: l'indication d'un ordre d'assemblage au cours du temps. Cet ordre peut être représenté par un graphe analogue à celui utilisé par Bar Hillel



Du fait que chaque opération d'assemblage aboutit à créer une arête de l'arborescence, leur ordre peut être indiqué directement sur celle-ci, en numérotant les arêtes :

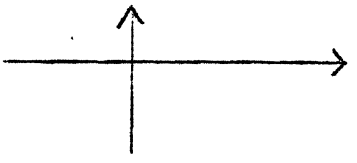


ou bien dans la représentation à parenthèses, sur le mot inférieur de chaque arête :

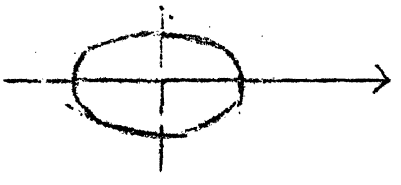
```
      2     1      5              3    4
(((le)(vieux)meunier)siffle((son)chien))
```

Mais dans la pratique, cet ordre n'est pas noté du tout, car on s'arrange pour qu'il soit lié à la forme de l'arborescence, ou bien imposé par les catalogues de règles.

Pour conclure, rappelons qu'il y a entre notre schématisation et la syntaxe de Togeby, une différence de nature comparable à celle qui oppose une paire d'axes de coordonnées



à la figure, par exemple une ellipse, qu'ils permettent de repérer



Les règles syntaxiques ne seront pas sacrifiées pour autant, mais on les rassemblera, dans la mémoire machine, à leur place normale, qui est le catalogue.

Il convient maintenant de voir si les graphes que nous venons de dessiner ont un caractère optimal.

IV, 3 LE CARACTERE D'ARBORESCENCE PROJECTIVE CORRESPOND AU MOINS A UN OPTIMUM RELATIF DE LA FONCTION DE RENTABILITE.

Les formes proposées en IV, 2, après examen de la syntaxe de l'école de Copenhague, ont, entre autres particularités, celles d'être des "arborescences projectives".

Nous commencerons par examiner en quoi consiste cette propriété. Nous verrons ensuite (ainsi qu'il a été dit en III, 4, D), de quelle façon varient les coûts lorsqu'on lui fait subir des entorses, avec de petites modifications des formes de graphes.

A/ LES ARBORESCENCES PROJECTIVES

Rappelons que selon notre terminologie, une arborescence mérite la qualification de "projective" si, et seulement si, il est possible de la représenter en notation de parenthèses [exemple du IV, 2 : (((Le)(vieux)meunier) siffle ((son)chien)).]

Toutes les arborescences n'ont pas cette propriété, bien au contraire. Soit N_1 le nombre d'arborescences quelconques différentes que l'on peut construire avec m mots, c'est-à-dire m sommets. Soit N_2 le nombre d'arborescences projectives distinctes que l'on peut construire avec m mots. Donnons quelques valeurs de ces nombres :

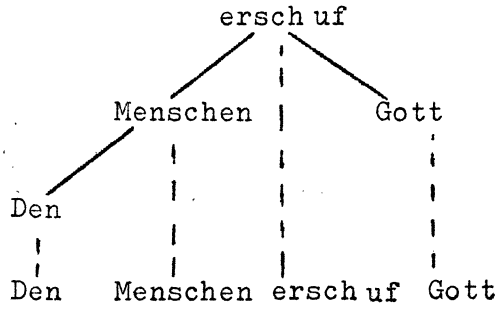
mots m	arb. quelconques N_1	arb. projectives N_2
10	10^9	$7.\ 10^5$
20	$5\ .\ 10^{24}$	$4.\ 10^{13}$
30	$6\ .\ 10^{42}$	$5.\ 10^{21}$
40	$3\ .\ 10^{62}$	$6.\ 10^{29}$

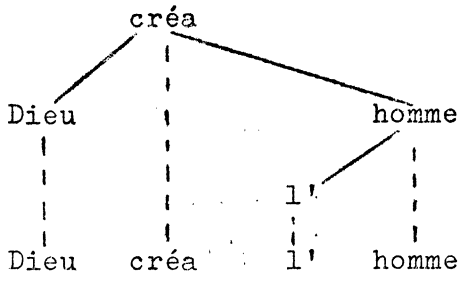
Ce tableau avait déjà été donné en III, 4, B. On avait montré alors que le nombre N de graphes de repérage de la famille influe sur la durée de l'assemblage, car l'un des facteurs multiplicatifs de cette durée, et donc du coût, croît quand N croît. La réduction considérable de ce nombre lorsque l'on passe de N_1 à N_2, c'est-à-dire d'arborescences quelconques à des arborescences projectives, matérialise l'un des bénéfices apportés par la projectivité.

Un second avantage de la projectivité tient au caractère linéaire de la notation de parenthèses qu'elle permet d'utiliser en machine.

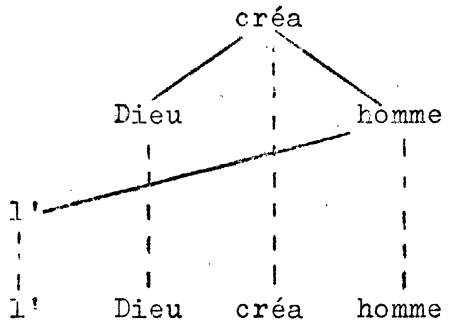
Ceci dit, la notation de parenthèses n'est guère parlante pour l'oeil, et il est plus commode de raisonner sur des arborescences à condition de pouvoir vérifier facilement si elles sont projectives. Il existe des critères de projectivité assez simples à l'emploi.

On peut montrer, par exemple, qu'une arborescence connexe est projective si, et seulement si, elle peut se projeter suivant l'ordre linéaire de la phrase de telle façon que la figure de projection ainsi constituée ne comporte ni croisement entre arcs, ni croisement entre arcs et projetantes :





Si la définition des graphes de repérage, en fonction des natures des mots et de leurs rapports, est convenablement donnée, il n'apparaîtra jamais de croisements, sauf pour des phrases mal construites, incorrectes, comme le serait par exemple une phrase à qui on aurait fait subir une permutation dans l'ordre d'énonciation des termes.



Ainsi, en introduisant une condition de non croisement, la projectivité crée un lien entre les adresses selon l'arborescence et les adresses selon l'ordre linéaire d'énonciation des mots.

Il est normal que cela facilite les changements de repère, et en particulier l'assemblage, puisque, au sens du III, 1, ce mot désigne la procédure de mise en correspondance des adresses selon le graphe avec les adresses selon l'ordre linéaire.

B/ CARACTERE D'OPTIMUM RELATIF

L'équivalent d'une petite variation influant peu sur le coût, serait la décision qu'une phrase, ou bien qu'un nombre assez limité de phrases, ne doivent pas être repérées selon des graphes en arborescences projectives. Si la liste des exceptions est donnée, il en coûte assez peu de les orienter vers un sous programme spécial.

Mais si les entorses au caractère d'arborescence projective se produisent au niveau des constructions, c'est-à-dire de façon à la fois imprévisible et systématique, tout le bénéfice de la projectivité est perdu. Pour une phrase de 40 mots N passe de 10^{29} à 10^{82}, ou même plus si le caractère arborescent disparaît lui aussi, et le coût augmente très brutalement.

IV, 4 L'ASSEMBLAGE : LES METHODES A CONFLITS DE PARENTHESES.

Les méthodes à conflits de parenthèses (ou, tout court, méthodes de conflits) ne sont que des sous variétés particulièrement rapides des méthodes par assemblage progressif définies en III,2.

Elles impliquent l'emploi de la notation parenthetique décrite plus haut, exemple :

(((Den)Menschen) erschu f (Gott))

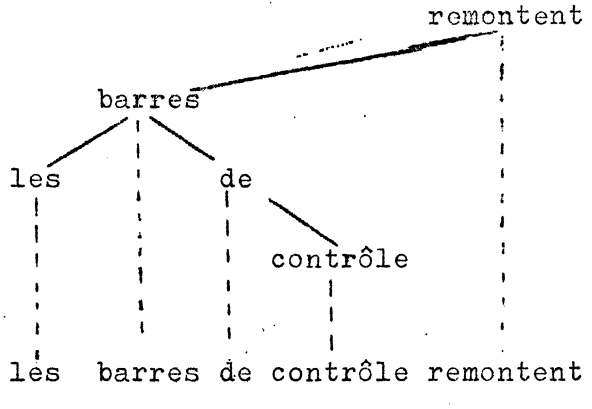
De ce fait, les méthodes à conflits de parenthèses ne peuvent être utilisées que dans les cas où les graphes de repérage sont des arborescences projectives.

A/ LE "PROBLEME DE L'ASSEMBLAGE" SE RAMENE A CELUI DE LA CONSTRUCTION D'UNE FIGURE DE PROJECTION

Rappelons le but des procédures d'assemblage en général (III, 1) : calculer automatiquement les adresses selon le graphe de repérage correspondant aux adresses linéaires des mots (ou groupes de mots) dans la phrase donnée.

Rappelons que cette mise en correspondance du graphe d'adressage et du repère linéaire est indispensable en vue de la consultation des catalogues de règles. En effet, pour réduire le prix de ces catalogues, on les a rédigés en y exprimant les adresses en fonction du graphe de repérage. Par contre, les hypothèses formulées par le dictionnaire automatique sur les natures des mots à homonymies sont exprimées en adresses selon l'ordre linéaire. Pour voir si ces hypothèses sont conformes à la grammaire et à la sémantique des catalogues, il faut les confronter avec les catalogues, donc mettre en correspondance les adresses linéaires et les adresses selon le graphe. D'où la nécessité du calcul dit d'assemblage.

Lorsque les graphes de repérage sont des arborescences projectives, les résultats du calcul, c'est-à-dire cette correspondance, pourra être matérialisée par la "figure de projection" (IV, 3, 4). Le problème de l'assemblage se ramène donc à celui de la construction de cette figure. Exemple :



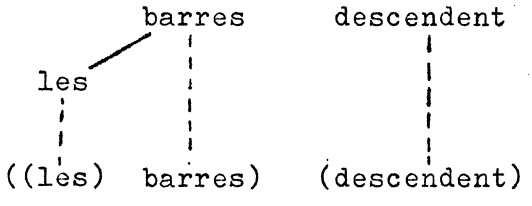
On sait (III, 1) que cette figure doit être construite par tâtonnement, en choisissant, parmi tous les tracés possibles, ceux qui ne contredisent aucune règle du catalogue. Il peut y avoir plusieurs solutions.

On se rappelle enfin le principe des méthodes à assemblage progressif (cf III, 2). Dans le cas présent, il s'exprime comme suit: la machine doit construire trait par trait la figure de projection, en vérifiant chaque fois que le trait nouvellement tracé ne contredit aucune règle du catalogue.

B/ OPERATION ELEMENTAIRE PERMETTANT DE CONSTRUIRE CHAQUE TRAIT DU GRAPHE DE REPERAGE

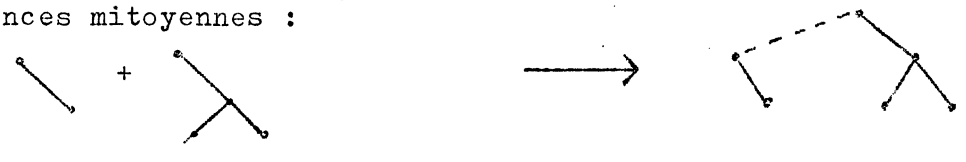
La figure de projection sera donc construite morceau par morceau, et en assemblant des morceaux par des traits.

Comme on veut utiliser la notation de parenthèses pour représenter tous les états intermédiaires de cette construction, cela impose que les morceaux soient toujours des arborescences projectives. Un mot tout seul n'est qu'une arborescence projective particulière. Donnons un exemple d'état intermédiaire pour la phrase: "les barres descendent".

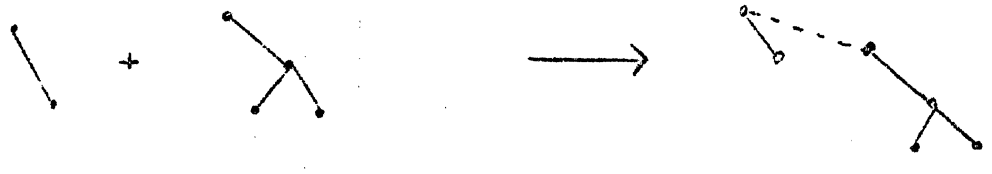


On désire en outre que le passage d'un stade intermédiaire à un autre ne se fasse jamais autrement que par déplacement d'une parenthèse, et une seule.

De cette exigence, il résulte que la seule manière permise d'ajouter un trait à la figure, consiste à réunir les têtes de deux arborescences mitoyennes :



ou bien :



Des deux têtes des arborescences composantes, l'une devient tête de l'arborescence résultant de l'assemblage, et l'autre lui devient subordonnée. D'où la terminologie des "conflits" : on dit que lors du tracé du trait reliant leurs deux têtes, les arborescences composantes entrent en conflit. Le conflit comporte, lorsque l'assemblage a lieu, un gagnant et un perdant. L'arborescence "gagnante" fournit la tête de l'arborescence résultante.

Avant de relier par un trait les têtes de deux arborescences mitoyennes, la machine regarde si le catalogue permet explicitement cet assemblage. Tout assemblage dont le catalogue ne parle pas est défendu.

Le catalogue peut aussi contenir des interdictions explicites. Lorsque l'assemblage n'est pas permis, il n'y a ni gagnant ni perdant, et on ne trace aucun trait.

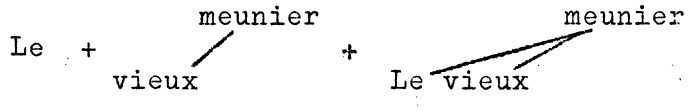
Lorsque le catalogue autorise un assemblage, il doit indiquer quel est le "gagnant". L'arborescence résultante se trouve alors bien déterminée.

On remarquera que les assemblages suggérés par Togeby, et dont nous avons donné des échantillons en IV, 2, ont toutes les caractéristiques des conflits définis ci-dessus. ces procédures de conflits apparaissent ainsi comme directement inspirées par la linguistique de l'école de Copenhague.

C/ LES CONFLITS DE PARENTHESES

Tout ce qui vient d'être dit en terminologie d'arborescences projectives peut être retranscrit, sans rien changer, en notation de parenthèses.

L'opération élémentaire de conflit :



s'exprime avec un seul déplacement de parenthèse

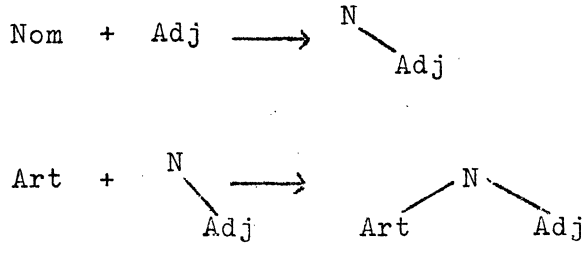
$$(\text{Le}) + ((\text{vieux})\ \text{meunier}) \longrightarrow ((\text{Le})\ (\text{vieux})\ \text{meunier})$$

la parenthèse extrême du "gagnant" venant annexer le "perdant".

Un ensemble de mots et de parenthèses contient une information équivalente à celle d'une figure de projection.

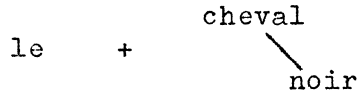
D/ LE CATALOGUE DE REGLES

Il contient des autorisations d'assembler du genre de :

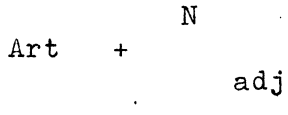


etc.

Soit par exemple à arbitrer le conflit :



le dictionnaire indique (au moins à titre d'hypothèse dans une alternative, pour le mot "le"), que "le" est un article, "cheval" un nom, "noir" un adjectif. Le conflit s'écrit donc :

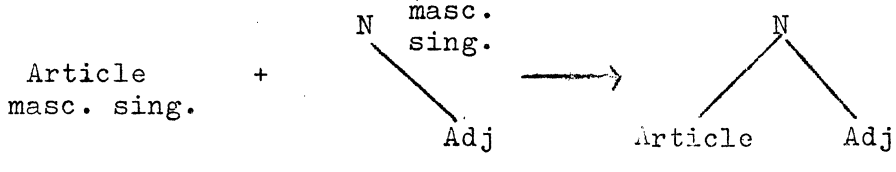


et la règle citée plus haut comme exemple permet de l'arbitrer.

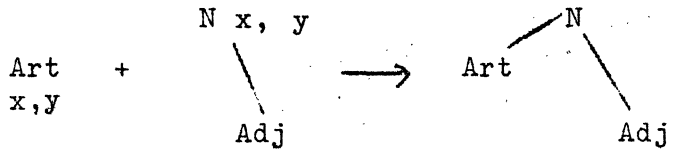
Donnons quelques indications générales sur l'organisation du catalogue :

a/ Il y a intérêt à adopter les conventions suivantes : l'autorisation de ne pas assembler peut toujours être considérée comme sous-entendue. L'autorisation d'assembler n'est jamais sous entendue, on doit la formuler explicitement dans le catalogue.

b/ Les catégories de mots doivent être décrites avec l'indication de toutes leurs caractéristiques pertinentes : exemple :

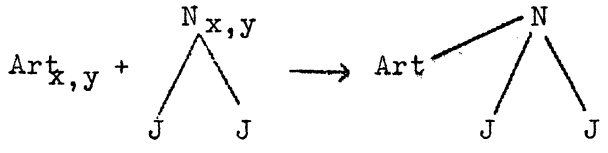


c/ Ces caractéristiques, fort nombreuses, augmenteraient beaucoup le nombre de règles si l'on ne recourait à des notations algébriques :



x est le genre, y le nombre, et cette notation indique que Art et N ont même genre et même nombre.

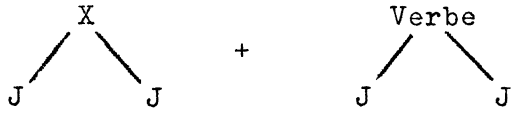
d/ Les caractères non pertinents ne font pas l'objet de vérifications. Pour cela on utilise divers procédés, dont les "Jockers" ne sont qu'un exemple :



si d_{lg} (N) ne contient pas d'article.

Cette règle s'interprète comme suit: les jockers représentent n'importe quoi, c'est à dire un nombre quelconque d'arborescences quelconques. La condition signifie que la dérivée première à gauche du N ne contient pas d'article. Par dérivée première à gauche d'un mot, on désigne l'ensemble des subordonnés directs de ce mot, situés à gauche de la projetante de ce mot.

e/ Donnons un exemple de règle un peu complexe. Prenons la règle sujet-verbe dans le cas où le sujet est à gauche du verbe. Elle s'écrit :



Conditions 1° X

1		y	z	s	...

(ce qui signifie ce qui suit : la tête de l'arborescence de gauche est un mot de catégorie non précisée x, de nombre non précisé y, de personne non précisée z, ayant le caractère sémantique non précisé s, mais ce mot doit être au nominatif, comme l'indique le 1)

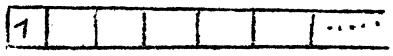
2° Verbe

		y	z	s'	...

Les y semblables figurent l'accord en nombre, les z l'accord en personne.

3° accord sémantique entre s (catégorie d'agent) et s' (catégorie d'action)

4° Les dérivées premières de V ne doivent pas contenir de



(ce qui signifie qu'on ne doit pas avoir déjà trouvé un sujet).

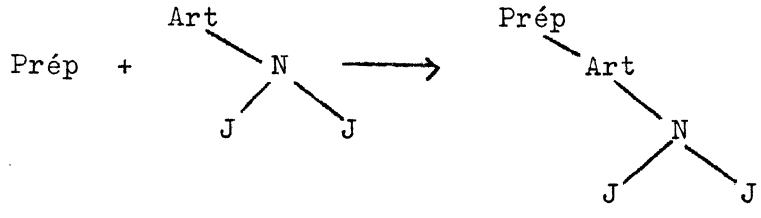
E/ DEFINITION DETAILLEE DU TRACE DES GRAPHES DE REPERAGE

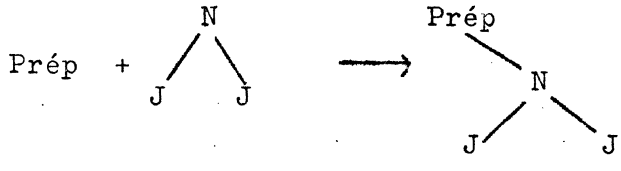
C'est en rédigeant le catalogue que l'on fixe définitivement le tracé des graphes de repérage. Pour ces choix on fera intervenir les divers éléments de coût cités en III,4.

Il est clair par exemple que des deux tracés

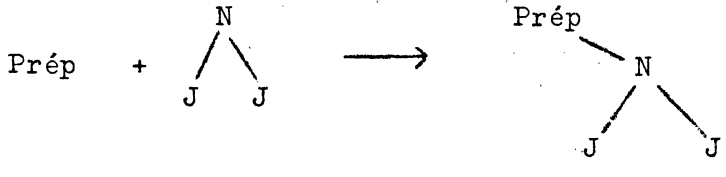
aucun ne se heurte à une impossibilité, mais que le second augmente légèrement le volume du catalogue.

Certains assemblages, qui exigent l'énoncé de deux règles lorsque l'on emploie le second tracé :

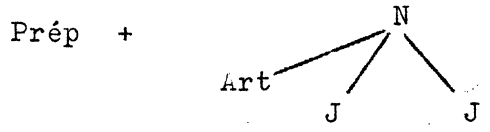




s'expriment avec une seule règle lorsque l'on emploie le premier tracé. En effet, comme le jocker peut remplacer n'importe quoi, la règle



inclut le cas particulier



si bien que l'on n'a pas besoin de mentionner ce cas, et que l'on fait ainsi l'économie d'une règle.

Tous les éléments du coût doivent être pris en considération. Les formes qui se révèlent les plus rentables sont souvent celles proposées par les linguistes.

IV, 5 CONSTRUCTION D'UN FILTRE PARTICULIER. MISE EN PLACE DES MECANISMES D'ASSEMBLAGE.

Même dans le cadre restreint des méthodes de conflits, il reste une très large latitude de décision, au moment où l'on met en place les mécanismes d'assemblage. Ceux que nous nous proposons de décrire comportent comme organes principaux :

- un sous programme "de cascade de conflits en régression". Nous l'appellerons aussi sous programme A;
- un sous programme "de déclenchement des cascades de conflits". Nous l'appellerons aussi sous programme B

A/ CASCADE DE CONFLITS EN REGRESSION

La notion de conflit a été définie et décrite avec beaucoup de détails en IV, 4. Rappelons simplement que chaque opération de conflit comporte déplacement d'une parenthèse

(X) (Y)

↓

(X (Y))

ou bien

(X) (Y)

↓

((X) Y)

après consultation d'un catalogue de règles.

X et Y peuvent être des expressions entières de mots et parenthèses

((...) m_1 (...)) ((...) m_2 (...))

↓

((...) m_1 (...) ((...) m_2 (...)))

Remarquons qu'en machine on peut se contenter d'indiquer les nombres de parenthèses de chaque sorte situées entre deux mots :

((le) bout (de((la) rue)))

devient

0,2 le 1,0 bout 0,1 de 0,2 la 1,0 rue 3,0

Les déplacements de parenthèses sont représentés par des variations correspondantes de ces nombres.

Donnons un exemple de cascade de conflits de parenthèses.

a/ état initial: un ensemble de mots et de parenthèses

(le) (bout) (de) (la) (rue) D

la lettre D indique le point de déclanchement de la cascade de conflits.

b/ blocage de toutes les parenthèses

Convenons, lorsque des parenthèses sont bloquées, de l'indiquer en dessinant autour d'elles des machoires. Cela donne ici :

[(le) (bout) (de) (la) (rue)] D

c/ libération progressive des parenthèses, de droite à gauche, avec conflits

[(le) (bout) (de) (la) (rue)] D

[(le) (bout) (de) (la)], (rue) D

[(le) (bout) (de)], (la) (rue) D

[(le) (bout) (de)] (la) rue) D

[(le) (bout)], (de) ((la) rue) D

[(le) (bout)] (de ((la) rue)) D

[(le)] (bout) (de ((la) rue)) D

[(le)] (bout (de ((la) rue))) D

⌐(le) (bout (de ((la) rue))) D

⌐((le) bout (de ((la) rue))) D

Le sous programme A peut être défini comme effectuant les opérations suivantes :

- prendre dans une mémoire réservoir RE un ensemble de mots et de parenthèses, muni d'une lettre D au point où la cascade doit être déclenchée. Transporter cet ensemble dans une "mémoire de travail" pour le traiter.

- effectuer, vers la gauche et à partir du point D de déclenchement, les opérations de cascade de conflits décrites pour l'exemple ci-dessus.

- le sous programme A prend fin lorsque, pour un conflit, le catalogue ordonne le statu quo.

B/ SOUS PROGRAMME DE DECLENCHEMENT

Le sous programme B a pour fonction :

- de "photographier" tous les états successifs distincts de la "mémoire de travail" du sous programme A (c'est-à-dire les états d'un ensemble de mots et de parenthèses s'y trouvant).

- effectuer sur chaque "photographie" diverses vérifications (test de fin de phrase, test de solution, etc.) Dans le cas banal, modifier l'expression en déplaçant le signe D d'un mot vers la droite, puis envoyer cette expression ainsi modifiée dans la mémoire réservoir RE.

ANNEXE

DISPOSITIF EXPERIMENTAL

A/ OBJET DE L'EXPERIENCE.

a/ Essai sur ordinateur du sous programme de cascades de conflits en régression, défini au IV, 5, A.

Ce sous programme constitue le mécanisme de base du filtre défini en IV, 1; IV, 2; IV, 3; IV, 4 et IV, 5. La notion de filtre a été définie en I, 5. Les problèmes de filtrage, en I, 2; I, 3 et I, 4. L'agencement des filtres, en I, 5. Le champ de rentabilité du filtre à conflits, en IV, 1.

Ce sous programme de cascades de conflits en régression met en oeuvre des graphes de repérage. On a exposé en II, 1; II, 2 ; II,3 et II, 4 les simplifications que l'emploi de ces graphes apporte lors de la construction de catalogues grammaticaux et sémantiques consultables sur ordinateur. On a décrit en III, 1, III, 2 et III, 3, les servitudes de calcul qui constituent la rançon de ces avantages.

Les opérations effectuées par ce sous programme de cascades de conflits en régression ont été analysées en IV, 5, A.

b/ Passage automatique de la notation de parenthèses à celle d'arborescences projectives.

Le principe du sous-programme qui effectue cette transformation a été donné au 5 des travaux pratiques de la journée de linguistique. Pour plus de détails sur les arborescences projectives, voir ci-dessus en IV, 3.

B/ ORGANIGRAMMES DE L'EXPERIENCE.

Ils seront donnés plus loin sous la signature d'Eric Morlet.

C/ ENTREES ET SORTIES D'INFORMATIONS.

Dans toute expérience partielle visant à tester des sous programmes, il se pose des problèmes anormaux d'entrées et de sorties d'informations. On doit alimenter les sous programmes dans des conditions comparables à celles de leur fonctionnement comme organes d'un ensemble complet.

Les cascades de conflits supposent une triple alimentation: d'une part, informations normatives (règles); d'autre part,

informations lexicales (nature des mots); enfin, phrases à traiter.

Employées seules, les cascades de conflits ne permettent pas de résoudre des homonymies. Il faut au moins, en même temps, le sous programme de déclenchements (IV, 5, B), il faut un organe proposant les diverses acceptions des mots (I, 5), il faut de préférence, en plus, d'autres filtres. N'ayant que les cascades de conflits en machine, nous avons préédité les phrases d'entrée, en indiquant les natures grammaticales correspondant aux acceptions qui convenaient effectivement pour les mots. Enfin, on a introduit en mémoire une grammaire très grossière et peu sélective, mais suffisante pour les phrases d'entrée. Parmi celles-ci, il y a surtout des phrases en français, plus quelques phrases en allemand, italien, néerlandais.

Les ensembles de mots et de parenthèses obtenus à la sortie du sous programme de cascades de conflits, servent d'alimentation d'entrée au second sous programme, qui les transformera en arborescences.

D/ CARACTERE MULTILINGUE DE L'EXPERIENCE.

Les procédures de conflit ne sont rien d'autre qu'un calcul de changement de repère. C'est pour souligner le caractère extrêmement général et abstrait du procédé, que nous avons introduit des phrases en plusieurs langues naturelles.

E/ CARACTERE PARTIEL DE L'EXPERIENCE.

Le matériel dont nous disposons ne permet pas mieux que des expériences partielles. La mémoire de l'IBM 650 normale est limitée à 2.000 mots. Il est hors de question, par exemple, d'y introduire des catalogues grammaticaux, sémantiques, lexicaux tant soit peu complets.

Même sans ces limitations cependant, nous préférerions effectuer des expériences partielles. L'objectif poursuivi est, on l'a dit, la réduction du coût des catalogues de règles (coût mesuré en temps de travail humain) au prix d'une augmentation du coût de leur consultation (mesuré en temps machine). Seules des expériences partielles, effectuées avant rédaction complète des catalogues, peuvent montrer où l'on doit s'arrêter dans cette voie, en donnant une idée approximative des coûts.

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

1. Ch. BALLY — "Linguistique générale et linguistique française" A. Francke, S.A. (Berne 1950).

2. A. LEROY et P. BRAFFORT — "Notice relative à l'élaboration d'un codage par phrases-clés pour la programmation d'un système de sélection automatique de documents" Note CEA n° 278.

3. S. CECCATO — "Principles and classifications of an operational grammar for Mechanical Translation" Internation.Conference for Standards on Common Language for Machine Searching and Translation, Cleveland, September 1959.

4. N. CHOMSKY — "Syntactic Structures" Mouton and C° - 's-Gravenhage 1957.

5. L. TESNIERE — "Eléments de syntaxe structurale" Klincksieck - Paris 1959.

6. K.E. HARPER and D.G. HAYS — "The use of machines in the construction of a grammar and computer program for structural analysis" Proceedings of ICIP - Paris, June 1959, Unesco.

7. Knud TOGEBY — "Structure immanente de la langue française" Travaux du cercle linguistique de Copenhague, volume 6 - Nordisk Sprog-og Kulturforlog - Copenhague 1951.

8. Aravind K. JOSHI — "Recognition of Local Substrings" Rapports de l'Université de Pennsylvanie N° 18.

9. Henry HIZ — "Steps toward Grammatical Recognition" Rapports de l'Université de Pennsylvanie N° 21 a.

10. R.S. SOLOMONOFF — "A new method for discovering the grammar of phrase structure language" Proceedings of ICIP, Paris, June 1959, Unesco.

11. Z. HARRIS — "Co-occurence and transformation in Linguistic Structure" - Language n° 33, 1957.

12. Z. HARRIS — "Linguistic transformations for information retrieval", preprints of papers for the International Conference on Scientific Information, National Academy of Sciences National Research Council, Washington D.C. 1958.

13. E. MARETTI — "How to represent and rule correlating" International Conference for Standards on Common language for Machine Searching and Translation, Cleveland, September 1959.

14. E. ALBANI — "Construction of the correlational net by means of digital computer" International Conference for Standards on a common Language for Machine Searching and Translation, Cleveland, September 1959.

15. A. SESTIER — "La Traduction automatique" Ingénieurs et Techniciens, mars 1959, avril 1959, mai 1959, juin 1959.

16. L. HJELMSLEV — "Omkring Sprogteoriens Grundlaeggelse" Festkrift udgivet of Københavns Universitet November 1943 - Copenhague.

17. Y. BAR HILLEL — (Heb. U.J.) "Report on the State of MT in U.S.A. and Great Britain, 1959" Technical report n° 1 prepared for US Office of Naval Research, Information System branch, Contract NONR 2578(00) NR 049-130, February 1959.

18. J. LAMBEK — "The Mathematics of Sentence Structure" Amer. Math. Monthly, 65 : 3 (March 1958) 154-170.

19. Richard S. GLANTZ — "Further Investigation of English Syntax with the Theory of syntactic Types" NBS Report 6856 - October 1, 1959.

20. A.D. BOOTH, L. BRANDWOOD and J. CLEAVE — "Mechanical Resolution of linguistic problems" Butterworths - London 1958.

°°°°°

EXPERIENCE DE FEVRIER 1960

Eric Morlet

Organigrammes et description du programme réalisé pour l'ordinateur 650

La lecture de ce qui suit n'est profitable qu'au lecteur connaissant les conventions d'organigramme.

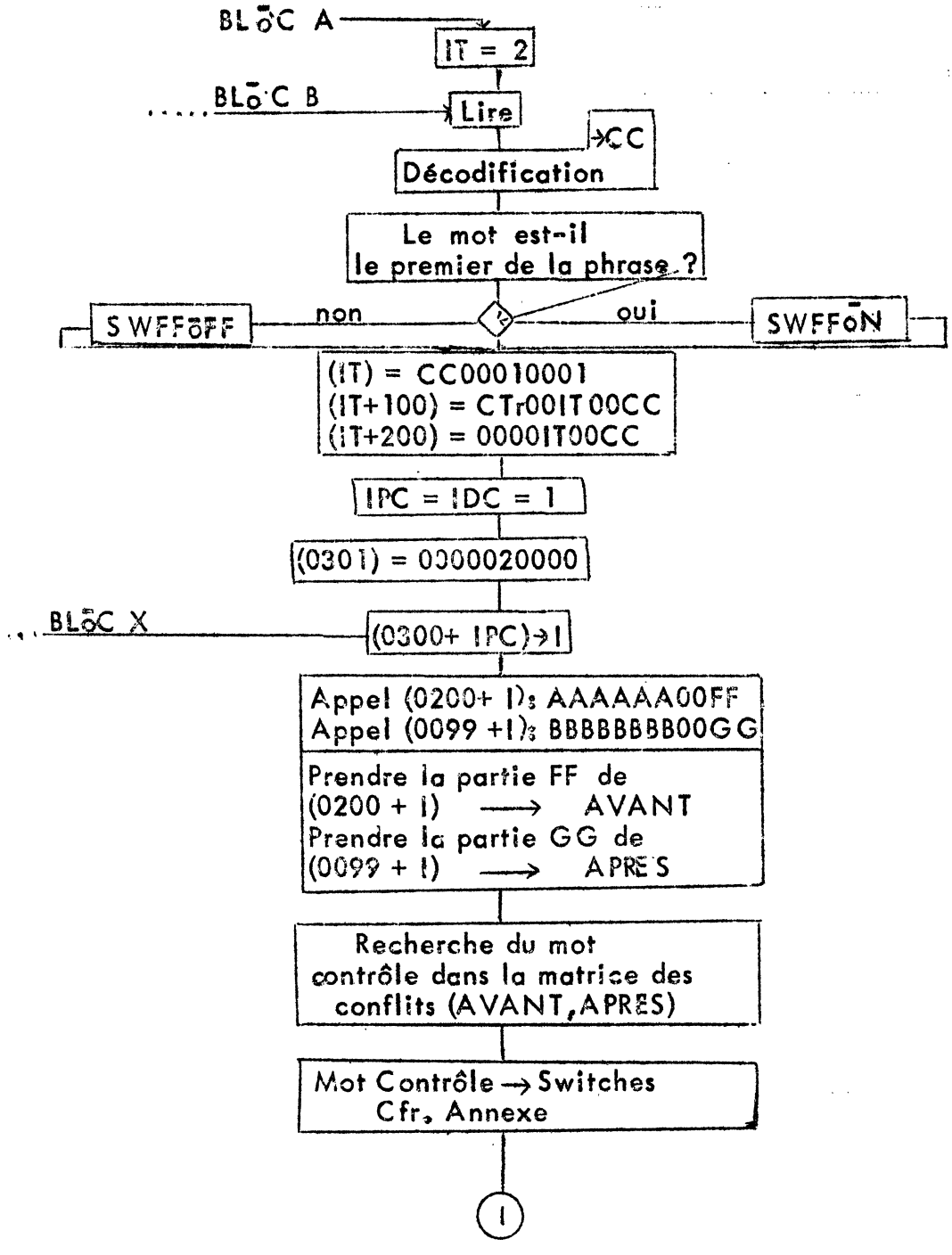
TABLES

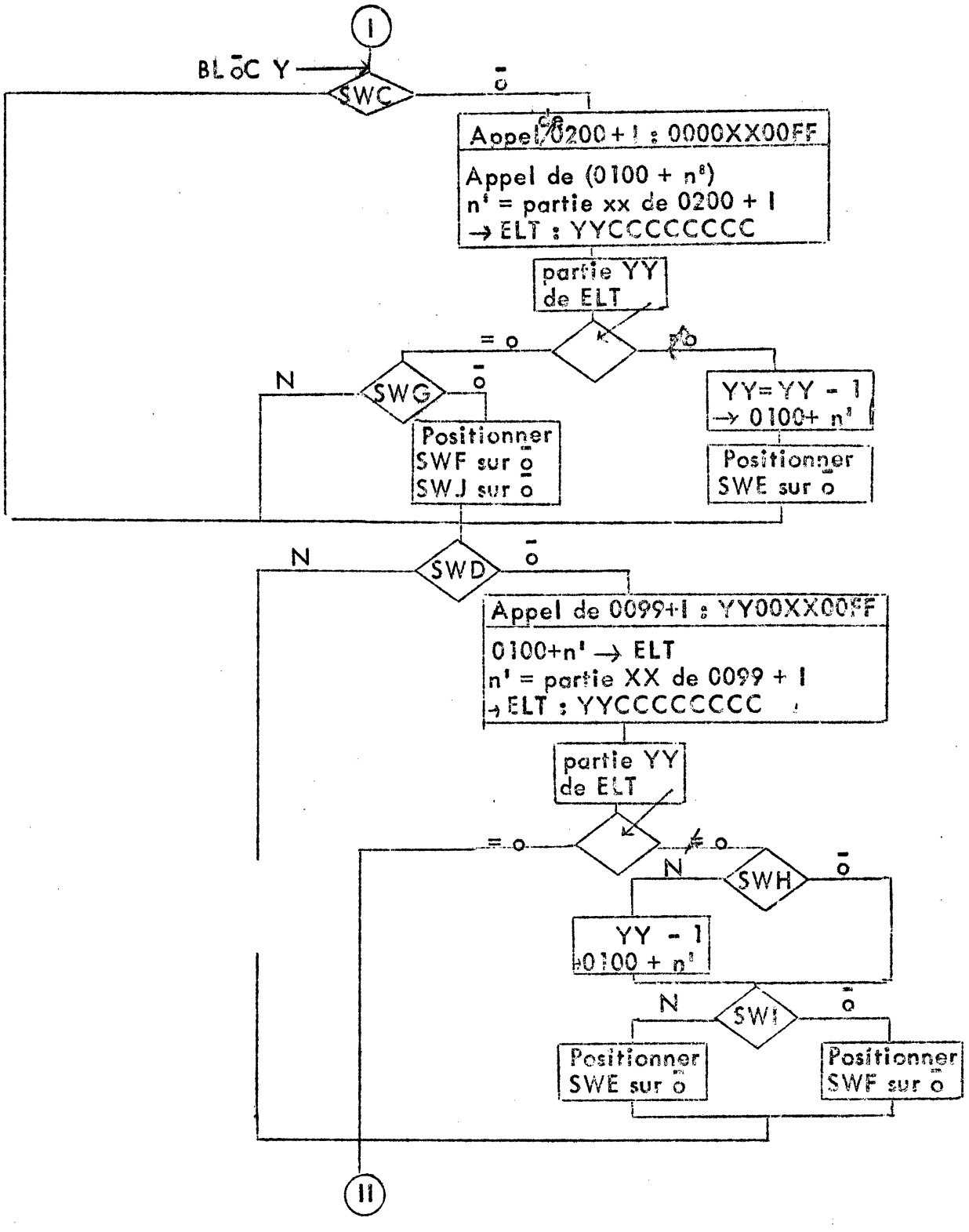
Emplacement	Structure
1-100	positions 10-9: Code Nature positions 8-7-6-5 : Nombre de parenthèses à la gauche du mot positions 4-3-2-1 : Nombre de parenthèses à la droite du mot
101-200	positions 10-9 : compteur positions 8-7-6-5 : Adresse du mot correspondant à la parenthèse extrême gauche positions 4-3-2-1 : Nature de ce mot
201-300	positions 10-9 : 00 positions 8-7-6-5 : Adresse du mot correspondant à la parenthèse extrême droite positions 4-3-2-1 : Nature de ce mot
300-350	Table des conflits en suspens (Table tournante) Soit le conflit : (A) (B), on trouve dans la table l'adresse de A
0351-0370 0401-0420 0451-0470 0501-0520	Matrice des consignes

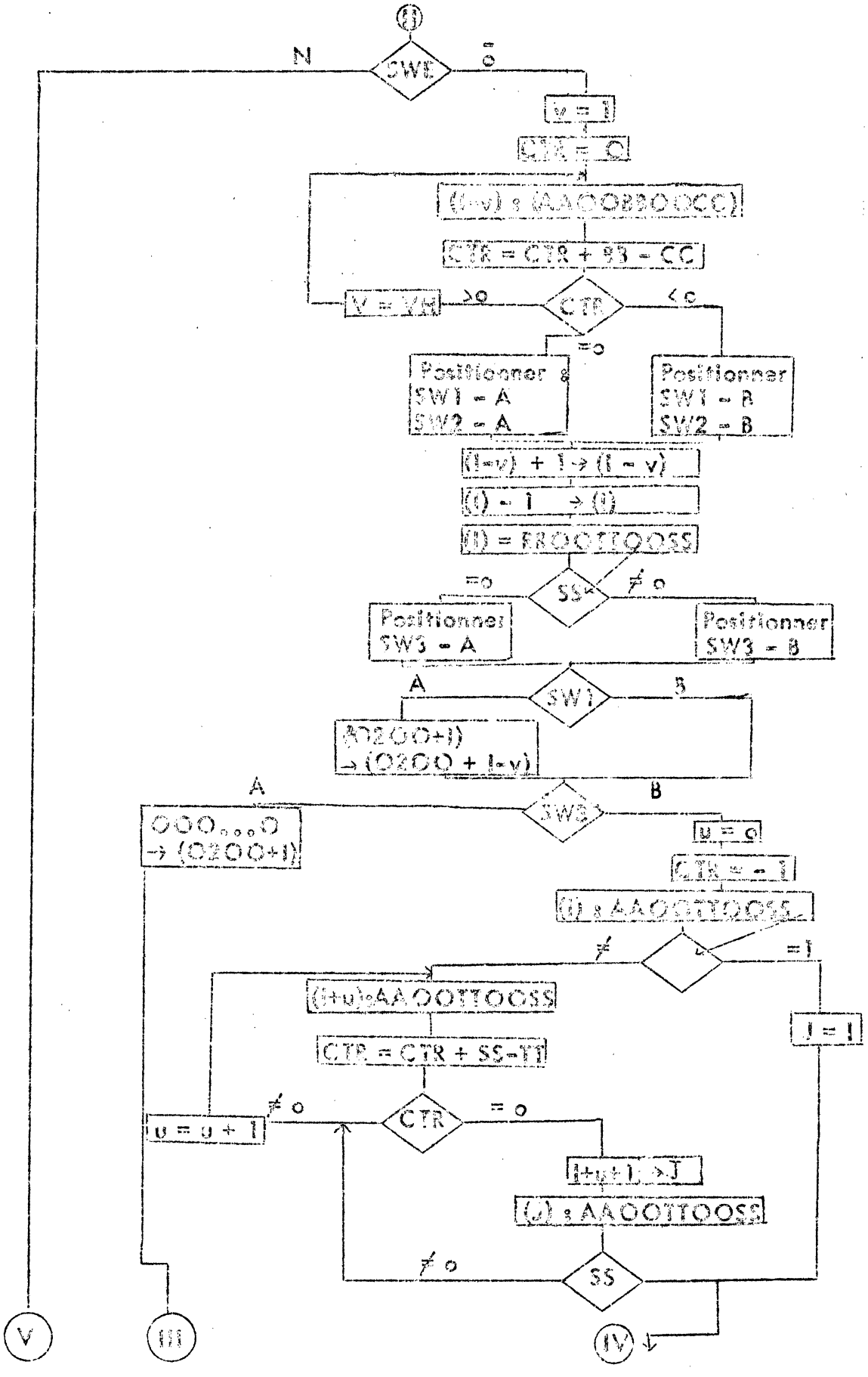
VARIABLES

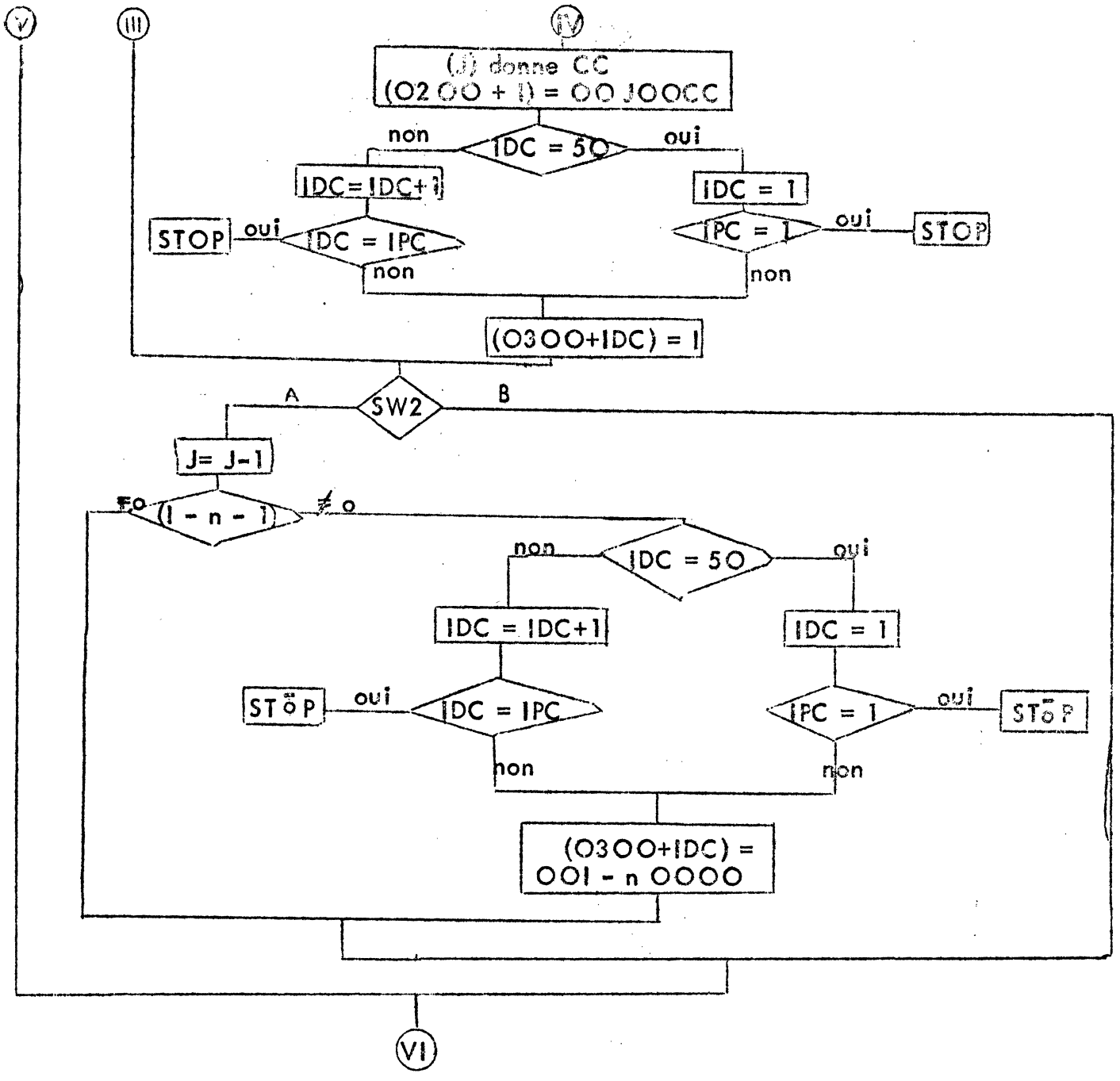
IT : Indice texte (Remarque : IT est initialisé à 2, le dernier "mot est systématiquement un point)
IDC : Adresse du premier conflit dans la table tournante
IPC : Adresse du dernier conflit dans la table tou nante
Avant : 1er indice d'une consigne
Après : 2ème indice d'une consigne.

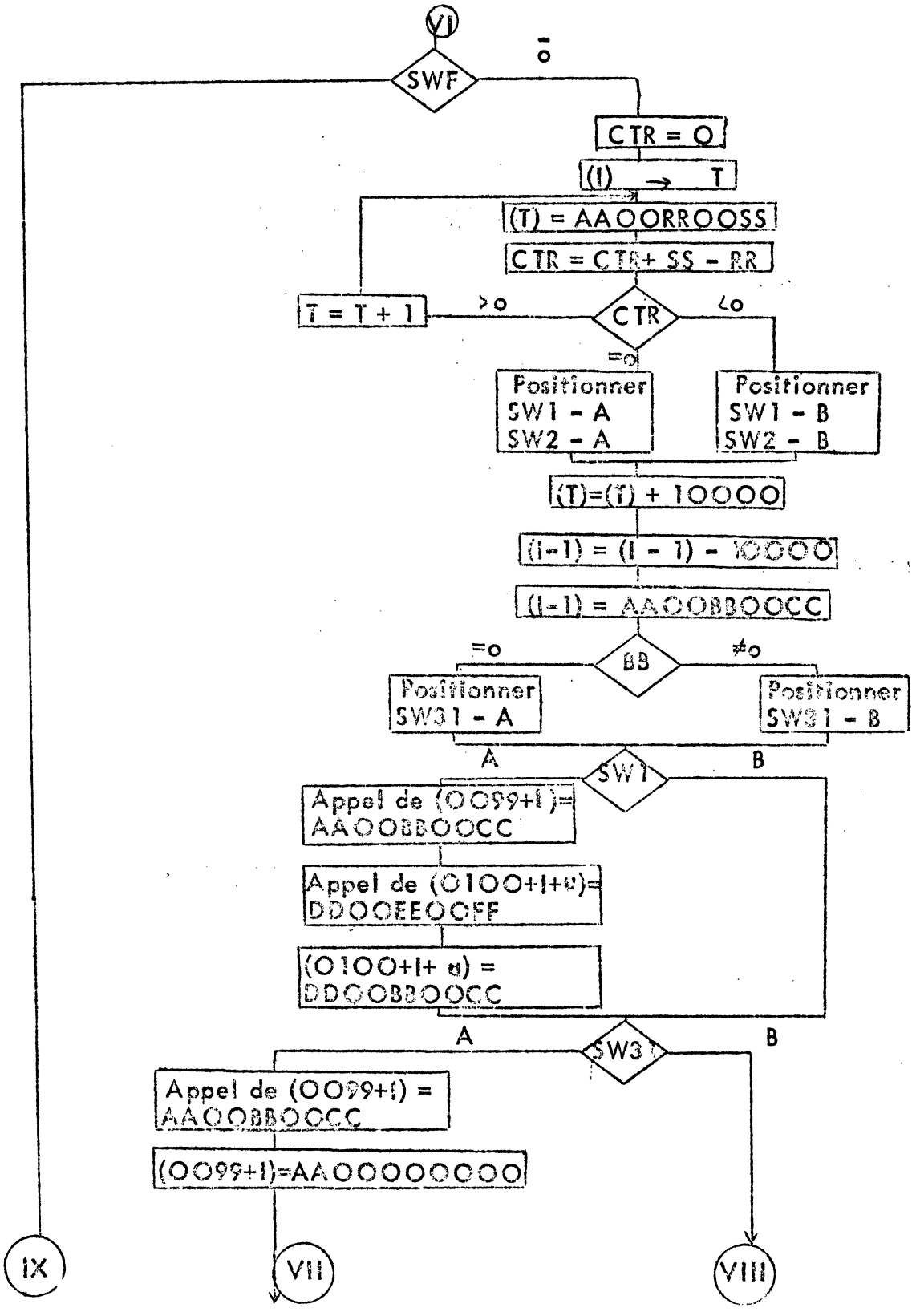
ORGANIGRAMME

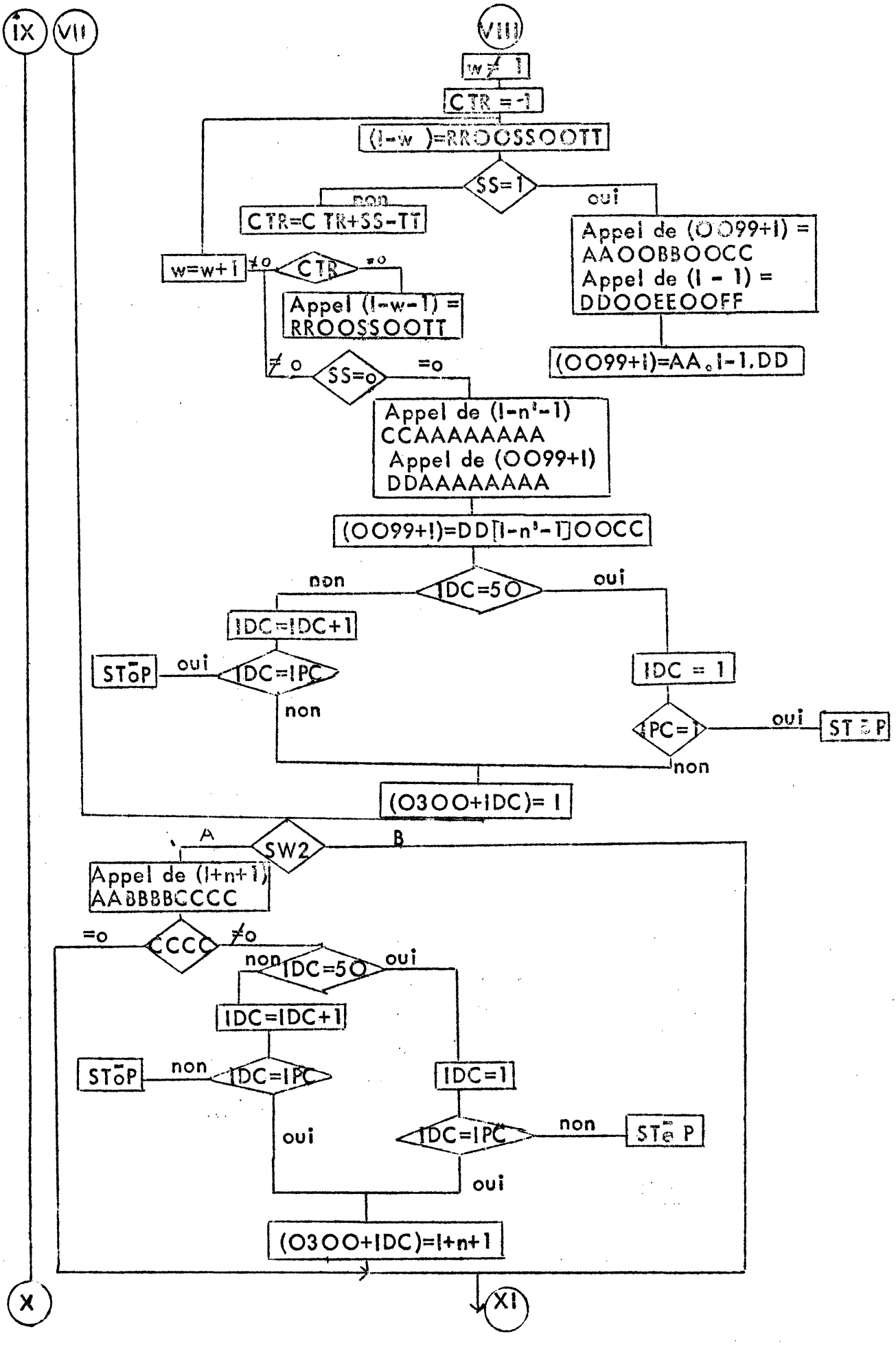


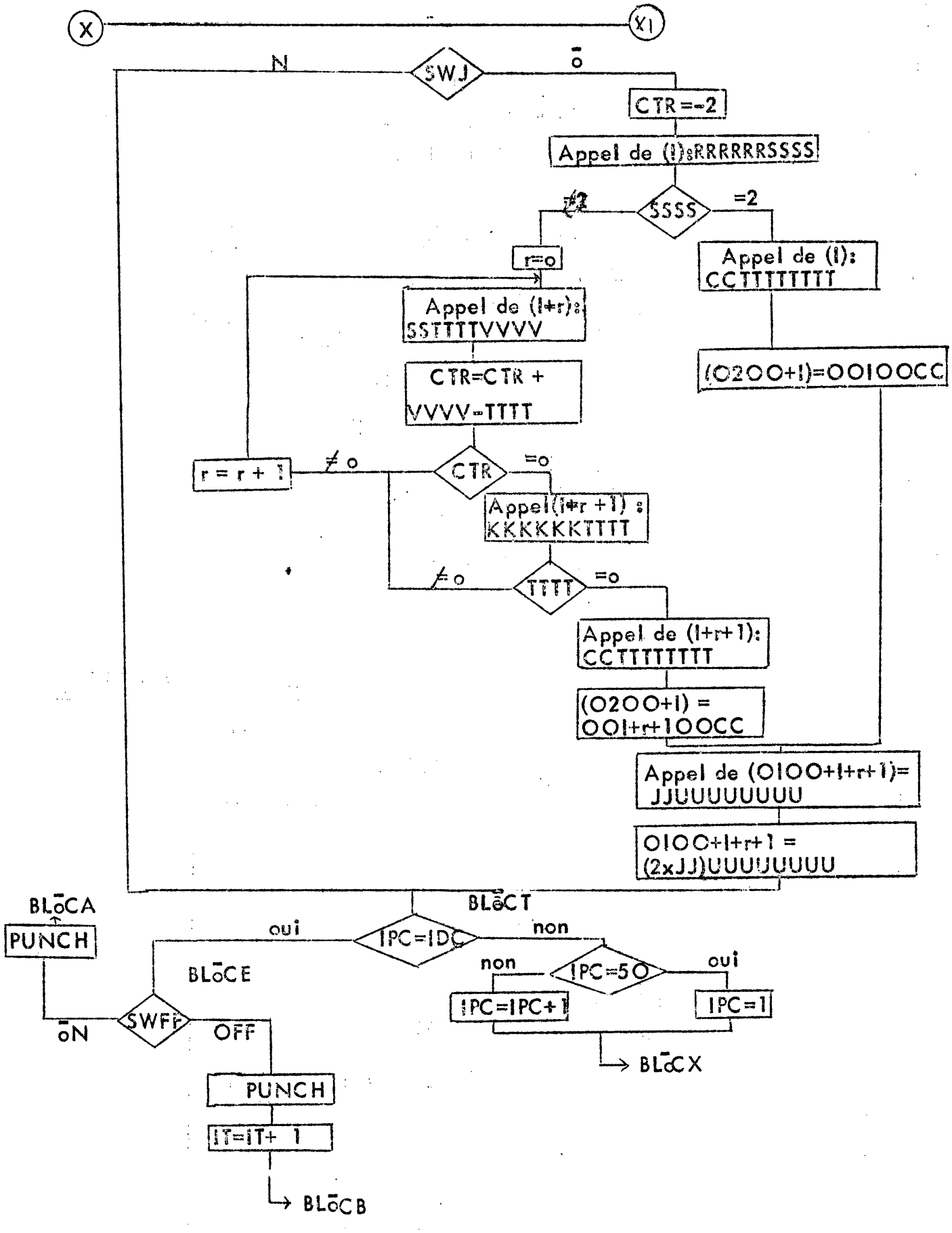

I
BLOC Y
SWC
Appel 0200 + I : 0000XX00FF
Appel de (0100 + n')
n' = partie xx de 0200 + I
→ ELT : YYCCCCCCCC
partie YY de ELT
= o
≠ o
N
SWG
Positionner SWF sur o
SWJ sur o
YY = YY - 1
→ 0100 + n'
Positionner SWE sur o
N
SWD
Appel de 0099+I : YY00XX00FF
0100+n' → ELT
n' = partie XX de 0099 + I
→ ELT : YYCCCCCCCC
partie YY de ELT
= o
≠ o
N
SWH
YY - 1
→ 0100 + n'
N
SWI
Positionner SWE sur o
Positionner SWF sur o
II


(i)
SWE
N
O
v = 1
CTR = 0
(i-v) : AAOOBBOOCC)
CTR = CTR + BB - CC
V = VM
>0
CTR
<0
=0
Positionner
SW1 - A
SW2 - A
Positionner
SW1 - B
SW2 - B
(i-v) + 1 → (i - v)
(i) - 1 → (i)
(i) = PROCTTOOSS
=0
SS
≠ 0
Positionner
SW3 - A
Positionner
SW3 - B
A
SW1
B
(0200+i)
→ (0200 + i-v)
A
SW2
B
000...0
→ (0200+i)
u = 0
CTR = - 1
(i) : AAOOTTOOSS
≠
=1
(i+u):AAOOTTOOSS
CTR = CTR + SS-TT
J = i
≠ 0
CTR
= 0
u = u + 1
i+u+1 > J
(J) : AAOOTTOOSS
≠ 0
SS
V
III
IV


V
III
IV
(J) donne CC
(O2 OO + I) = OO JOOCC
non
IDC = 5O
oui
IDC = IDC+1
IDC = 1
STOP
oui
IDC = IPC
IPC = 1
oui
STOP
non
non
(O3OO+IDC) = 1
A
SW2
B
J= J-1
Fo
(I - n - 1)
≠ o
non
IDC = 5O
oui
IDC = IDC+1
IDC = 1
STOP
oui
IDC = IPC
IPC = 1
oui
STOP
non
non
(O3OO+IDC) =
OOI - n OOOO
VI


VI
SWF
=o
CTR = Q
(I) → T
(T) = AAOORROOSS
CTR = CTR+ SS - RR
T = T + 1
>o
CTR
<o
=o
Positionner SW1 - A SW2 - A
Positionner SW1 - B SW2 - B
(T)=(T) + 10000
(I-1) = (I - 1) - 10000
(I-1) = AAOOBBOOCC
=o
BB
≠o
Positionner SW31 - A
Positionner SW31 - B
A
SW1
B
Appel de (0099+I)= AAOOBBOOCC
Appel de (0100+I+u)= DDOOEEOOFF
(0100+I+ u) = DDOOBBOOCC
A
SW3
B
Appel de (0099+I) = AAOOBBOOCC
(0099+I)=AAOOOOOOOO
IX
VII
VIII


IX
VII
VIII
w≠ 1
CTR =-1
(I-w)=RROOSSOOTT
SS=1
non
oui
CTR=C TR+SS-TT
Appel de (OO99+I) =
AAOOBBOOCC
Appel de (I - 1) =
DDOOEEOOFF
w=w+1
≠o
CTR
=o
Appel (I-w-1) =
RROOSSOOTT
(OO99+I)=AA.I-1.DD
≠ o
SS=o
=o
Appel de (I-n'-1)
CCAAAAAAAA
Appel de (OO99+I)
DDAAAAAAAA
(OO99+I)=DD[I-n'-1]OOCC
non
IDC=5O
oui
IDC=IDC+1
STOP
oui
IDC=IPC
non
IDC = 1
IPC=1
oui
STOP
non
(O3OO+IDC)= I
A
SW2
B
Appel de (I+n+1)
AABBBBCCCC
=o
CCCC
≠o
non
IDC=5O
oui
IDC=IDC+1
STOP
non
IDC=IPC
oui
IDC=1
IDC=IPC
non
STOP
oui
(O3OO+IDC)=I+n+1
X
XI


X
X1
SWJ
N
o
CTR=-2
Appel de (I):RRRRRRSSSS
SSSS
≠2
=2
r=o
Appel de (I):
CCTTTTTTTT
Appel de (I+r):
SSTTTTVVVV
(0200+I)=0010OCC
CTR=CTR +
VVVV-TTTT
r = r + 1
≠ o
CTR
=o
Appel(I+r+1) :
KKKKKKTTTT
≠o
TTTT
=o
Appel de (I+r+1):
CCTTTTTTTT
(0200+I) =
001+r+100CC
Appel de (0100+I+r+1)=
JJUUUUUUUU
0100+I+r+1 =
(2xJJ)UUUUUUUU
BLOCA
PUNCH
BLOCT
oui
IPC=IDC
non
non
IPC=50
oui
BLOCE
IPC=IPC+1
IPC=1
oN
SWFP
OFF
BLOCX
PUNCH
IT=IT+ 1
BLOCB

JOURNEE D'AUTOMATISATION

THE "L'UNITE" DOCUMENTATION SYSTEM

Th. W. te Nuyl (x)

When preparing this lecture a word written by Charles Dickens in "The Cricket on the Hearth" came to my mind, to wit : "If I come to tell a story I must begin at the beginning, and how is it possible to begin without to begin at the...." Dickens was able to give the word for the beginning of his story but when one has to talk about mechanized documentation it is not so easy to know with what one should begin.

It could be said : "Well, with the document"; but this answer would only be a very vague indication. It should be said at the same time what numbers of documents have to be considered, what the type of the documents would be, wether thick, like books,or thin, like patent specifications. In addition something should be said about the complexity of the documents. Thus it could be that for only a few documents, like the Scrolls of the Dead Sea, an extensive documentation is made with the use of machines, whereas for a large number of less important documents a limited system is used.

This leads to another aspect, to wit : the nature of analysis. This may be a deep one with the use of specific characteristics or less deep by using more general features or even only the indication of the document.

One could also start with the search. Here again a number of aspects can be considered such as : how often searches are made, and at what speed an answer must be given to the customer, perhaps asking his question over the telephone. For the way of searching it is of importance how the recorded features are arranged; "terms after items" or "items arranged after terms" which means that in the first case for each document all the features are recorded after the document reference number, requiring scanning of the whole file or a class of the whole file, when making a search, whereas in the second case, like in the Peek-a-boo system, the document reference numbers are grouped after the characterizing features. This way of arrangement allows a random access to the file and co-ordinating any number of "terms" when making a search.

(x) N.V. De Bataafsche Petroleum Maatschappij
(Royal Dutch Shell Group)

Last but not least comes the aspect of the required answer : Must the answer be precise ? May a pack of documents be revealed, which have then to be studied ? And may this pack contain false answers, due to the fact that the records have been kept small in numbers by superimposing features. This is the case when using a direct code, by which each punch on a punched card has a specific meaning.

When coming back to the question where to begin, we see that in fact we should start with the documentalist who has to face all of these aspects and who has to bring them into balance.

No machine can perform such a thing and documentation therefore always depends on the human being.

Before discussing the system as used in the patent division of the Shell International Research Maatschappij N.V. in the Hague, some information may be given on the history of the development thereof.

The whole project has grown out of the desire of the then head of the patent library to have machines performing a great deal of the monotonous administrative work required for the patent library.

Out of the preliminary studies, now about ten years ago, emerged ideas on the use of punched cards for recording information on chemical compounds given in patent specifications.

It was thought to be advisable to arrange items after terms and to try to make the number of items after the various terms of about the same magnitude.

In October 1956 a research committee was formed and out of the discussions grew the idea to use index words as such, without any coding, on 40columns Powers-Samas cards. By superimposed punching of up to about seven words which show interrelations,a card would be obtained allowing to make in one run searches based on a combination of question words.

The new card was called "L'Unité" card and it was made by means of the Samas reproducer, allowing so-called group gang-punching, controlled by a punch in the cards to be reproduced into a single L'Unité card.

The mechanized documentation has now been developed to a point where there can be distinguished three different applications :

a/ the mechanization of administrative work;

b/ the development of a chemical code and

c/ the use of index word cards.

Research work on machines started in September 1957.

Messrs. Powers-Samas were prepared to construct for us a special self-settingselecting device covering a field of 25 adjacent columns. This equipment has become available in September 1958 and has proved to be quite satisfactory.

In order to obtain as quickly as possible a certain amount of material it was decided to start with available copies of abstracts of own Shell inventions.

In these abstracts informative expressions were underlined - not only single words, but also complex-words or even expressions formed by a number of words - with the only limitation that not more than 19 letters could be used. These index-words were then recorded on the punched cards. The sequence of the index-words was indicated by means of punches in column 1. Since more than 7 words had sometimes to be used for one abstract, these had to be grouped into sets of interrelated words, each set being recorded on one L'Unité card.

In addition index-words for specific features could be taken up on separate index-word cards. All of the index-word cards were put into alphabetical order. By means of the available tabulator an inventory could be made in which the frequency of the use of each index-word was given.

From this inventory all possible single words -uniterms - were selected and listed in a "dictionary". Further a so-called "correlated dictionary" was prepared showing after each uniterm all complex index-words used in the system which contained the particular uniterm.

It was noticed that the frequency of use of most index-words was very low, often not beyond ONE.

This fact led later to the decision that less specific index-words should be used and for a number of technical fields even only uniterms were used. The available 20 columns were grouped in two fields. Up to 14 words could now be used per L'Unité card. It was then again found that many terms were only used once or only a few times. Since there is no necessity of using machines for finding a document characterized in the system by certain index-words which are only used once or only a few times, it was then decided to separate the frequent words from the non-

frequent words and to make L'Unité cards only for the frequent words, but replacing them, if necessary, by standard words or descriptors of up to ten letters. At the same time synonyms were taken care of.

In order to test the system we have repeated a number of searches, already carried through with our classified abstracts.

Sometimes we found a little bit more and sometimes we missed the answer, because we had limited the information on a single L'Unité card to 7 words; and in a few cases we had not used the correct characteristic words when making the punch instructions. We now have more experience in the selection of index-words.

Perhaps I could first show you a few pictures which will make it clearer to you how we have been working so far :

Here we have a punch instruction, I told you that we started with abstracts of our own cases and that we have underlined words therein. Later on, we have also handled other documents for which we have made special punch instructions as shown on this picture. (fig 1).

You can see a number of index-words to characterize the United States patent, bearing our internal reference number and our internal classification number. Here we have words up to 19 letters. And here you can see the numbers 1, 2, 3 etc. giving the place of the words in the abstract. You can so-to-say read : it is a hydro-forming process with sulphur-addition in connection with the catalyst life and sulphur containing stock. Here you will see another column in which there is indicated the number of the sentence, because in an abstract we can have more than one sentence. We want to put on one card only interrelated words. So words for one sentence are all on one card and words for another sentence on another card and and that is indicated here by the sentence number. The words I put on the punched cards you can see here. (fig. 2). The "one" is represented by one punch and the "h" is represented by 2 punches, an upper-punch, called A, B or 0 and a lower-punch indicating one of the figures 1 - 9. The other letters of the alphabet are also represented by 2 punches. So you can read, so-to-say the card and interpret these punches and print here on top the meaning, so that the contents of the card can be read. Here on the right half we have the class, the country and the document number. And here in column 39 we have a punch, which is a very important one, because it allows us to count the cards on the tabulator. Here in column 40 is a particular control punch for making later on the L'Unité cards.

Next - Here you see the whole collection of the words

shown before on the punch instruction and seven in total.(fig.3). First of all the operator is making a number-card; you can read the number on the top here, it is the same as the number on the other cards. See here the punch, the A-40; this punch is missing in the last card, were we put in column 40 the number of the words which has to come on the L'Unité card. This corresponds with the number here in column one of the last index-word card. When there is given the figure here, in column 40 the "A" punch is released. And here at the bottom we have the L'Unité card.

Next picture - Here we have the peculiar L'Unité card containing superimposed all the index-words together (fig. 4). Here in column 40 you see the 7 which allows us to check wether the L'Unité card is correct, because we must now have 7 punches in column 1. This allows us to make a machine check.

These cards are now our L'Unité file cards used in the search.

Next - (fig 5) Further we have written the punch instructions by machine because we do not want to check with the cards, because you can damage them, but by this punch document, as we call it, made with the tabulator, we can easily check wether the cards are correct. The 7 here must correspond with the 7 there.

Next one - (fig. 6) Here you see - I do not want to go into details - the organization-flow-sheet of the papers, the punch instructions and the punching of index-words, the interpreting, the tabulating, the production of L'Unité cards, etc...

I want to draw your attention to the fact that we have two ways of punching. One is with normal punching equipment, the automatic key-punch, but the other way,via tape,is by means of writing machines producing 5 channel-tape. By doing this we get at the same time the punch document, so that we do not have to go to the tabulator to obtain the punch document. The operator can see now what is done and if a mistake has been made, this can be corrected immediately.

We have the hope that no further checking will be necessary, but it appears that all machines are making mistakes and this is another place where the human being comes into the picture to correct the mistakes the machines are making. Sometimes something is indicated on the punch document but it does not appear on the tape. Thus a certain check is necessary.

The next - (fig 7) It is about the same picture showing, however, more details, but I do not want to elaborate on it.

Next one - Here you see the production of dictionary cards from index-word-cards (fig. 8). It is necessary to have

a list of the words used in the system and you have to take care of new words taken up in the system, but that is not so easy. I draw your attention to the point that you must have a dictionary of the words because you must put your questions exactly in the same way as you put them on file. You cannot use a synonym, you must use the exact word and the exact letters. When I made my first search, I used a word in the plural and I missed the word which was in the singular.

Next - Here you see a page of our dictionary (fig. 9) Here you have words in alphabetical order : hand, handle, handled etc. and after each single term, that is a sort of a uniterm, a complex expression, like hand-operated-pump. First I could go to the word "pump" and then I would find this expression after the word "pump". Here we have the word "dipping-head" and we will find it also after "dipping". When making a search, we are putting in a list all the uniterms to be considered in connection with the search. And then we look in the dictionary for the complex index-words which we could use in the search. Sometimes the list of complex index-words are rather long. Then they are not very useful, but in such a case you only use the shorter lists. Mostly you can find your answer by means of words occuring in short lits.

Next Please (fig. 10) - We have two ways of making a search. The L'Unité method and the tabulation method. We make a question card by means of the words found in the dictionary. You need not ask for the whole word, you can also take only parts of them so that you can cover a number of words having the same prefix and leaving out different endings. By superimposing question words, you are making a question card of the same type as a L'Unité card. Then we have to compare the L'Unité file cards with the question card by means of the special selecting device constructed for us by Messrs POWERS-SAMAS. Then we get the selected L'Unité cards and a list can be made of the document numbers. Then you give the list of numbers to the documentalist who can then have a look at the original documents. For the other way, called by us the "tabulation method", I go to the index-word file and select all cards bearing pertinent index-words. These cards are then put in the order of the document numbers and thereby all the index-words for one document come together. Then we go to the tabulator to have prepared a list of the index-words.

Next - You can see the result of such a search (fig 11) The question was wether there are documents on the use of boric acid as a catalyst in catalytic cracking. So we have take file words like : catalytic cracking, boric acid and acid vapour. These words are arranged in numerical order i.e. in the order of the document number. The words are also counted. So you can see two words and three words. Mostly the highest score will give the best answer, but this is not always true and we have had a case where the highest score was no good at all. But

in general the higher scores give the best answers, so we find here that boric acid is used in cracking or that a boric catalyst is used in catalytic cracking. So these cases give an answer to the question. This list is then given to the documentalist and he-she can then study whether there is an answer to the question or not.

Here are some pictures to show the department. Here are two automatic key-punches (fig. 12). The next picture shows an interpreter and a reproducer by means of which the L'Unité cards are made (fig. 13).

Next - Here are two sorters (fig. 14), one showing the special selective sorting device, made by Messrs. POWERS-SAMAS. The question card is put into the feeding box. By turning a wheel, the card is fed into the machine and sensed. Those pins which find a hole are elongated by simply turning a wheel. When making a search, the pattern set up by the elongated pins is compared with the punches in the file cards. If there is a match, the file card is selected.

On the next picture you will see the tabulator by means of which inventories are made since the machine can count the cards having a same index-word (fig. 15). This counting is very important because when making a search, we know how often a word occurs in the system. This allows us of getting more quickly an answer by using words which have only been used a few times.

There is further shown an interpolator by means of which packs of cards, each bearing a search word, are compared to find whether the same document number occurs in both packs. In such a case we know that the two question-words will be found in the same document.

The last picture shows the newer type of punch instructions (fig. 16). Words up to 10 letters are used and put on two different fields, one for words beginning with the letters a to (and including) i and the other beginning with j to (and including) z. There are special columns for indicating the sequence of the words in the sentence and the number of the sentence and in addition it has been indicated when words form together a single complex expression. Since not more than 7 words may be put on a single field, there is indicated at the left, for checking purposes, how many words will come on each field.

I hope I have made it clear to you that there are different ways to get an answer with the system : The searching by means of L'Unité cards, the comparison of packs of cards in numerical order by means of the interpolator and the tabulation method. The system allows us to search at will by serial scanning (L'Unité cards) or by correlating relevant parts of the file through random access to the index-words.

In conclusion I would like to say a few words on statistical data we have collected on the use of index-words.

For our own patents, we have used about 7 index-words of 19 letters per document.

For about 4500 documents we used about 30.000 index - word cards. The dictionary contains now about 15.000 complex words, composed of about 5700 single words or uniterms.

In a system of 2.370 patents on polyvinylchloride we have worked with batches, using index-words of up to ten letters. The analysis was somewhat deeper than with the patents on Shell inventions. For 1.000 documents we used about 24.000 index-word cards based on 7.500 index-words.

There was a gradual decline in the increase of the dictionary. For the first hundred documents we had 1.500 words, around 1.000 documents we have 700 new words and around 2.000 documents 500 new words per 100 documents.

In our opinion with our present way of working the frequency of use of the index-words is too low to give a true machine method. A further study of the frequency of use of index-words will certainly lead us to a more appropriate mechanized system.

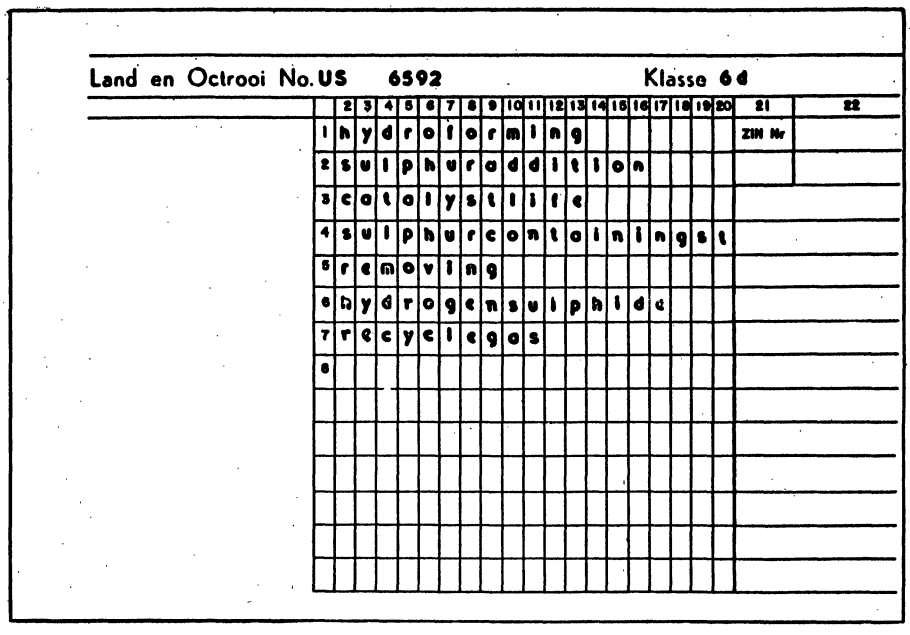


Fig. 1 - Punch instruction.

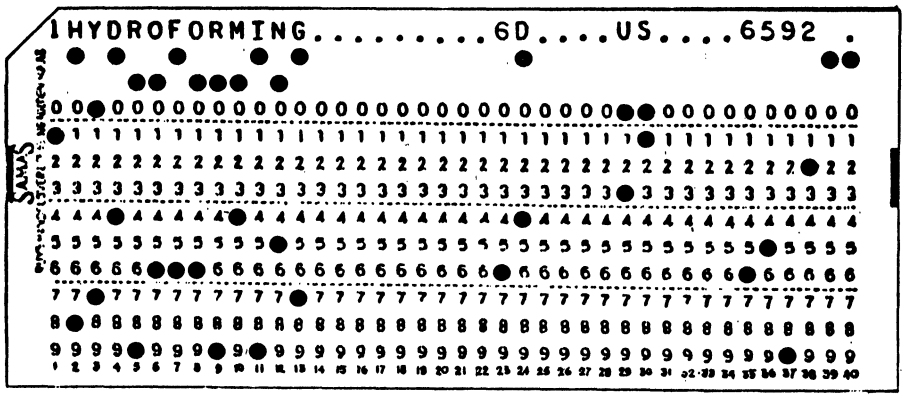


Fig. 2 - Word put on a punched-card.

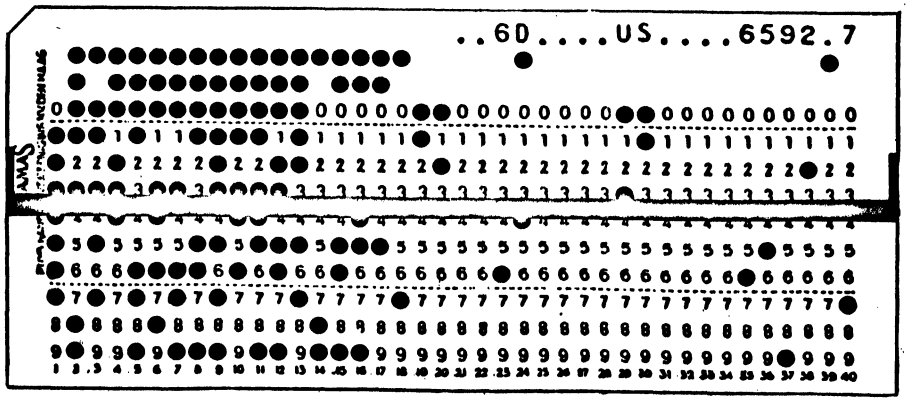


Fig. 4 - L'unité card.

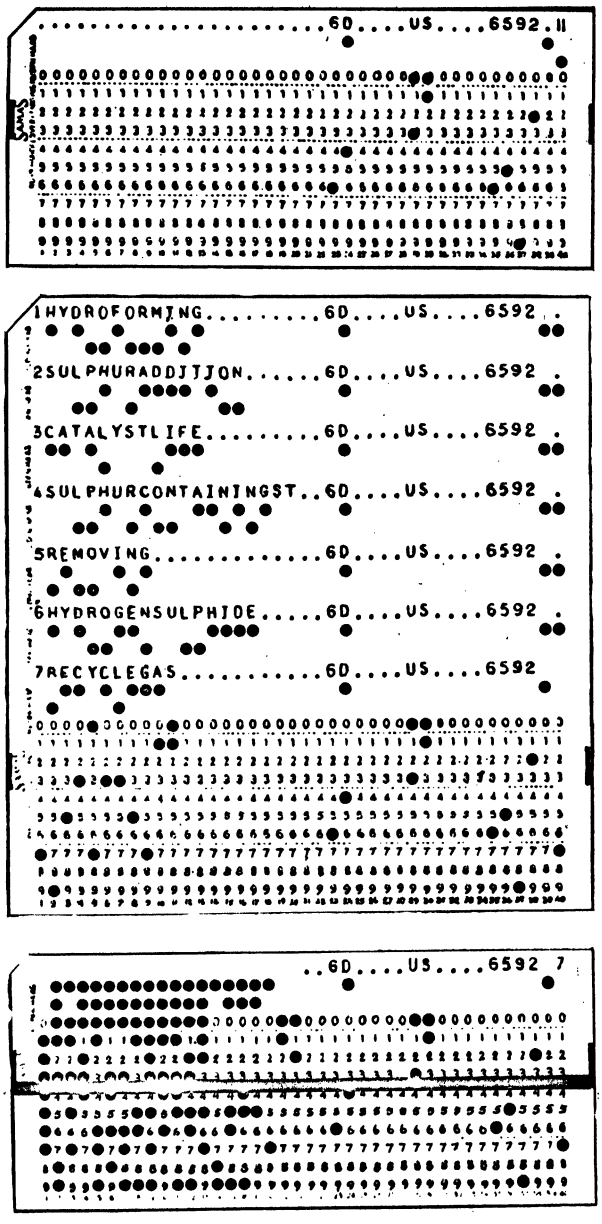


Fig. 3 - Whole collection of the words shown on punch instruction.

	6D	31	6592
1HYDROFORMING	6D	31	6592
2SULPHURADDITION	6D	31	6592
3CATALYSTLIFE	6D	31	6592
4SULPHURCONTAININGST	6D	31	6592
5REMOVING	6D	31	6592
6HYDROGENSULPHIDE	6D	31	6592
7RECYCLEGAS	6D	31	65927

Fig. 5 - Punch-document is Punch-instruction written by machine.

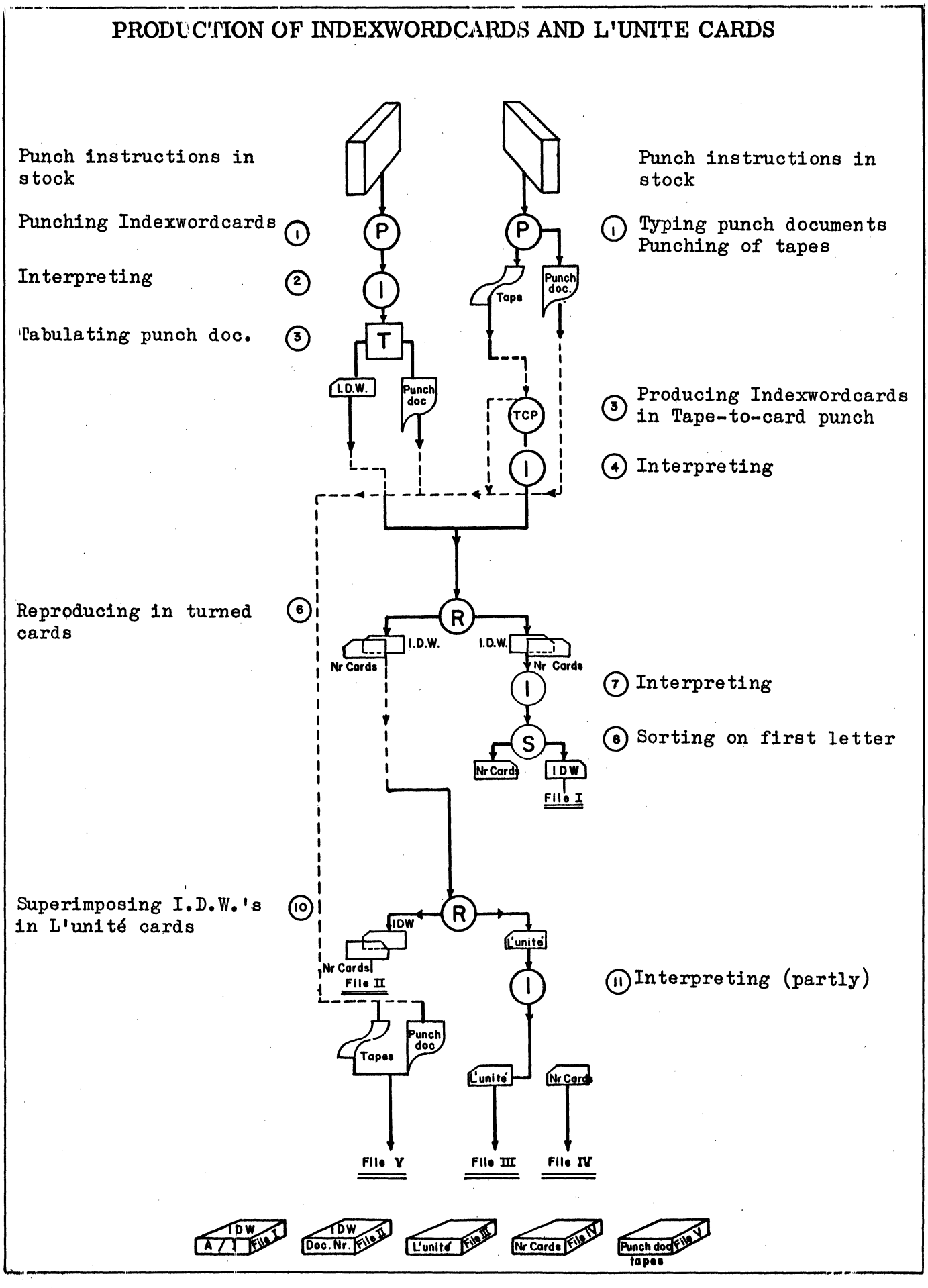


Fig. 6 - Organization flow-sheet.

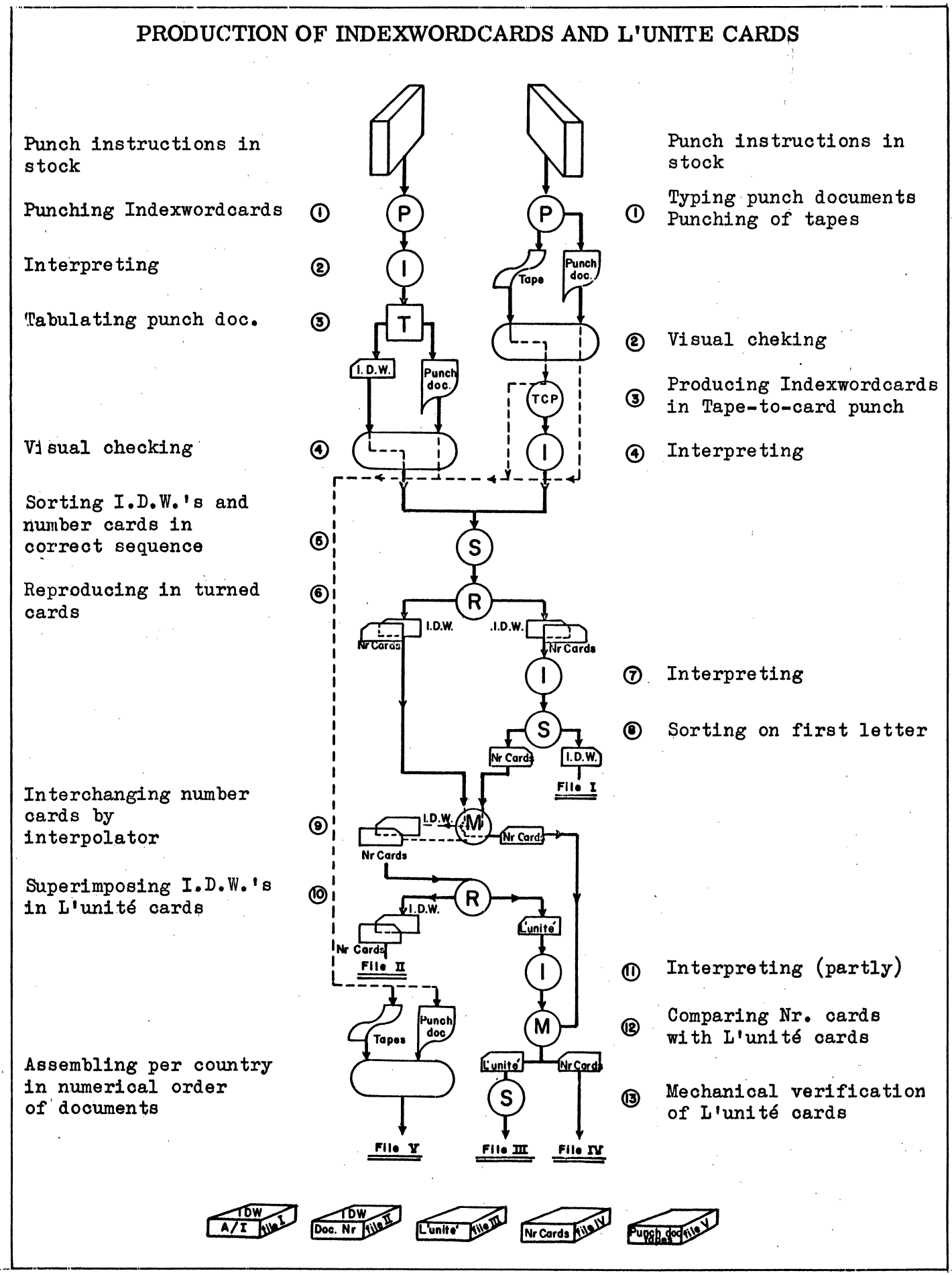


Fig. 7 - Organization flow sheet with more details.

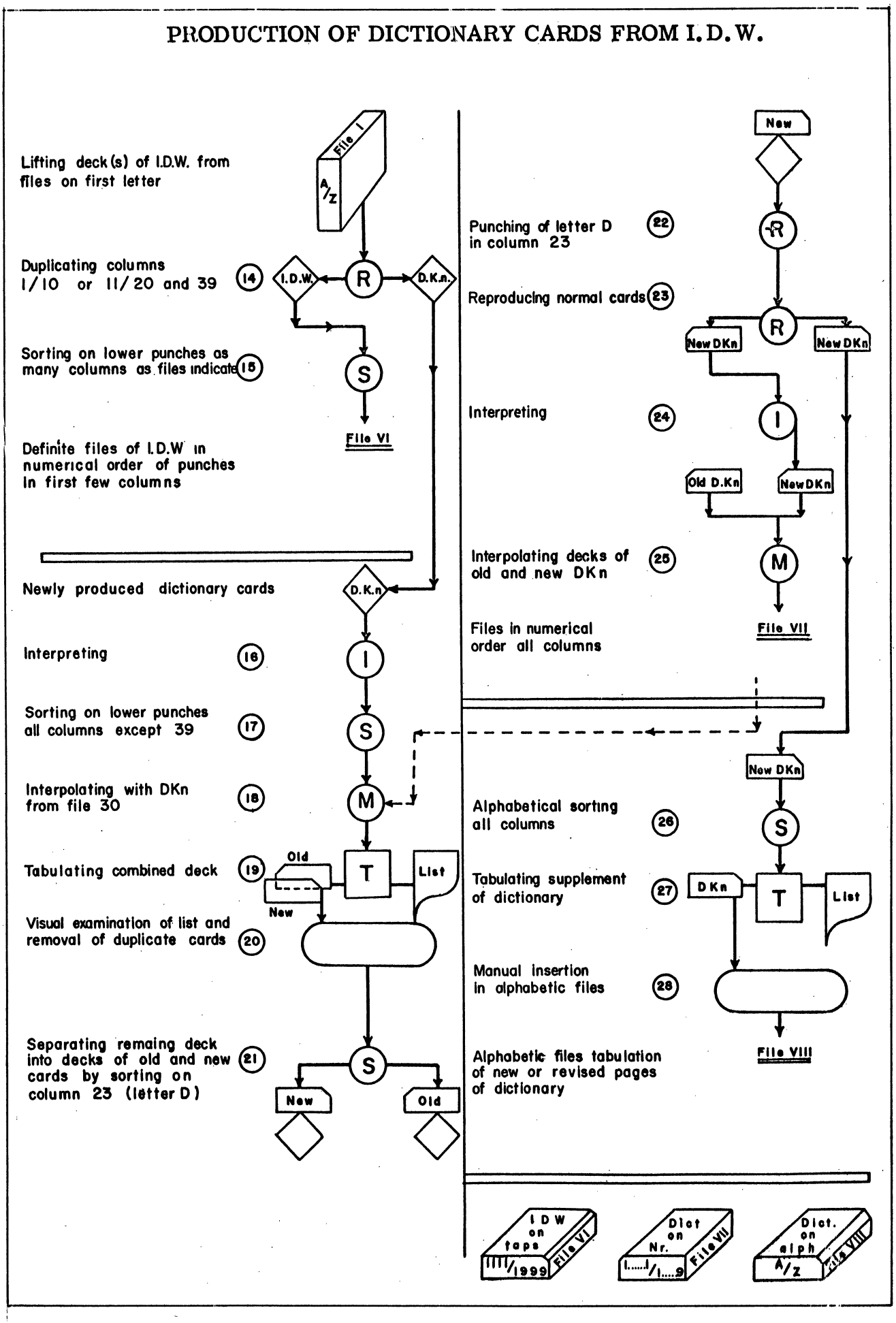


Fig. 8 - Production of dictionary cards.

HAND
HANDAPPLICATOR
HANDOPERATEDPUMP

HANDLE
HANDLE
SINGLEHANDLECONTROL

HANGING
HANGINGBAR
HANGINGCURTAIN
PIPEHANGINGMANIFOLD

HARD
HARDASPHALT
HARDMETAL
HARDMETALBUSHING
HARDMETALPINS
HARDMETALROLLER
HARDPARTICLES
HARDRESIDUE
HARDRESIDUEWAX
HARDRUBBERLAYER
HARDWAXCOMPOSITION

HARDENED
HARDENED
HARDENEDCASTOROIL
HARDENEDINTHESAND

HARDENING
AGEHARDENING
CASEHARDENING
FLAMEHARDENING
HARDENINGAGENT
HARDENINGOIL
RATEOFHARDENING
SURFACEHARDENING

HARDNESS
BRINELLHARDNESS
HARDNESS
HIGHHOTHARDNESSVALU
HOTHARDNESS

HARMFUL
HARMFULDEPOSIT

HAVING
HAVINGASURFACEAREA

HAZARD
FIREHAZARD
NAVIGATIONHAZARD

HCL
HCLACCEPTOR

HCN
BYREACTINGWITHHCN

HD
HDOIL

HEAD
CASINGHEADASSEMBLY
CATHEAD
CONVEXCYLINDERHEAD
CYLINDERHEAD
DIPPINGHEAD
FINNEDPOWERHEAD
FLANKSOFRAILHEAD
FLOATINGHEAD
FLOATINGHEADCOVER
FORAWELLHEAD
HYDROSTATICHEAD
SAMPLINGHEAD
SENSINGHEAD
TEEHEADPONTOON
UNDERSIDEOFRAILHEAD
UPPERSIDEOFRAILHEAD
WELLHEAD
WELLHEADASSEMBLY

HEARTH
OPENHEARTH
OPENHEARTHFURNACE

HEAT
AGEDBEFOREUSEBYHEAT
ELIMINATINGHEATLOSS
EXCESSHEAT
GOODHEATCONDUCTOR
HEAT
HEATACCUMULATING
HEATACCUMULATINGRIB
HEATBALANCE
HEATCARRIER
HEATCONDUCTIVITY
HEATCURABLE
HEATCURABLECOMPOSIT
HEATCURING
HEATDISTORTIONTEMP
HEATENGINE
HEATEXCHANGE
HEATEXCHANGER
HEATEXCHANGERTUBE
HEATEXCHANGETUBES
HEATEXCHANGINGTUBE
HEATINSULATING
HEATINSULATINGSTRUC
HEATINSULATINGTANK
HEATOFCONDENSATION
HEATOUTPUTCONTROL
HEATRADIATION
HEATRAYCONDUCTOR
HEATRESISTANCE
HEATRESISTANT
HEATRETENTIVESOLID

Fig. 9 - Page of dictionary.

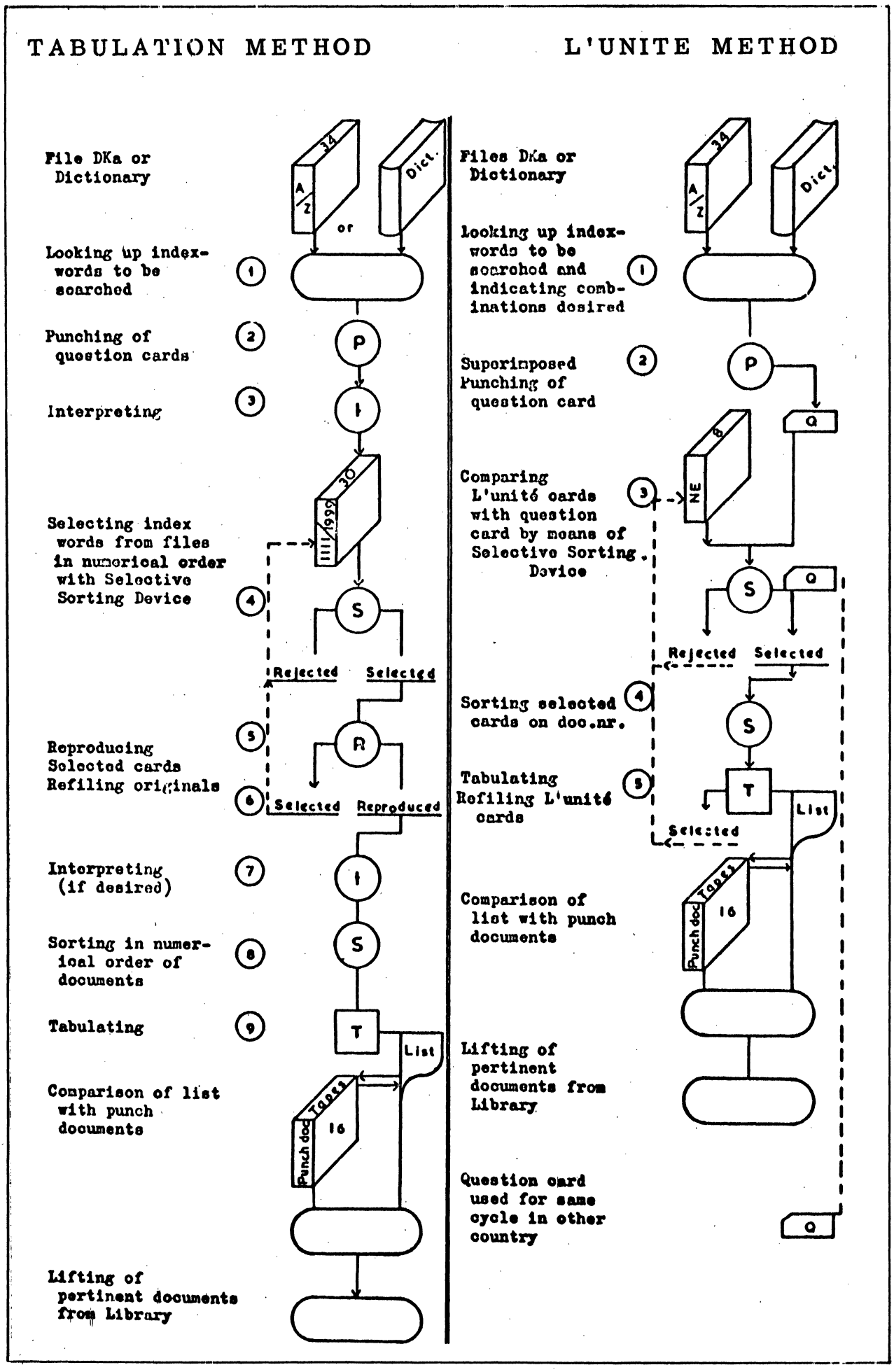


Fig. 10 - Two search methods.

1CATALYTICCRACKING	6F1	72	40291
			1
4BORICOXIDE	6F1	31	41141
5CRACKING	6F1	31	41141
			2
1CATALYTICCRACKING	6F1	31	41391
3BORICACIDCATALYST	6F1	31	41391
			2
5CATALYTICCRACKING	6F1	31	41451
			1
4CRACKING	6F1	31	41491
			1
2CATALYTICCRACKING	6F1	31	41501
			1
2CATALYTICCRACKING	6F1	31	41541
			1
1CRACKING	6F1	31	41641
			1
1BORIAALUMINACATALYS	6F1	31	41671
2CATALYTICCRACKING	6F1	31	41671
7BORICACIDVAPOUR	6F1	31	41671
			3
1CATALYTICCRACKING	6F1	31	41741
			1
1CATALYTICCRACKING	6F1	31	41751
			1
1CATALYTICCRACKING	6F1	31	41821
			1
1CATALYTICCRACKING	6F1	31	41921
			1
1BORIAALUMINACATALYS	6F1	31	42011
3CATALYTICCRACKING	6F1	31	42011
			2
1CATALYTICCRACKING	6F1	31	42361
			1

Fig. 11 - Result of search according to tabulation method.

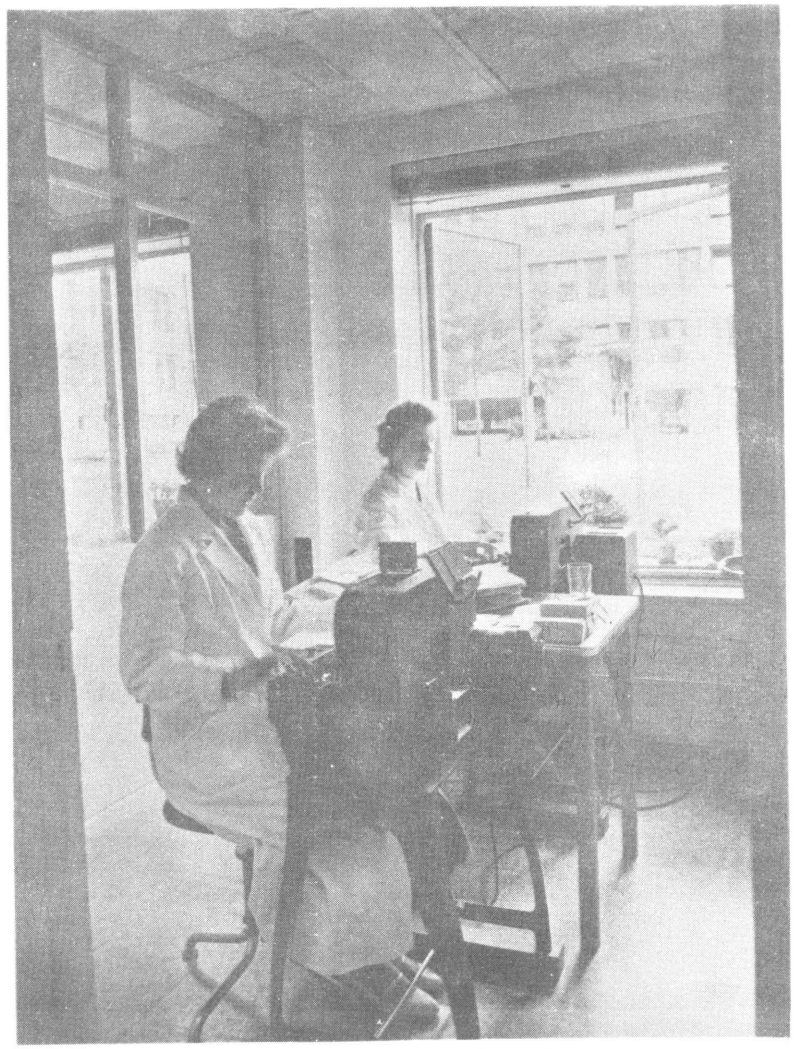


Fig. 12 - Two automatic key-punches.

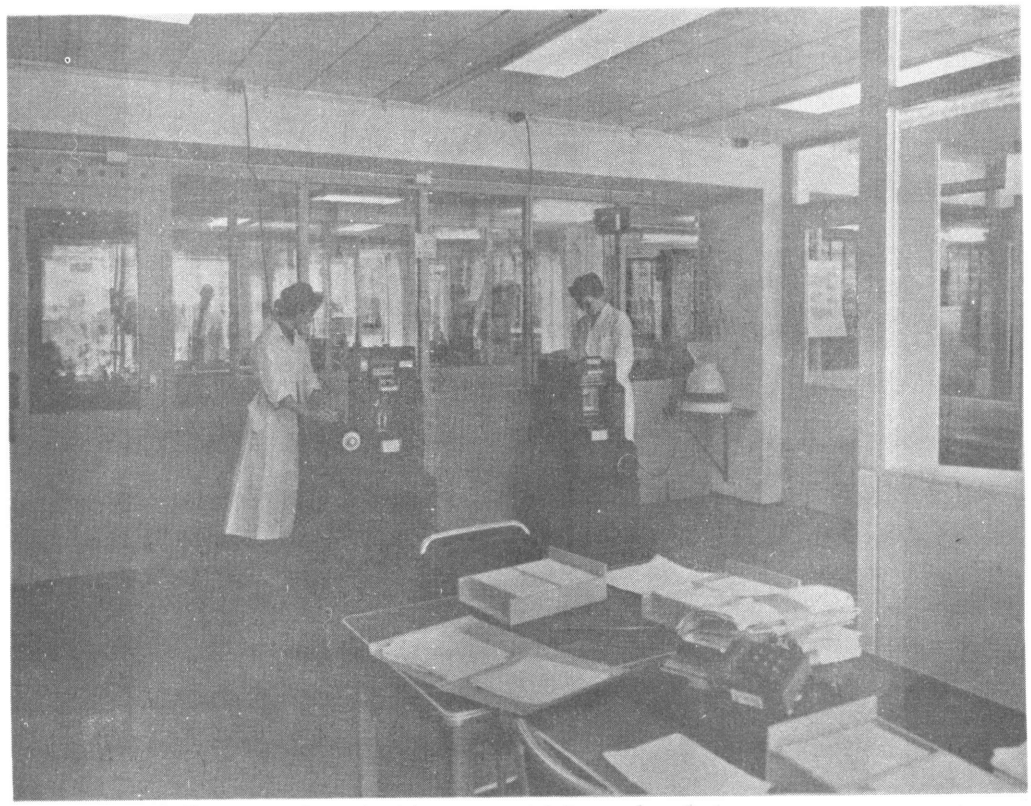


Fig. 13 - Interpreter (right side)
reproducer (on the left)

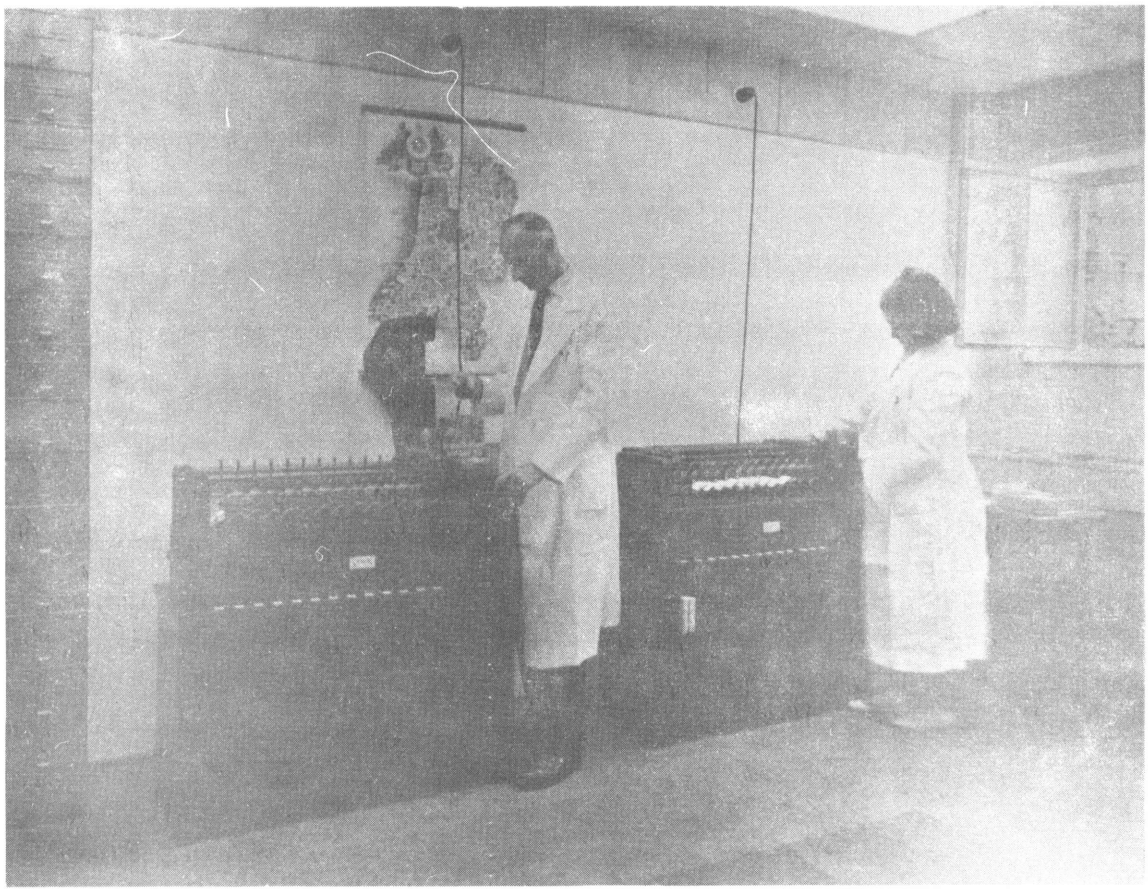


Fig. 14 - Two sorters, the one on the left showing the selective sorting device.

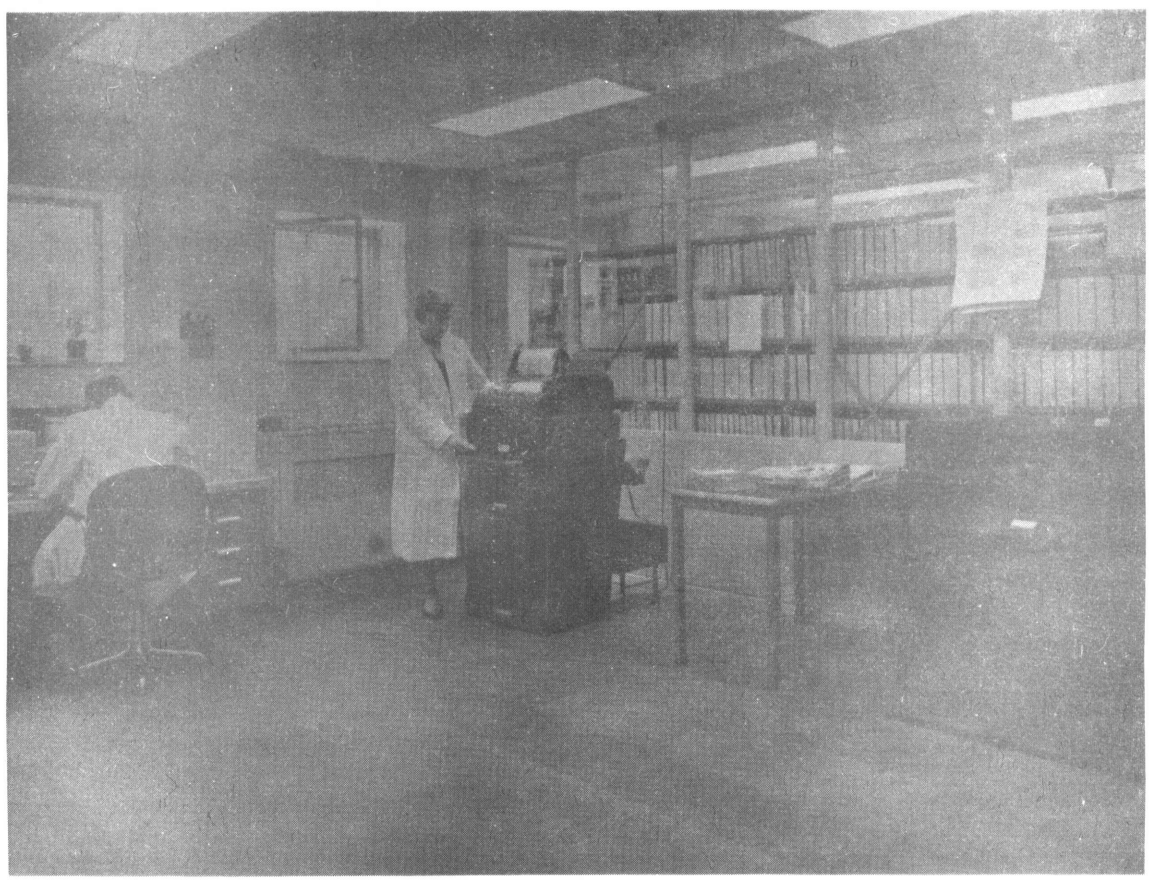


Fig. 15 - Tabulator interpolator

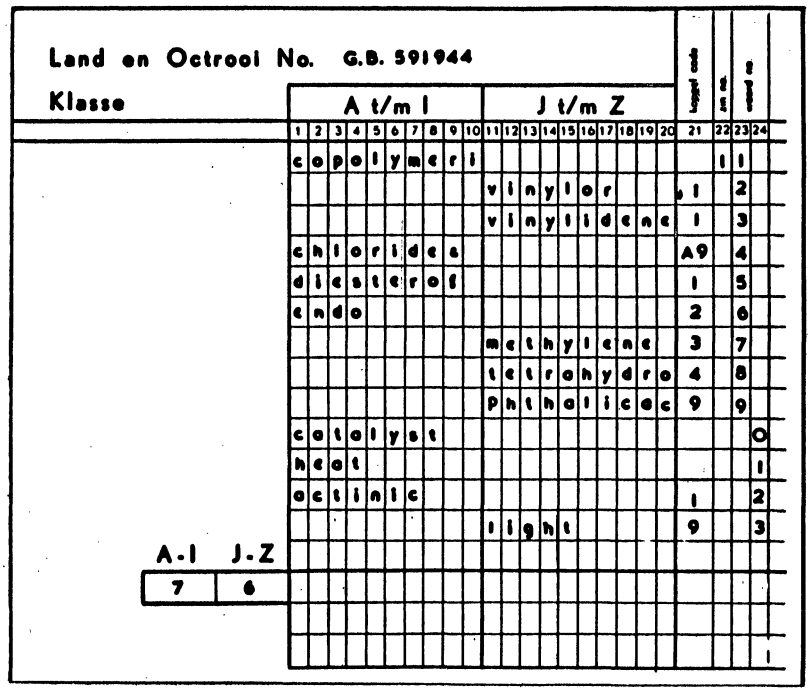


Fig. 16 - Punch instruction.

QUELQUES PROBLEMES POSES PAR LE TRAITEMENT DE L'INFORMATION NON NUMERIQUE

J. Poyen *

Je ne voudrais pas ici faire un cours général sur les machines à calculer, ou plus exactement à traiter l'information. Je supposerai connues de vous les caractéristiques générales de ces machines, leurs possibilités et leurs restrictions, tant dans la classe du matériel dit classique : perforatrices, reproductrices, trieuses, interclasseuses, tabulatrices, que du matériel beaucoup plus puissant, pour la désignation duquel n'existent pas de mots généraux, mais simplement des noms caractérisant chacune de ces machines en particulier : 650, Gamma AET, 704, Gamma 60 etc.

Je n'essaierai pas non plus de faire le tour d'horizon des applications de ces machines dans le domaine qui nous intéresse ici, celui du traitement de l'information non numérique, car cette liste serait beaucoup trop importante pour le temps dont je dispose, et de toutes façons incomplète, le champ d'application de telles machines s'étendant chaque jour.

Parmi les problèmes que l'on vous a exposés durant cette semaine, certains sont encore du domaine de la recherche (analyse automatique de textes, traduction des langues etc...); d'autres, par contre, sont effectivement dans la phase de réalisation (mécanisation de la recherche documentaire par exemple), en admettant les documents déjà codés.

J'aimerais essayer de mettre en lumière quelques points précis d'applications mécanographiques, travaux d'approche de solutions futures ou réalisations concrètes, et de vous montrer comment ont été résolues certaines difficultés, inhérentes au traitement de l'information non numérique.

Parmi les centres mécanographiques les plus intéressants, nous pouvons tout d'abord citer le centre de Besançon, dirigé par M. QUEMADA. Les travaux en cours de réalisation à ce laboratoire concernent les principaux types de compilations lexicologiques ou lexicographiques

- Index des mots

- Concordance des textes littéraires

- Inventaires lexicologiques

- Diverses recherches documentaires du type "répertoire historique du vocabulaire".

* Compagnie des Machines Bull.

Le Laboratoire, équipé en matériel Bull, comporte :

- 3 perforateurs de bande utilisés pour perforer la bande à partir d'un clavier similaire à celui d'une machine à écrire.

- 1 lecteur de bande utilisé pour lire les informations enregistrées sur la bande sous forme de perforations.

- 2 lecteurs connectés à 2 poinçonneuses qui servent à transcrire automatiquement sous forme de perforations dans des cartes, l'information contenue sur la bande.

- 1 traductrice qui sert à imprimer en clair dans la partie supérieure de la carte les informations perforées dans celle ci.

- 1 reporteuse traductrice qui peut imprimer sur une carte, à une ligne quelconque, l'information contenue sur cette même carte ou dans une autre carte.

- 1 reproductrice pour reproduire des cartes perforées.

- 1 trieuse électronique D 3 pouvant sélectionner des cartes comportant les perforations recherchées.

- 1 interclasseuse permettant de sélectionner et interclasser des cartes.

- 1 tabulatrice.

Dès l'abord un problème fort important a dû être résolu. En effet, les machines mécanographiques classiques, utilisées généralement pour la comptabilité, devaient être adaptées au traitement désiré.

Cette adaptation, souvent complexe, n'avait encore jamais été réalisée.

Les modifications les plus importantes concernaient les signes graphiques du français et la transformation des dispositifs de tri, d'interclassement et d'impression.

Compte tenu des possibilités des machines, qui appartiennent à la classe petit matériel, les signes suivants ont été adoptés pour l'impression des textes français :

a à â b c ç d é è ë ê f g h i ï î j k l m n o ô p q r s t u ù û ü v w x y z

soit 26 lettres plus à â ç é è ë ê ï î ô ù û ü

ponctuation . , ; : ! ? " - / '

chiffres arabes 0 1 2 3 4 5 6 7 8 9

figure 1

La distinction entre (et) a dû être sacrifiée, et ces deux signes sont représentés par /. D'autre part, la distinction entre majuscules et minuscules est impossible pour la totalité des signes alphabétiques. L'impression est donc faite dans un seul système au choix, en totalité majuscule ou minuscule.

L'utilisation des machines perforatrices de bandes rendait possible l'introduction des 59 codes différents, puisque le clavier de la machine émettrice dispose d'un nombre de touches suffisant, alors que cela était impossible à partir du clavier de la poinçonneuse de cartes qui ne comporte que 36 touches.

En effet, le code mécanographique standard ne comporte que 36 combinaisons au total, d'une ou deux perforations par colonne.

Pour atteindre le nombre de combinaisons nécessaires, il a fallu employer un code à 3 perforations par colonne, procédé entièrement nouveau.

Parallèlement, les imprimantes classiques ne disposant que de 36 caractères, celles-ci ont été modifiées pour la traduction et l'impression des fichiers. Les dispositifs d'impression de la tabulatrice et de la reporteuse ont également été modifiés.

Parmi les différents travaux effectués à ce centre nous allons prendre comme exemple la réalisation de l'indexage des mots.

Nous entendons par là un indexage systématique de tous les mots d'un texte donné pour en obtenir par exemple la liste imprimée suivie des différentes références du mot dans le texte. Ce genre de travail est fait entièrement automatiquement.

Le plan de travail est le suivant :

- Création d'un fichier de base où chaque carte comporte un mot.

- Classement du fichier par ordre alphabétique (les diverses formes flexionnelles étant regroupées sous un indicatif commun).

- Impression du fichier.

a) Perforation du texte : le texte à mettre sous forme de fichier est d'abord perforé sur bande perforée. Les références permanentes (auteur, références du texte complet) sont perforées une seule fois au début du texte, les références variables (page, n° de vers par exemple) sont frappées au début de chaque séquence correspondante, précédées des codes de service. Le texte est perforé d'une manière suivie, la frappe sur la barre espace provoquant également la perforation d'un code de service particulier. La bande se présente alors de la façon suivante :

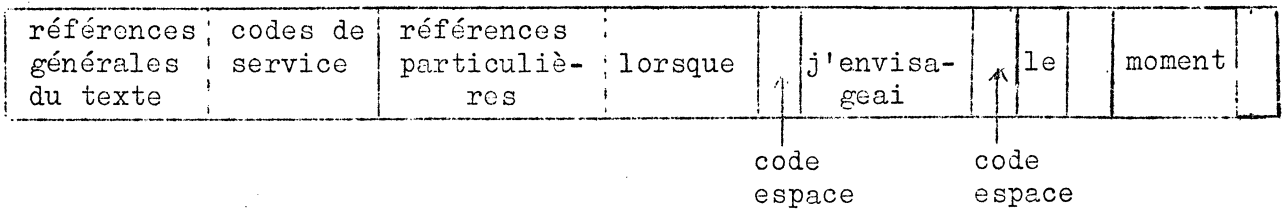


Figure 2

b) Création du fichier. La bande ainsi constituée est placée sur la poinçonneuse télébande. A partir de celle-ci le texte sera automatiquement perforé sur cartes, à raison d'un mot par carte, les références étant automatiquement reportées sur chaque carte et les codes de service provoquant le changement de cartes et les changements de codes nécessaires.

c) Le fichier de cartes perforées sera alors traduit, c'est-à-dire que sur chaque carte seront imprimés automatiquement les divers renseignements perforés sur la carte.

d) A cette étape, une vérification peut être effectuée, par exemple en imprimant sur la tabulatrice les différents mots dans l'ordre du texte.

e) Classement. Se fait en 2 temps :

- Classement, en trieuse, des différentes cartes mots dans un ordre alphabétique rigoureux.

- Classement dans l'ordre alphabétique définitif, les variantes orthographiques et les différentes formes flexionnelles étant rangées sous la forme de référence. Pour cela, à l'aide de l'interclasseuse et de la reproductrice, on perforera sur chaque carte mot un numéro matricule correspondant au mot. Un tri numérique de ces matricules donnera au fichier l'ordre alphabétique définitif.

f) Impression. On imprime alors en tabulatrice le fichier des cartes-mots suivis de leurs références dans le texte. Il est possible d'imprimer simultanément à côté du mot son numéro d'ordre dans la liste et le nombre de ses emplois (de cartes mots dans le texte).

g) Index de fréquences : en employant une poinçonneuse récapitulatrice connectée à la tabulatrice, on élabore, parallèlement à l'impression du fichier, des cartes fréquences portant le nombre total d'emplois du mot. Ce nouveau fichier, après reclassement, permet d'obtenir une liste dégressive de la fréquence d'emploi des mots par exemple.

h) Index des rimes. On peut également élaborer des index de rimes :

- soit en sélectionnant les cartes mots-rimes lors de la création du fichier et en les affectant d'un code particulier dans un emplacement déterminé à l'aide de la reproductrice.

- soit lors de la perforation du texte sur bande, en plaçant une perforation spéciale après le dernier mot du vers.

Différents tri permettront d'imprimer par exemple :

- Les différentes rimes d'un texte

- Les couples de rimes d'un texte

- Les index de rimes d'après l'ordre des initiales, des finales, etc...

Temps d'exécution

Soit à indexer les mots d'une tragédie classique de 1.900 vers, soit environ 18.000 mots.

- Etablissement de la bande		18 heures
- Perforation des cartes		3 heures
- Traduction des cartes		4 heures 30
- Impression de contrôle		2 heures
- Classement alphabétique		7 heures
- Immatriculation		
interclassement	1 heure	
perforation	2 heures 30	
sélection	1 heure	
	4 heures 30	4 heures 30

a reporter: 39 heures

	report :	39 heures
- Classement numérique		3 heures
- Impression définitive		2 heures
	TOTAL	44 heures

Une impression dans l'ordre alphabétique total réduirait le temps à 37 heures.

Mais l'avantage considérable de cette technique est d'être très aisément cumulative. Par exemple pour réaliser un index général d'un auteur à partir des index des différentes oeuvres, il suffira d'interclasser les différents fichiers des différentes oeuvres.

Soit un auteur de 10 tragédies, soit 180.000 mots. L'indexage de chaque pièce a demandé 47 heures, soit au total 470 heures. La préparation d'un index général demandera en plus :

interclassement	12 heures
impression	20 heures
	32 heures pour 180.000 références

En mécanographie le travail le plus long et le plus onéreux est la préparation du fichier de base. Mais une fois ce fichier élaboré, il est possible de l'utiliser à des fins les plus diverses et de trier suivant un nombre considérable de critères. Alors que les méthodes traditionnelles imposent au chercheur d'établir une copie de son fichier chaque fois qu'il envisage un regroupement suivant un nouveau critère, une carte mécanographique peut être reclassée excessivement rapidement suivant tel ou tel critère. Les copies de fichiers ou de parties de fichier peuvent être obtenues très facilement pour des applications particulières et le nombre et la nature des critères de chaque mot peuvent être variés sans difficultés et pratiquement à l'infini pour chaque utilisation. On pourrait très bien envisager des fichiers de base établis une fois pour toutes et utilisés dans des laboratoires différents pour des applications propres à chacun d'eux.

La description donnée ici ne l'a été qu'à titre d'exemple. Il est évident qu'une foule d'autres applications et utilisations peuvent être envisagées dans le même domaine. Cette exploitation, conçue avec un petit matériel, pourra parvenir à son plein épanouissement le jour où elle pourra être transposée sur gros matériel, le Gamma 60 par exemple.

Nous allons maintenant examiner 2 solutions proposées par la Compagnie des Machines BULL à un même problème de mécanisation documentaire, l'une utilisant du matériel classique, l'autre beaucoup plus

intégrée, mais aussi évidemment plus onéreuse, faisant appel à un grand ensemble à traiter l'information, le Gamma 60.

Caractéristiques principales du centre de documentation à mécaniser.

La littérature à classer comporte environ 12.000 documents par an. Soit, pour une documentation portant sur 10 ans, 120.000 documents. Chaque document sera représenté par un article d'un modèle normalisé portant toujours les seules caractéristiques essentielles du texte conservé. La codification permettant d'aboutir à un article en partant d'un document doit exiger le minimum de travail intellectuel. On admet qu'un analyste du centre de documentation a déterminé la liste des mots-clés du document. Ce sont ces mots qui doivent figurer dans "l'article document". La codification de ceux-ci doit être rapide et automatique, de façon à n'exiger aucune formation spécialisée des personnes chargées de l'accomplir et que ces personnes ne puissent faire aucune erreur dans l'application ou l'interprétation du code.

La sélection doit être non seulement automatique, mais encore rapide, complète et infaillible, quel que soit le thème de sélection proposé, tout en correspondant d'aussi près que possible à la "question" posée, dans le cadre du système d'analyse choisi. On s'efforcera notamment de ne pas être gêné par les synonymes du vocabulaire.

SOLUTIONS PROPOSEES

Solution petit matériel

La première solution proposée comprend le matériel suivant :

- 1 machine de perforation
- 1 poinçonneuse reproductrice duplicatrice - vitesse 120 cartes/minute
- 1 interclasseuse - vitesse 250 cartes/minute
- 1 traductrice

La carte : Les cartes seront des cartes perforées ordinaires 80 colonnes. Chaque document sera représenté par une seule carte. Celle-ci devra porter les renseignements suivants :

- Année de parution

- Revue-page

- Numéro de document

- Langue

- Numéro de l'analyste

- Mots clés caractérisant le document.

Les mots clés : L'analyste rédige le résumé du document et souligne les mots clés. Dans l'exemple traité, le nombre de mots clés caractérisant un document ne dépasse pas 22, les sujets étant très spécialisés.

Le nombre total de mots clés existants, constituant la richesse du vocabulaire, sera prévu élevé.

Codification des mots clés : Il était impossible de prévoir sur la carte la transcription en toutes lettres des mots clés selon le code alphabétique normal. L'encombrement ne permet pas de mettre 22 mots sur une carte de 80 colonnes. La solution d'un nombre variable de cartes par document était également à rejeter en raison de la longueur et de la complexité de la recherche dans de telles conditions. En outre, le problème des synonymes n'aurait pas été résolu et la longueur variable des termes aurait rendu l'exploitation très difficile. L'affectation d'une zône à un mot est aussi inacceptable en raison de la place demandée. Les renseignements bibliographiques occupant 14 colonnes, il ne restait que 66 colonnes pour enregistrer en code 22 mots clés.

La codification choisie fait appel à un système entièrement différent de ceux utilisés habituellement. C'est un code original à deux dimensions. On attribue à chaque mot une combinaison numérique composée de 9 chiffres s'écrivant dans un carré de 3 chiffres de côté

X X X

X X X

X X X

Les chiffres utilisés sont 0 9

Dans cette combinaison, les 3 nombres de 3 chiffres lus verticalement dans chaque colonne doivent être tels que le même chiffre ne figure jamais 2 fois et qu'en lisant le nombre de haut en bas on aille du chiffre le plus faible au chiffre le plus élevé.

Exemple : 1
3
5

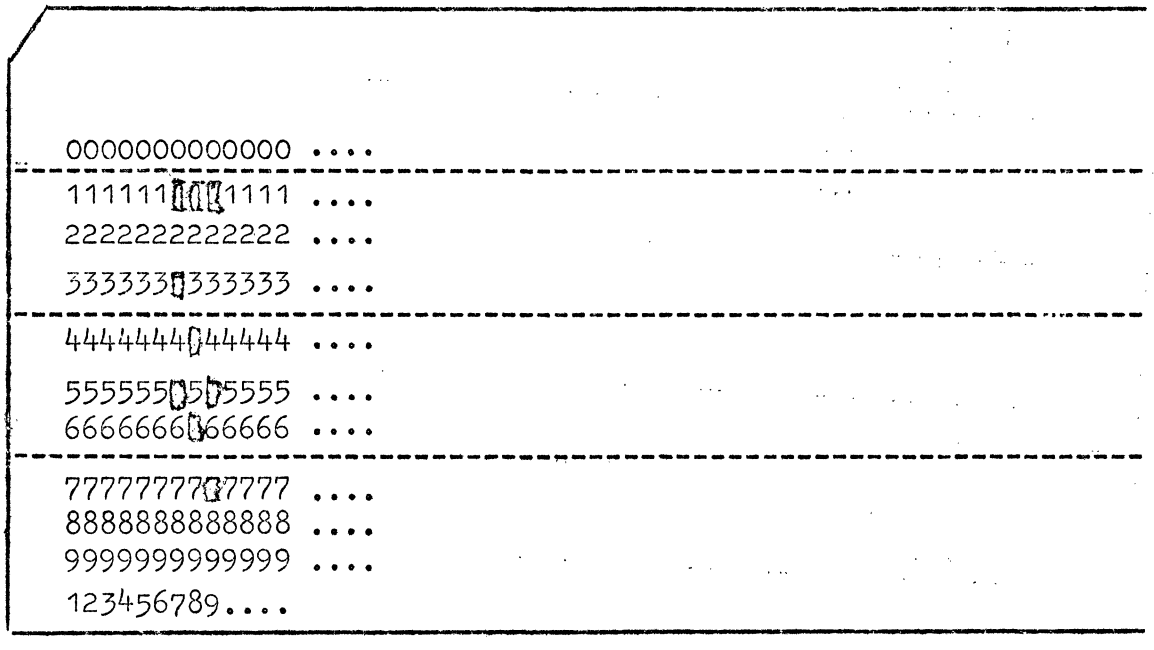
De fait, étant donnée la configuration des cartes perforées et compte tenu de ce que chaque chiffre différent a une position géographique verticale différente sur la carte, la combinaison représentant un mot occupera 3 colonnes de carte perforée. Par exemple:

1 1 1

3 4 5

5 6 7

sera représenté par :



Pour des raisons de sécurité d'exploitation, on a préféré interdire la présence de 2 perforations contigües dans la même colonne

1

2 interdit

3

En dressant le tableau des perforations possibles pour chaque colonne

0 2 4 5 6 7 8 9 6

0 3 5 6 7 8 9 5

---- --- --------- etc----

On trouve un total de plus d'un million de combinaisons permettant l'utilisation d'autant de mots clés différents.

Chaque carte est donc constituée de la façon suivante :

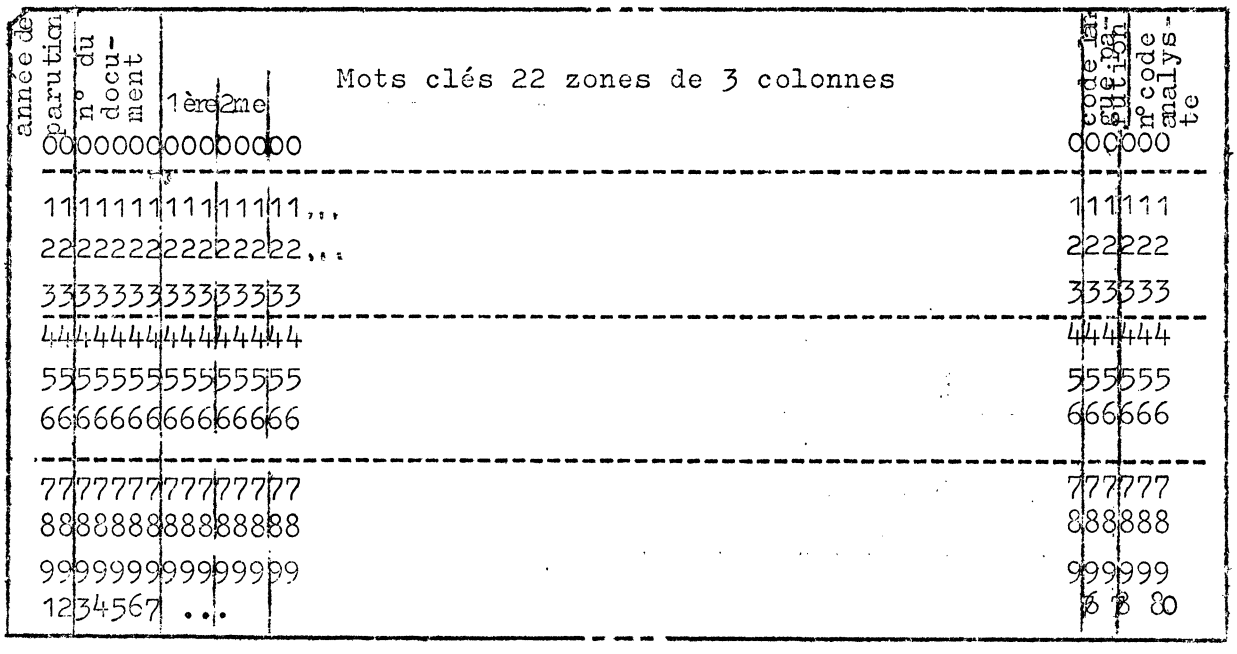


Figure 4

Codification automatique

En dépit de ses avantages, un tel procédé de codification comporterait en lui même un inconvénient grave, qui est sa complexité, si les codes des mots clés devaient être pris dans un dictionnaire ordinaire. En effet, la constitution et l'emploi manuel d'un tel dictionnaire seraient une cause de perte de temps et laisseraient persister

à l'usage des risques permanents d'erreurs, d'autant plus graves qu'elles seraient impossible à déceler, personne ne pouvant prétendre connaître le code par coeur.

Aussi la codification a-t-elle été rendue purement automatique de la façon suivante :

On réalise un fichier de base qui sera le dictionnaire de code. Sur chaque carte de ce fichier on perfore automatiquement une des 1.000.000 de combinaisons du code. Ce fichier de base pourra donc comporter jusqu'à 1 million de cartes. Sur chaque carte, on perfore en outre le mot clé qui correspondra au code porté par la carte. Ce mot clé est perforé en perforation alphabétique normale.

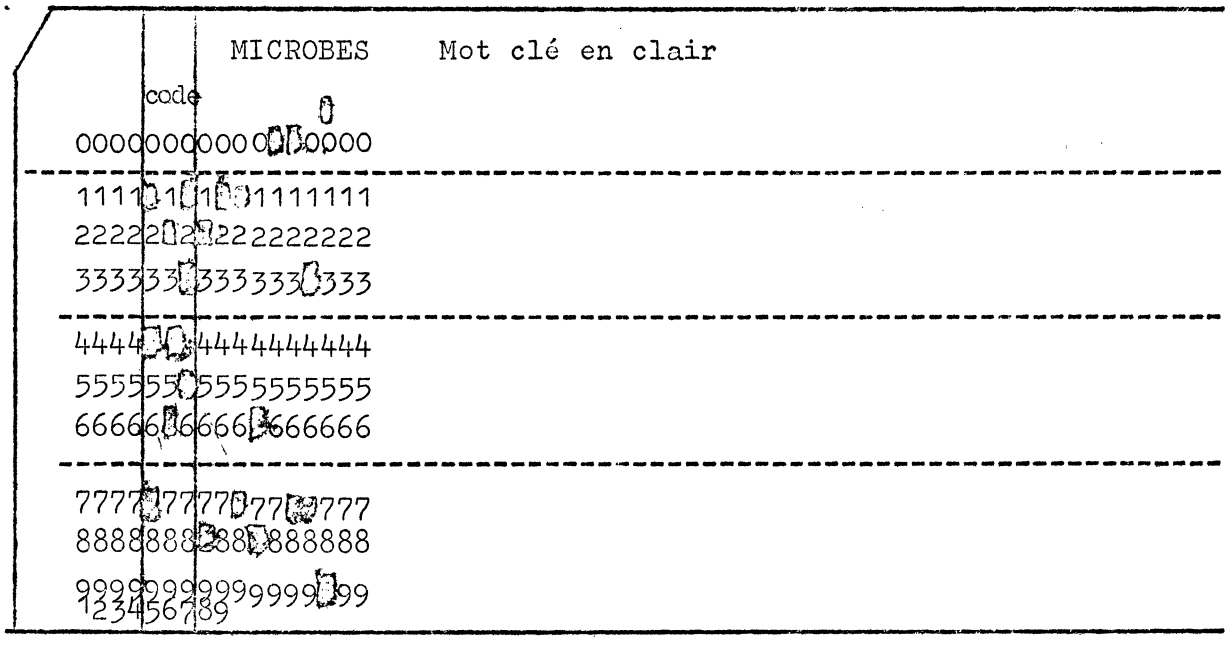


Figure 5

Le problème des synonymes est résolu en perforant 2 codes identiques pour deux mots clés différents mais synonymes.

CREATION DES CARTES "DOCUMENT"

1) Perforation et vérification sur les cartes "document" des coordonnées bibliographiques

2) Extraction manuelle du fichier "code" des cartes portant le code des mots clés recherchés.

3) Report sur chaque carte document du code des mots clés qu'il comporte

4) Réintégration des cartes codes dans le fichier trié par ordre alphabétique.

5) Les nouvelles cartes document sont rangées dans le fichier "Documentation"

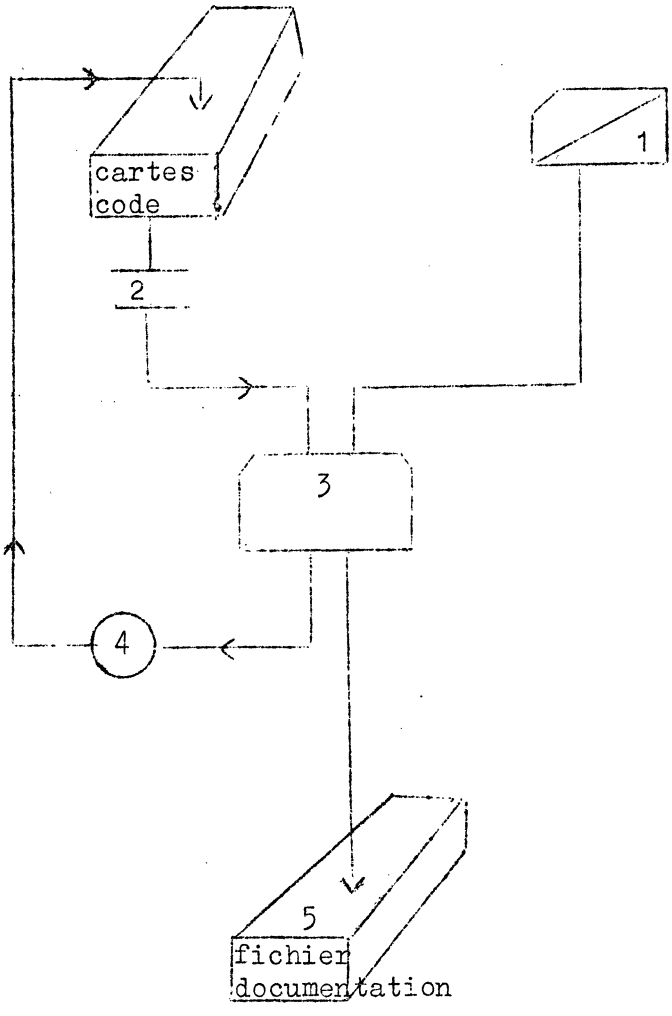


Figure n° 6

On a ainsi constitué un véritable dictionnaire.

Le fichier est rangé par ordre alphabétique des mots clés, chaque carte étant traduite en clair.

Constitution des cartes documentation codées.

Pour coder la carte relative à un document particulier, on extrait du fichier de base les cartes-codes des mots clés du texte. Si l'on ne trouve pas dans le fichier le mot clé cherché, c'est qu'il s'agit d'un mot nouveau. On prend alors la première carte disponible

dans le fichier "codes non affectés" et on perfore en code alphabétique normal le mot clé nouveau qui reçoit ainsi un code.

Toutes les cartes mots clés sont alors placées devant la carte à compléter et passées en PRD. On obtient ainsi, codifiée, la carte propre au document. (cf. figure 6).

Recherche : La recherche s'effectue automatiquement à l'aide d'une interclasseuse. Il suffit d'extraire du fichier la carte code mot clé correspondant au mot clé sur lequel on veut sélectionner et de faire passer en machine le fichier documentation complet, derrière la carte code. La machine extrait automatiquement toutes les cartes contenant ce mot-clé, quelle que soit la zône où il est perforé. Si la recherche porte sur plusieurs critères, on recommencera pour les mots suivants sur les premières cartes sélectionnées. Il est à remarquer que la sélection s'effectue de plus en plus rapidement, le nombre des cartes sélectionnées décroissant très vite.

N.B. Le fichier principal peut être soit en vrac, soit classé par grandes rubriques.

Temps machine : un fichier de 120.000 cartes représentant 10 ans de documentation peut être exploré en entier en une journée de travail de 8 heures.

Il est à noter l'originalité de cette solution étant donné le matériel très restreint utilisé.

Solution Gamma 60

Ici, la machine étant beaucoup plus puissante, la solution proposée sera évidemment plus complète.

Je vais essayer de vous donner, tout d'abord, une très brève description des principes de fonctionnement du Gamma 60.

Cet ensemble électronique se compose tout d'abord d'une mémoire centrale, qui sert en quelque sorte de réserve d'information pour l'ensemble de la machine. Cette information peut être indifféremment des données ou du programme.

La mémoire centrale, à tores magnétiques, peut avoir une capacité de 4096 à 32.768 catènes, un catène ayant une capacité de 4 signes (lettres, signes de ponctuation, etc...) ou 6 chiffres décimaux.

A la mémoire centrale est connectée un organe, le distributeur de programme qui joue le rôle d'agent de circulation de l'information à l'intérieur de la machine.

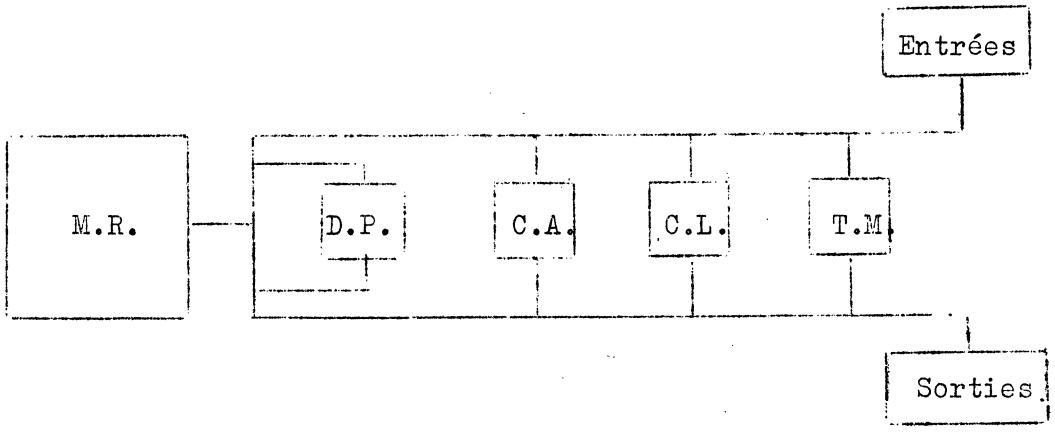


figure n° 7

La machine comporte des éléments de calcul (calculateur arithmétique, logique, comparateur général etc...) et des éléments de stockage de l'information (tambours magnétiques, rubans magnétiques etc...). Les organes d'entrée-sortie (lecteurs et perforateurs de cartes ou de bande perforée, imprimantes) fonctionnent à 300 cartes/minute pour les lecteurs et perforateurs de bande, à 300 lignes/minute pour les imprimantes et à 300 caractères seconde pour la bande perforée. L'originalité du Gamma 60 réside dans les points suivants :

- d'abord, il peut comporter un nombre d'éléments de stockage de l'information (tambours magnétiques, dérouleurs de ruban magnétique) et d'éléments d'entrée sortie aussi grand que l'on veut. On peut donc adapter la taille de la machine à chaque problème particulier.

- ensuite, plusieurs traitements peuvent être exécutés simultanément à l'intérieur de la machine. En effet, son fonctionnement est le suivant : lorsqu'un élément quelconque de la machine est appelé au travail, il puise dans la mémoire rapide les informations dont il a besoin pour commencer à travailler et à partir de ce moment-là fonctionne en autonome. Un autre élément peut donc travailler en même temps que le premier, puisant dans la mémoire rapide ses informations et travaillant en autonome et ainsi de suite. C'est le distributeur de programme qui joue comme nous l'avons dit le rôle d'agent de la circulation pour les différentes demandes de travail des éléments. Le Gamma 60 peut donc, non seulement traiter simultanément des parties différentes d'un même programme, mais encore dans le même temps résoudre des problèmes complètement différents, sans aucune relation entre eux.

- Dans la solution proposée, il s'agit de mettre en mémoire sans aucune codification conventionnelle propre à la machine, toutes les caractéristiques bibliographiques et documentaires d'un document, ainsi que la totalité du résumé analytique du document.

- La machine sera capable de fournir la liste complète de tous les numéros d'immatriculation des documents répondant à un thème ou une question donnée :

- thème défini par mots clés
- bibliographie des travaux d'un auteur
- relevé des travaux publiés dans un périodique
- relevé des travaux publiés entre telle et telle année.

Il y aura donc une grande multiplicité d'explorations possibles :

- par matières
- par auteur
- par origine géographique
- par date d'origine

et, au besoin, tous ces travaux peuvent être combinés entre eux et sans limitation.

- La machine pourra extraire des mémoires et fournir à chaque destinataire une documentation imprimée complète comprenant :

- les données bibliographiques complètes des documents sélectionnés.

- la reproduction des résumés analytiques extraits des mémoires.

- La machine pourra opérer simultanément les travaux de recherche et d'extraction correspondant à un grand nombre de questions différentes (100 ou 1.000 questions différentes simultanément) et la documentation imprimée par la machine sera strictement sélective, sans risques d'interférences pour chacune de ces questions.

Le temps nous manque pour examiner les détails de la solution. Disons simplement qu'elle est entièrement et complètement automatique et que le fichier utilisé pour la recherche documentaire est composé d'articles mots clés comportant les références des documents.

Exemple Typhus
125
132
175 A
.
.
etc...

D'autres articles comportant les différents critères sont également intégrés au fichier.

Exemple Revue numéro ...
156
205
315
.
.
.

La recherche se fait en perforant sur bande en clair les différentes questions avec leurs différents critères. Le procédé de recherche lui même est complètement original.

L'introduction d'un nouveau document se fait en perforant sur bande en clair les différents critères du document : Nom, prénoms des auteurs, année de parution, langue, nom de laboratoire, journal ou revue, numéro de classement systématique, mots clés caractérisant le document, résumé analytique. La bande introduite en Gamma 60, les différents critères sont réorganisés automatiquement et chaque caractéristique du document est intégrée dans le fichier recherche.

A la fin de la recherche, les différentes caractéristiques des documents sélectionnés sont réorganisées à nouveau et l'état obtenu est présenté par demandeur et par numéro de question de demandeur si l'on admet qu'une même personne puisse poser plusieurs questions simultanément. Cet état, présenté avec en face de chaque référence de document sélectionné le résumé analytique de celui-ci peut être envoyé tel quel au demandeur.

Le Gamma 60 utilisé comprend :

- 1 unité centrale
- 2 blocs de mémoire rapide
- 1 lecteur de bande perforée
- 1 imprimante 300 lignes par minute
- 1 tambour magnétique
- 4 dérouleurs de ruban.

Il est évident que si l'on désire traiter simultanément dans le même temps un nombre de questions beaucoup plus grand portant sur un fichier beaucoup plus vaste, il suffit d'augmenter le nombre d'éléments de mémoires et d'entrées sorties de la machine, sans avoir à modifier en rien l'organisation adoptée si le besoin ne s'en fait pas sentir.

Dans un domaine comme celui de la documentation où le volume d'information traité s'accroit chaque jour sans que l'on soit toujours à même de prévoir la rapidité de cet accroissement, un tel ensemble est un outil précieux : son organisation interne lui permet d'être toujours à la mesure des problèmes traités, car il croit en même temps qu'eux. L'accroissement du nombre de sorties et les simultanéités rendent possible l'obtention d'un nombre de résultats toujours plus grand, sans augmenter pour cela le temps nécessaire à leur obtention.

Par contre, l'importance et le coût de ce matériel font qu'une telle solution n'est valable que pour un organisme documentaire excessivement centralisé travaillant pour un très grand nombre d'usagers.

Dans le cadre de cet organisme documentaire centralisé, on peut imaginer qu'en même temps que les travaux "routiniers" de classement et recherche documentaire se poursuivent sur la machine, celle-ci

traite en simultanéité des problèmes de recherche pure, du type dont il a été question tout au long de cette semaine.

Après vous avoir exposé les principes de ces différentes réalisations, j'aimerais aborder maintenant une partie plus technique et essayer de mettre en évidence quelques uns des problèmes constants rencontrés pour le traitement, sur gros calculateurs, de l'information non numérique.

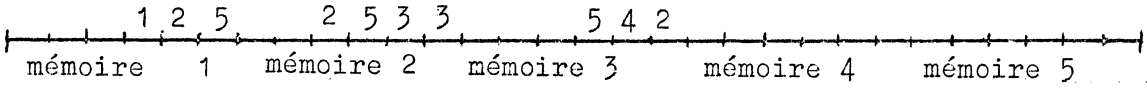
Toujours faute de temps, je me contenterai d'exposer dans le détail un ou deux points précis.

Qui dit traitement sur un gros calculateur dit évidemment enregistrement dans les différentes hiérarchies de mémoire de celui-ci, de l'information à manipuler.

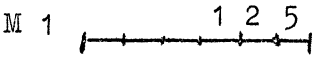
Or, une des différences essentielles entre l'information numérique et l'information non numérique est que celle-ci est en général "continue" par opposition à l'information numérique qui, elle, est discontinue.

En effet, le traitement d'un problème d'analyse numérique par exemple porte sur des nombres ou des tableaux de nombres dont chacun d'eux constitue une entité.

La mémoire rapide d'un calculateur est construite de telle sorte qu'elle soit divisée en mémoires élémentaires ou "mots" de mémoire, de capacité fixe : les organes de la machine sont construits de façon à travailler toujours sur un mot de mémoire - ou plusieurs mots associés, mais toujours en nombre fixe - Par exemple, dans tout ou partie d'un problème, on travaillera sur des nombres occupant 2 mémoires partiellement ou complètement remplies. Par exemple, les nombres 125, 2533 et 542 sont enregistrés à l'intérieur de la machine, chacun dans une mémoire particulière.



Lorsqu'on désire traiter un de ces nombres, on appellera la totalité de la mémoire le contenant en la désignant par son numéro, ou adresse

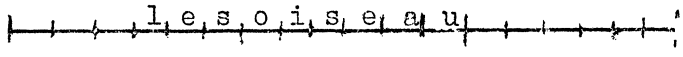


Pour le traitement de l'information littérale, par contre, le problème est différent et plus complexe. Supposons que l'on veuille enregistrer à l'intérieur d'une machine une phrase de texte, pour travailler sur cette phrase ultérieurement.

"Les oiseaux migrateurs, arrivés hier, se sont envolés ce matin".

Admettons que la machine soit telle que l'on puisse introduire, sans transformations ou codes spéciaux, les différents signes utilisés.

On pourrait évidemment enregistrer cette phrase à raison d'un mot ou signe de ponctuation par mémoire

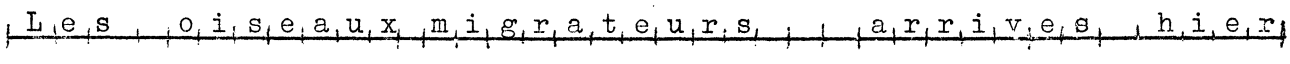


Ici déjà le mot oiseaux ne tient pas dans une mémoire. Il faudrait donc prendre pour chaque mot une capacité de mémoire égale au plus long d'entre eux.

Supposons 26 lettres, par exemple. Si l'on met par exemple 6 lettres par mémoire, chaque mot occupera donc 5 mémoires. Mais la longueur moyenne des mots étant de 6 lettres, on se rend compte de la place perdue.

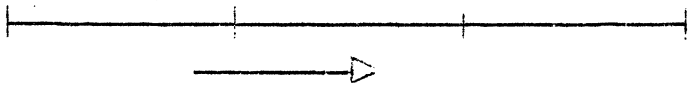
Pour de l'information numérique on pourrait, tenant compte de la précision, prendre une longueur fixe = 2 mémoires.

Mais ici ce n'est pas possible. Il faut donc enregistrer l'information à l'intérieur de la mémoire de façon continue



Mais alors on perd, pour un mot, la notion d'adresse unique et on ne peut, ayant l'adresse d'un mot, en déduire celle du mot précédent ou suivant. Comment reconnaître, aller chercher, un mot à l'intérieur de la suite de mémoires représentant la phrase ? Il serait possible évidemment de dresser une table de tous les mots enregistrés avec l'adresse complète de chacun d'eux (numéro de mémoire et position dans la mémoire du début et de la fin du mot). Mais on imagine sans peine l'encombrement et la longueur d'une telle table.

Une bien meilleure solution est l'écriture d'un court programme spécial, constitué en sous programme, qui sera chargé d'explorer les mémoires position à position, s'arrêtant chaque fois qu'il rencontre un blanc ou un signe de ponctuation et sélectionnant le mot ainsi délimité pour un traitement ultérieur.



Ce sous-programme, utilisé constamment au long du traitement, sera très travaillé au point de vue temps de déroulement.

D'autres sous-programmes généraux seront écrits également pour sélectionner le mot précédant celui que l'on vient de traiter, ou le mot suivant, ou la phrase précédente etc... On aura ainsi toute une série de sous-programmes généraux pour résoudre les problèmes constants au long du programme.

Lorsque des caractéristiques auront été déterminées pour des mots, elles seront enregistrées sous forme numérique à l'intérieur des mémoires. Les mots eux-mêmes pourront être transformés en cours d'exploitation pour faire apparaître certaines de leurs caractéristiques sous une forme plus directement assimilable par la machine.

On arrive ainsi, petit à petit, à une transformation totale de l'information introduite.

Si le principe de cette transformation est basé sur des études théoriques (organigrammes de principe), sa forme elle même ne peut être déterminée que par une connaissance profonde des machines et des techniques de programmation et est du domaine de spécialistes.

Une foule d'autres problèmes d'apparence mineure, mais dont dépend la performance donc le coût du programme, peuvent être résolus de façon similaire.

Mon ambition a été de vous donner ici un aperçu des possibilités des calculateurs électroniques et du matériel à cartes perforées dans le domaine de l'information non numérique. J'ai essayé également de vous faire sentir qu'étant donné le caractère universel de telles machines, elles peuvent s'adapter facilement à n'importe quelle forme de problème. Je voudrais également vous avoir montré qu'il n'est pas toujours nécessaire de pemser gros matériel ou machine à traiter l'information pour résoudre de tels problèmes et que, compte tenu des possibilités financières, le matériel classique, peu onéreux, permet d'apporter des solutions qui, si elles sont parfois moins élégantes que celles apportées par le gros matériel, permettent néanmoins des réalisations intéressantes.

Les machines, telles qu'elles existent actuellement, peuvent traiter les problèmes d'information non numérique. Dans l'état actuel des choses, on ne peut attendre l'aboutissement des recherches de principe sur la représentation des documents et sa mécanisation.

Une mécanisation actuelle, même incomplète, permet de faire face aux problèmes urgents, tout en continuant les recherches de solutions optimales.

POINTS COMMUNS ENTRE LES PROBLEMES POSES

PAR

LA TRADUCTION AUTOMATIQUE ET LA DOCUMENTATION AUTOMATIQUE

J. IUNG *

L'exposé comporte deux parties principales :

Dans la première partie, nous exprimons quelques idées assez générales sur l'identité des problèmes posés par la traduction et la documentation.

Dans la seconde partie, nous développons de manière assez détaillée plusieurs exemples illustrant les méthodes qui ont été utilisées pour résoudre les problèmes posés par la traduction automatique et nous montrons comment ces méthodes conduisent à des résultats intermédiaires, qui sont du plus grand intérêt pour la documentation automatique.

En conclusion, nous donnons les raisons qui nous conduisent à penser que la réalisation de la traduction automatique précèdera celle de la documentation automatique.

Nous allons tout d'abord commencer par définir ce que nous entendons par traduction automatique et documentation automatique.

La traduction est le passage d'un texte écrit dans une langue à un texte écrit dans une autre langue, le sens du texte initial devant être respecté. La traduction effectuée par des êtres humains exige souvent dans le domaine scientifique la collaboration d'un linguiste et d'un scientifique. Dans le cas de la traduction automatique, nous supposerons que le travail est complètement effectué par la machine, c'est-à-dire qu'il n'existe ni "pré-éditeur", ni "post-éditeur".

La documentation automatique est le passage d'un texte écrit dans une langue à un texte écrit généralement dans la même langue, mais pas nécessairement; ce second texte doit être une "réponse" au premier

---

* Groupe d'Automatisation des Fonctions Documentaires,
Centre d'Etudes Nucléaires de Saclay.

et, dans une certaine mesure, lui être complémentaire. Comme dans le cas de la traduction automatique, nous supposerons que tout le travail doit être effectué par la machine.

Pour bien fixer les idées, nous allons donner un exemple de question abrégée et un exemple de réponse à cette question :

Question abrégée : Dans le cadre de l'avant-projet de construction d'une pile atomique chaude, on est amené à se demander si la conductivité thermique des combustibles nucléaires céramiques dépend de l'irradiation.

Réponse abrégée : La conductivité thermique des corps suivants : oxyde d'uranium etc. est donnée en fonction de l'irradiation dans les courbes suivantes : ... On voit que les résultats obtenus n'étaient pas reproductibles. On explique pourquoi (nature différente de la structure des différents échantillons solides et, par conséquent, des résultats de l'irradiation). On fait quelques considérations sur le choix du combustible nucléaire dans les piles chaudes, sur les problèmes d'échange thermique posés, et également sur les méthodes de mesure de la conductivité thermique.

On notera que la machine doit fournir un texte réponse et non une liste de références. A partir de la réponse donnée, on peut poser d'autres questions, en particulier, demander les "sources" des résultats et, par conséquent, imaginer un "dialogue" homme-machine.

C'est sur les bases des définitions précédentes que nous allons développer notre comparaison.

Nous remarquons immédiatement un point commun entre les deux opérations : dans l'un ou l'autre cas, nous partons d'un texte écrit pour aboutir à un autre texte écrit, c'est-à-dire nous utilisons la représentation graphique du langage; mais, dans la transformation qui s'effectue, c'est "l'idée", le "sens" ou, comme on dit en français, "l'esprit" des textes qui joue le rôle principal.

Le point commun étant le langage, nous allons rappeler brièvement les différentes caractéristiques, ou plus exactement les différentes interprétations du langage.

On peut d'abord considérer le langage comme une suite de symboles acoustiques ("phonèmes") ou graphiques ("lettres", "mots"). On peut étudier de manière statistique l'apparition ou la répartition de ces symboles dans une langue. Suivant le point de vue adopté, on arrive à la théorie de l'information et à la linguistique statistique.

On peut adopter un point de vue philosophique et considérer le langage comme l'instrument qui permet la pensée. On peut encore dire

que le langage est le moyen qui permet la communication de la pensée entre les individus, et c'est surtout cet aspect du langage qui nous intéressera ici.

Il faut noter enfin que le langage est le fait, la propriété, la "sécrétion" de l'ensemble des individus. Des interprétations précédentes, nous retiendrons principalement que le langage permet la communication de la pensée entre les individus, mais qu'il exprime l'ensemble de la société.

Nous allons chercher maintenant les raisons pour lesquelles la communication de l'information se fait mal, ou ne se fait pas, alors que le langage, parlé ou écrit, devrait permettre facilement cette communication.

Il y a tout d'abord une raison banale et évidente, qui est la mauvaise diffusion matérielle des informations, volontaire (secrets militaire ou industriel) ou involontaire (mauvaise organisation). Nous n'en parlerons pas ici, et nous supposerons par la suite que la diffusion des informations est parfaite.

La nature différente des langues utilisées constitue un second obstacle à la transmission de l'information; ceci pose le problème de la traduction.

Mais, dans le cas idéal de la diffusion parfaite de textes en une langue connue, le problème de la communication de l'information reste posé. Trop souvent, on a expliqué de manière simpliste le problème documentaire en prétendant que tout serait résolu si chaque individu avait le temps de tout lire. Cette affirmation toute gratuite a conduit les constructeurs de machines à penser que le problème documentaire serait résolu le jour où il existerait sur le marché des dispositifs pourvus de grandes mémoires pouvant être explorées très rapidement dans leur totalité. En fait, une telle machine serait tout juste capable de restituer sans aucune modification l'information qu'on aurait introduite : dans le cas de la machine à traduire, la machine enregistrerait des traductions déjà faites par des êtres humains, et fournirait à la demande une des traductions d'un des textes qui se trouvent dans la mémoire; dans le cas de la machine documentaire, la machine enregistrerait un ensemble de textes, et fournirait à la demande un des textes auxquels aurait déjà pensé l'utilisateur de la machine, ceci dans le cas le plus favorable. On voit mal l'intérêt de telles machines.

En réalité, je pense qu'il faut chercher ailleurs l'origine du problème documentaire. A mon avis, la mauvaise communication des "idées" provient du fait que chaque lecteur d'un document adopte un point de vue particulier; cette "attitude" lui permet de saisir certains aspects, certaines parties de l'information transmise, mais lui masque automatiquement d'autres parties. On peut dire à la limite que chaque individu est incapable d'exploiter complètement l'information qui lui est communiquée.

Nous allons nous arrêter un peu sur cette question qui est le fondement même de la recherche documentaire, et qui est l'origine de tous les échecs de mécanisation. Pour mieux comprendre le problème, nous allons "détailler" les différentes opérations de la production du "document" scientifique.

Au point de départ, nous avons un scientifique qui écrit un article; il rédige ce document à l'intention d'autres personnes (autres scientifiques, ou lui-même à une autre époque); pour cela, il est obligé de supposer connu de la part de ses correspondants un certain nombre de faits qu'il ne va pas répéter. D'autre part, il exprime un certain nombre d'idées, et ces idées dépendent de sa propre expérience, de ses propres connaissances; or, cette expérience, ces connaissances, ne lui appartiennent pas, pas plus que le langage qu'il utilise ne lui appartient.

Au point d'arrivée, nous avons un autre scientifique; cette autre personne "lit" l'article et, de ce fait, replace le texte dans un autre contexte, qui dépend de l'expérience et des connaissances propres du lecteur. Ce "contexte" peut d'ailleurs être très proche, ou très éloigné du point de départ.

Mais dans aucun des cas nous ne pouvons séparer le texte de l'expérience de la personne qui l'écrit ou le lit. Si nous le faisons, nous laissons échapper toute la signification du texte et, dans ce cas, il est impossible de "transformer" le texte en tenant compte de sa signification, comme ce doit être le cas en traduction ou en documentation automatique. Un texte scientifique ne prend de sens que replacé dans un certain contexte; comme il y a une infinité de manières de replacer le texte dans le contexte scientifique, et comme le contexte scientifique évolue en fonction du temps, on peut dire que chaque texte peut avoir une infinité de sens et que chaque sens dépend du contexte considéré et de l'époque considérée. ** Il est évidemment impossible à un être humain isolé de "saisir" tous ces sens, mais on peut supposer qu'un ordinateur de type spécial pourrait le faire et, par conséquent, exploiter de manière beaucoup plus efficace l'information.

Il y a cependant une condition à cela : c'est que le "contexte" dont nous avons parlé, et qui se confondait avec "l'expérience", les connaissances de l'ensemble des êtres humains, soit équivalent à l'ensemble des textes écrits, seul contexte possible pour une machine. Ceci reste à démontrer.

Ces problèmes contextuels sont apparus clairement pour la première fois lorsqu'on a entrepris d'automatiser réellement, c'est-à-dire complètement, la traduction. Aucun essai d'automatisation complète de la documentation n'ayant été entrepris à cette époque, ce sont les spécialistes de la traduction automatique qui, après les tâtonnements du début, s'aperçurent que les problèmes de la traduction ne pouvaient être résolus qu'en tenant compte du sens du texte, et que le sens dépendait, en définitive, d'un contexte plus ou moins étendu. Voici, par

---

** Tout ce que nous venons de dire au sujet d'un texte scientifique reste valable pour le "mot" des linguistes : un mot n'a de sens que comparé à d'autres mots à une époque donnée, ou comparé à lui-même à une autre époque.

exemple, quelques problèmes qui se posent lors de la traduction : résolution des ambiguïtés de lexique (temps = time, Zeit; temps = weather, Wetter); traitement des expressions idiomatiques (marché "noir", idées "noires"; se faire du mauvais "sang", bon sang !), traitement des tournures elliptiques ("magnetic coil" = bobine produisant un champ magnétique ou, plus exactement, bobine parcourue par un courant produisant un champ magnétique); traitement des homographes (son = his, son = sound, son = bran).

Dans ce qui va suivre, nous allons développer de manière assez détaillée deux exemples de méthodes de traduction automatique.

Le groupe de Cambridge "Cambridge Language Research Unit" nous fournira le premier exemple pratique montrant comment on peut résoudre certaines ambiguïtés en tenant compte d'un contexte plus ou moins lointain. Il faut noter que ce groupe utilise les mêmes méthodes pour la traduction et la documentation automatiques. (1) Le groupe de Cambridge utilise comme base de départ un "thesaurus", c'est-à-dire un dictionnaire de synonymes. A chaque groupe de synonymes correspond une idée, une notion, ou plus exactement un point de vue qui peut, suivant les cas, être, ou ne pas être, défini par un mot. Pour le groupe de Cambridge, ce sont ces groupes de mots qui jouent le rôle principal dans la traduction automatique et ce sont les "têtes de chapitre" qui se correspondent langue à langue, et non les mots qui constituent ces "chapitres".

Nous allons donner un exemple d'utilisation du thesaurus. Auparavant, nous noterons que les chercheurs de Cambridge, pour gagner du temps, ont adopté le thesaurus de Roget (2), fait au siècle dernier, et qu'ils ont séparé, au moins au début, l'analyse des rapports sémantiques de celle des rapports grammaticaux. Il faut également remarquer que le thesaurus de Roget n'est absolument pas scientifique, et qu'il serait nécessaire de le revoir complètement si on voulait réellement l'utiliser comme base de travail.

Ceci étant dit, prenons un exemple précis, et voyons comment peut être utilisé le thesaurus d'après l'école de Cambridge.

Soit la phrase suivante :

** In the magnetic mirror approach, a straight section of tube is wound with external fiels coils so arranged to produce an axial magnetic field which is weak in the central region, but strong at the two ends.

La suite des opérations est la suivante :

On cherche dans la partie "dictionnaire" du thesaurus les mots significatifs du point de vue du lexique (substantifs, verbes, adjectifs, adverbes) et on note les numéros des têtes de chapitres figurant à la suite de chaque mot.

** Cette phrase est tirée de l'ouvrage de BISHOP : Project Sherwood (p. 52, première phrase du premier paragraphe).

Ainsi pour les vingt mots soulignés de la phrase précédente, on trouve 80 têtes de chapitres.

Par exemple au mot field correspondent les huit notions suivantes : ***

034 - Notion de temps, ayant un rapport avec la notion d'effet, de but
Mots voisins anglais : occasion, opportunity, room, scope...,
Mots français voisins : occasion, place...,

180 - Notion d'espace abstrait indéfini.
Mots anglais voisins : scope, range, latitude, expansion, way...,
Mots français voisins : Champ, espace, étendue...,

181 - Notion d'espace abstrait fini
Mots anglais voisins : region, close, court...,
Mots français voisins : enclos, cour...,

344 - Notion de matière fluide spécifique : plaine (opposé à lac, océan...)
Mots anglais voisins : pasturage, park, lawn, green...,
Mots français voisins : champ, parc, pelouse...,

371 - Notion de matière organique, vie, agriculture...
Mots anglais voisins : meadow, garden....,
Mots français voisins : champ, prairie, jardin...,

625 - Notion de volonté individuelle, perspective, affaires.
Mots anglais voisins : capacity, sphere, orb, line...
Mots français voisins : domaine, sphère, compétence...,

728 - Notion de volonté individuelle, antagonisme, lutte.
Mots anglais voisins : arena, scene, platform...,
Mots français voisins : théatre (d'un exploit), scène (d'une action)

780 - Notion de propriété
Mots anglais voisins : land, plantation, domain...,
Mots français voisins : champ, terrain...,

Les 80 numéros correspondant aux différentes notions que peuvent représenter les 20 mots soulignés de la phrase sont ensuite regroupés et classés dans l'ordre numérique. Cette opération permet immédiatement de repérer les numéros qui reviennent plus d'une fois; ces numéros représentent les notions communes à plusieurs mots de la phrase et donnent une première indication sur le "sujet" de l'article et, par suite, sur le choix entre les notions différentes correspondant à un seul mot.

*** Sans compter les expressions idiomatiques qui contiennent le mot "field".

Dans le cas précis de cette phrase, 5 numéros reviennent deux fois :

Il s'agit de :

154 - Notion de cause (introduite par les mots : produce et end)
181 - Notion d'espace abstrait (introduite par les mots field et region)
222 - Notion d'espace : dimensions (introduite par les mots : axis et center)
311 - Notion d'espace : mouvement (introduite par les mots : coil et wind)
248 - Notion d'espace : mouvement (introduite par les mots : coil, wind)

On note immédiatement que 4 des 5 notions précédentes appartiennent à une même classe : la classe espace (le thesaurus de Roget est divisé en 6 classes générales) et que ceci suffit à orienter automatiquement les choix.

L'étude des antonymes peut se faire également de manière automatique : dans le cas du thesaurus de Roget non modifié, les antonymes sont repérés par des numéros qui se suivent.

Ainsi dans la liste des 80 notions précédentes, on trouve :

180-181 - espace indéfini, espace fini (introduit par region et field)
245-246 - forme droite, forme courbe (introduit par straight et coil)
278-279 - mouvement dirigé, mouvement dévié (introduit par straight et wind)
159-160 - force, faiblesse (puissance) (introduit par strong et weak)
390-391 - épicé, insipide (goût) (introduit par strong et weak)

Enfin, on peut tenir compte de mots qui reviennent plusieurs fois dans la phrase; c'est le cas du mot magnetic, auquel peuvent correspondre les notions :

157 - Notion de puissance (classe : relations abstraites)
175 - Notion de : influence (classe : relations abstraites)
288 - Notion de : attraction (classe : espace)
615 - Notion de : persuasion (classe : volonté)

Si on combine les résultats obtenus sur les synonymes, les antonymes et la répétition des mots, on voit que les catégories principales mises automatiquement en évidence (relation abstraite, cause : numéros 153 à 179 dans Roget; et espace : numéros 180 à 315 dans Roget) suffisent déjà à faire un choix dans les cas que nous avons détaillés; pour field, élimination des notions 034, 344, 625, 728, 780 (restent 180 et 181); dans le cas de strong, weak, élimination de 390 et 391 (notions ayant trait au goût).

En résumé, on peut dire que dans cette première phase des opérations, on replace les notions exprimées par la phrase dans une ou plusieurs classes assez générales, mais cependant beaucoup plus limitées que l'ensemble des connaissances, ce qui permet d'éliminer tout d'abord à coup sûr un ensemble de notions représentées par les mots, mais n'ayant rien à voir avec le sujet traité.

Nous ne décrirons pas en détail les opérations suivantes effectuées par le groupe de Cambridge pour préciser encore les notions traitées dans un texte. Disons simplement qu'elles sont toujours basées sur les numéros représentant les ensembles de mots groupés sous une même notion : le principe consiste à faire les intersections de ces ensembles importants mis en évidence dans la première partie des opérations, de repérer les mots communs, de prendre les têtes de chapitre correspondant à ces mots communs, de faire les intersections de nouvelles têtes de chapitre avec les anciens, etc. En fait, la méthode conduit à une représentation mathématique ("lattice") et à l'établissement d'un certain nombre de règles permettant de limiter et d'orienter l'exploration de la machine.

L'essentiel à retenir est que dans aucun cas la machine ne s'en tient au texte. Sa première opération consiste à replacer le texte dans un contexte plus général (dans l'exemple précédent, tous les mots ayant trait à une action dans l'espace). Ensuite, la machine explore le contexte qu'elle a déterminé, se forgeant pour ainsi dire un vocabulaire pour chaque texte qu'elle a à traduire ("microglossaire"). Ce n'est que lorsque ces opérations sont terminées que la machine passe au stade suivant, qui peut être, soit la traduction, soit la recherche documentaire.

On peut remarquer que la phrase prise comme exemple précédemment peut très bien servir de base pour la recherche de tous les documents ayant trait aux appareils utilisant une configuration de miroirs magnétiques pour confiner un plasma : la définition de la recherche par une phrase entière et l'utilisation d'un thesaurus permet d'éliminer toutes les fausses combinaisons qui seraient apparues si on avait simplement défini la recherche par les "mots-clés" magnetic et mirror : en effet, cette association de mots peut désigner, en particulier dans la littérature russe, certains dispositifs d'optique corpusculaire; d'autre part, ces deux mots, associés à d'autres, peuvent très bien désigner des textes ayant trait par exemple à la magnéto-optique.

Dans le premier exemple, on a abordé le problème de la traduction automatique en tenant compte uniquement du lexique, en laissant de côté la syntaxe et la morphologie. Or, il est facile de voir que les correspondances de lexique et les correspondances grammaticales (morphologiques et structurales) ne peuvent être étudiées séparément. Ainsi, dans l'exemple étudié précédemment, on a trouvé que les mots produce et end appartenaient tous deux à une même notion (154 de Roget : notion de fin, but, dessein). Or, le mot "end" exprime ici une notion spatiale (extrêmité, bout) *: la fausse combinaison trouvée aurait pu être évitée si on avait tenu compte de la fonction du mot produce (verbe et non

---

* Cette notion est bien indiquée dans le Roget, mais assez curieusement, uniquement pour l'expression "from end to end".

substantif), et de la structure de la phrase : "end" se rattache à la dernière proposition de la phrase.

Avant de passer à la description d'une deuxième méthode de traitement d'un texte, nous allons encore donner un court exemple, emprunté à Cëccato (3), mettant en évidence l'étroite dépendance du lexique, de la grammaire et de la sémantique.

Soient à traduire les phrases suivantes, écrites dans des langues n'ayant pas d'article (latin et russe) :

Librum lego et scriptorem laudo

Я читаю книгу и хвалю писателя.

La traduction correcte est : Je lis un livre et je fais l'éloge de l'auteur (ou : de son auteur).

Or, une machine, tenant compte uniquement des mots de la phrase et de la structure de cette phrase, traduira :

Je lis un livre et je vante un auteur.

Cette traduction fausse provient du fait que la machine ignorait que les auteurs écrivent des livres, et le mot "livre" de la première proposition appelait automatiquement l'existence d'un auteur, c'est-à-dire de "l'auteur" mentionné dans la deuxième proposition. L'utilisation du "nuage sémantique" (mots groupés autour d'une notion) entourant le mot dans les têtes de chapitre du thesaurus de Roget aurait suffi à éviter l'erreur de la machine (author et book appartiennent à la notion 593 : communication des idées par des livres, au même titre d'ailleurs que : volume, tome, issue, paper, et : writer, journalist...)

Nous allons maintenant développer en détail une autre méthode de traitement des textes, plus empirique que la méthode du thesaurus, mais aussi plus proche de la réalité. Cette méthode est celle de la Rand Corporation : elle a été décrite l'an dernier par HAYS et HARPER à la Conférence de l'UNESCO (4); et nous allons reprendre ici leur description, et également leur exemple. Nous allons montrer que leur méthode de traduction automatique conduit à une représentation par diagrammes, très proche de celle utilisée par LEROY et BRAFFORT dans leur expérience de documentation automatique de Francfort (5).

Les chercheurs de la Rand Corporation traitent donc uniquement des problèmes de traduction automatique. Ils ont choisi une méthode d'approche purement empirique. Ils ont commencé par choisir un sujet d'étude ("corpus") assez limité : l'ensemble des textes de la physique et un couple de langues unique : russe vers l'anglais. Ils ont ensuite abordé directement la résolution du problème de la traduction automatique; après

avoir préparé un glossaire du vocabulaire rencontré et indiqué pour chaque mot russe un équivalent anglais et des indications d'ordre morphologique, un premier programme de traduction sur machine a été préparé, et une première traduction, mauvaise, produite. Cette première traduction a alors été donnée à des physiciens connaissant parfaitement les deux langues, qui ont indiqué toutes les erreurs, leur origine, et les moyens de les éviter. Ceci a permis de réaliser un nouveau programme, une nouvelle traduction, meilleure que la première, d'où un nouvel examen de la part des linguistes, et ainsi de suite, les traductions devenant de plus en plus correctes, le corpus de plus en plus étendu, le vocabulaire de plus en plus riche, la nouvelle grammaire de plus en plus complète et surtout de mieux en mieux adaptée aux exigences de la machine. Les programmes, de leur côté, sont devenus de plus en plus complexes.

. Les résultats auxquels sont parvenus les chercheurs de la Rand sont les suivants : l'analyse du texte de départ est la partie la plus importante et la plus difficile du travail. Cette analyse doit être fondée sur les "dépendances" des mots entre eux; deux sortes de dépendances sont utilisées : les dépendances grammaticales, lorsque la flexion d'un mot dépend d'un autre mot (ex. : un charmant garçon, une charmante fille), et les dépendances sémantiques, lorsque la signification d'un mot est associée dans une phrase à la signification d'un autre mot. Une fois qu'un dictionnaire des dépendances sémantiques a été empiriquement établi, il est possible d'entreprendre l'analyse de la structure des différentes phrases ou, plus exactement, de faire apparaître, à partir de la suite linéaire des mots (ou plus exactement de la suite linéaire dans le temps des phonèmes, imposés pour des raisons pratiques) la structure réelle des phrases. ⁂

Nous allons maintenant développer un exemple illustrant la méthode automatique de détermination des structures, lorsque les dépendances grammaticales et sémantiques sont établies une fois pour toutes. La suite des opérations est la suivante :

On numérote les mots de la phrase de 1 à n, en respectant l'ordre de ces mots.

Au début, on admet que chaque mot précède le suivant :

1 p 2 p 3 p 4 ... p N (1 précède 2 précède 3 etc.)

Ensuite, la machine étudie les couples de mots qui se précèdent l'un l'autre, les compare aux types de dépendances qui se trouvent dans sa mémoire, c'est-à-dire, elle cherche à établir les dépendances sans la phrase.

⁂ Les idées développées ici par les chercheurs de la Rand s'inspirent des idées de ADIUKIEWICZ, OSWALD et HARRIS.

Pour mener à bien ce travail et déterminer la structure de la phrase, la machine doit obéir aux deux règles suivantes :

I - X précède Y si :

1. le numéro de X est inférieur à celui de Y
2. tous les numéros entre X et Y dérivent de X ou de Y

   Remarque: W dérive de Z si:
   W d X d Y d Z c'est-à-dire:
   (W dépend de X qui dépend de Y qui dépend de Z)

3. X et Y ne dérivent pas l'un de l'autre
4. au moins l'un des deux, X ou Y, est indépendant.

II - X dépend de Y si :

1. le couple g (X) - g (Y), dans l'ordre, est inscrit dans les tables de dépendance
2. X p Y ou Y p X
3. X n'a pas été trouvé auparavant dépendant d'autre chose.

Premier exemple :

Je	vois	une	maison	rouge.
1	2	3	4	5

- Première exploration :

1°) On examine le couple 1 - 2 : on trouve 1 d⟶2
2°) On examine le couple 2 - 3, en vain
3°) On examine le couple 3 - 4 : on trouve 3 d⟶4
4°) On examine le couple 4 - 5 : on trouve 5 d⟶4

Après cette première exploration, conduite de gauche à droite à partir du premier mot à gauche, la machine cherche les nouvelles antériorités qui ont pu apparaitre à la suite des dépendances et dérivations mises en évidence : elle découvre ainsi :

2 p 4 (puisque 3 d⟶4)

D'autre part, les antériorités 1 p 2, 3 p 4 et 4 p 5 ont disparu, en vertu de la règle I.3

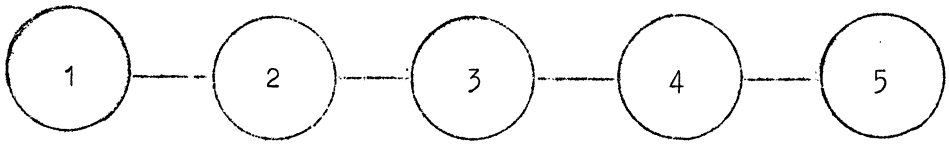
- Deuxième exploration :

La machine explore la nouvelle antériorité découverte : 2 p 4 et trouve la dépendance 4 d⟶2

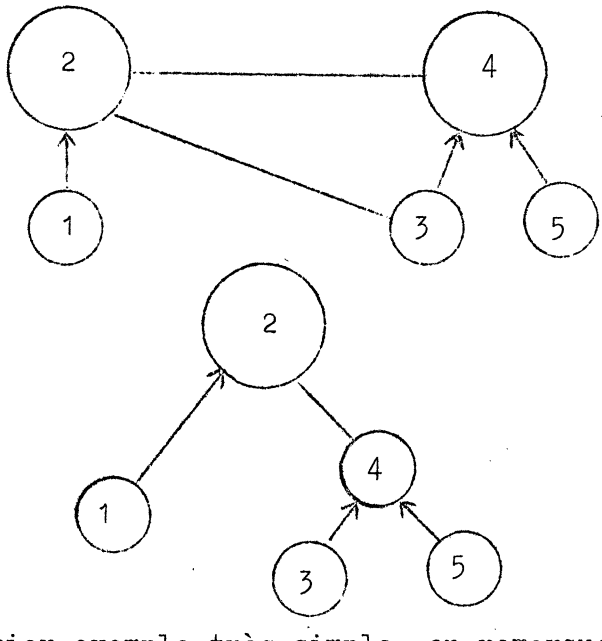
Cette nouvelle dépendance fait disparaître l'antériorité 2 p 4, puisque 3 dérive de 2 (3 d⟶4 d⟶2) (règle 1.3) 2 p 3.

Les différentes opérations deviennent plus claires si on adopte une représentation par diagrammes, les traits sans flèches représentant les antériorités, les traits avec flèches représentant les dépendances. Les mots indépendants étant représentés par des grands cercles, les mots dépendants par des petits cercles.

On obtient ainsi la suite I qui fait apparaitre la structure :



SUITE I



Dans ce premier exemple très simple, on remarquera que toutes les dépendances trouvées étaient d'ordre grammatical :

1 - d⟶ 2 : le sujet je dépend du verbe à la première personne : vois

3 - d⟶4 : l'article féminin singulier une dépend du substantif féminin singulier maison

5 - d⟶ 4 : l'adjectif singulier rouge dépend du substantif singulier maison

4 - d⟶ 2 : le substantif maison dépend du verbe transitif vois

Nous allons maintenant prendre un second exemple, un peu plus compliqué, celui que HAYS et HARPER ont appliqué au russe, mais que nous étendrons au français, à l'anglais et à l'allemand.

Il s'agit de la phrase suivante :

L'accroissementobservé de la résistance superficielle avec la
1 2 3 4 5 6 7 8 9

fréquence peut être expliqué par l'augmentation de la profondeur effective
10 11 12 13 14 15 16 17 18 19 20

de pénétration.
21 22

The observed increase in surface resistance with frequency
1 2 3 4 5 6 7 8

can be explained by an increase in the effective depth of penetration.
9 10 11 12 13 14 15 16 17 18 19 20

Наблюденный рост поверхностного сопротивления
I 2 3 4

с частотой может быть объяснен возрастанием
5 6 7 8 9 IO

эффективной глубины проникновения.
II I2 I3

Die in Abhängigkeit der Frequenz beobachtete Zùnahme des
1 2 3 4 5 6 7 8

oberflächichen Widerstands kann dùrch eine Zùnahme der effektiven
9 10 11 12 13 14 15 16

Eindringùngstiefe erklärt werden.
17 18 19

Nous allons donner la suite détaillée des opérations de la machine en ce qui concerne la phrase allemande pour bien montrer la "remise" en ordre effectuée.

Nous donnerons ensuite les diagrammes trouvés pour les quatre phrases.

Traitement de la phrase allemande :

- Première exploration :

On trouve :

3 d——→2 4 d——→5 6 d——→7 9 d——→10

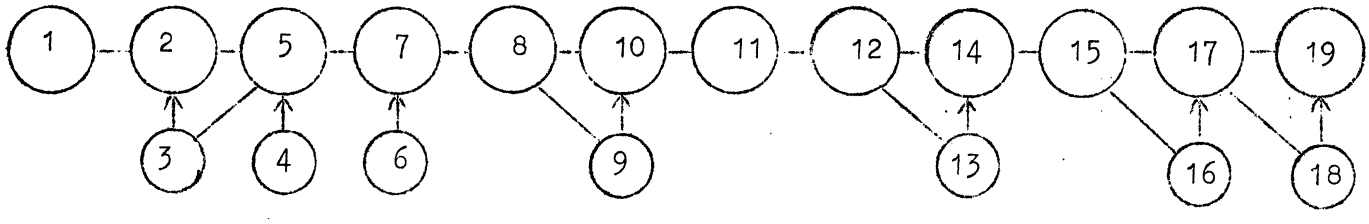
13 d——14 16 d——17 18 d——19

Disparaissent : les 7 antériorités correspondant aux dépendances trouvées, et :

3 p 4 (2 dépendants)

Apparaissent :

2 p 5 3 p 5 5 p 7 8 p 10 12 p 14 15 p 17 17 p 19



- Deuxième exploration :

On trouve :

5 d⟶3 8 d⟶10 15 d⟶17 14 d⟶12

Disparaissent : les quatre antériorités correspondant aux quatre dépendances trouvées, et :

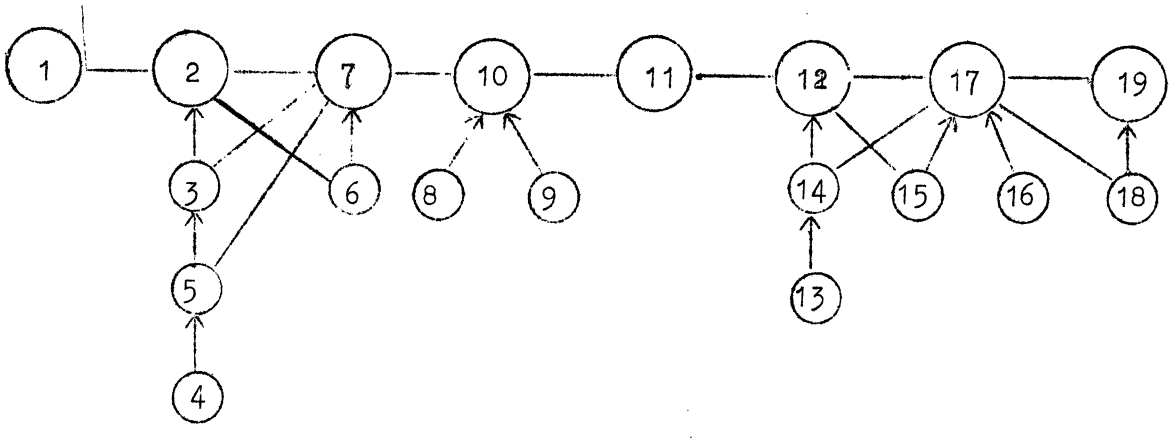
2 p 5 (dérivation), 8 p 9 (2 dépendants), 12 p 13 (dérivation),

14 p 15 (2 dépendants), 15 p 16 (2dépendants)

Apparaissent :

2 p 6 2 p 7 3 p 7 7 p 10 14 p 17 12 p 17 12 p 15

d'où le sohéma :



- Troisième exploration :

On trouve :

2 d——→7 (remarque : on aurait pu également envisager la dépendance : 2 d——→6) 10 d——→7 17 d——→14

Disparaissent : les 3 antériorités correspondant aux 3 dépendances trouvées, et :

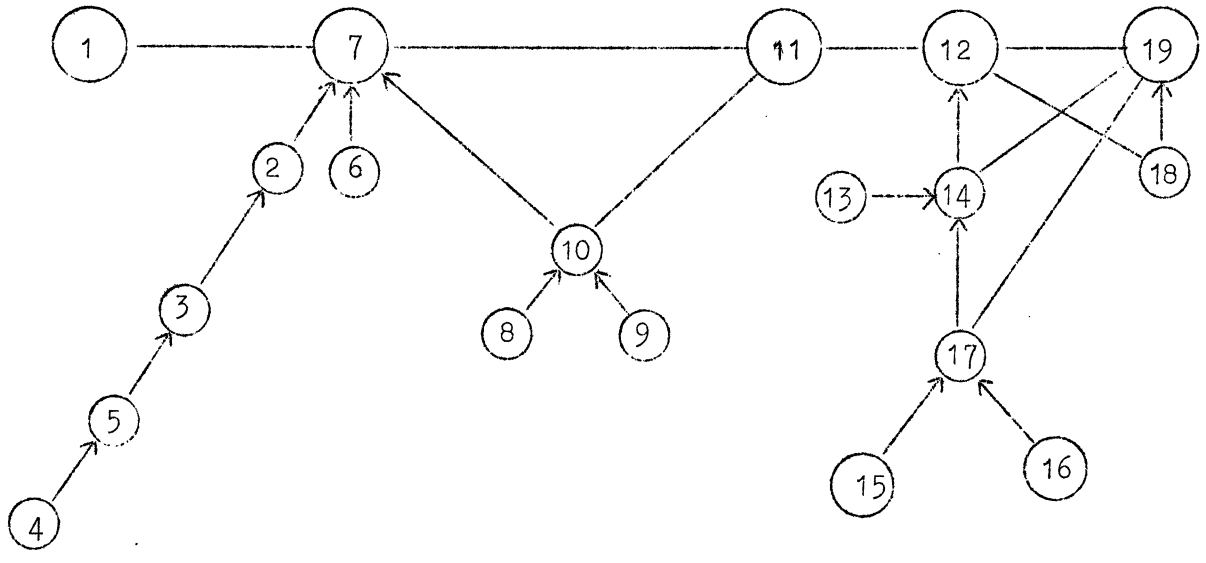
2 p 6 (2 dépendants), 3 p 7 (dérivation), 5 p 7 (dérivation),

17 p 18 (2 dépendants), 7 p 8 (dérivation), 12 p 15 (dérivation)

Apparaissent :

1 p 7 7 p 11 12 p 18 12 p 19 14 p 19.

d'où le schéma :



- Quatrième exploration :

On trouve :

1 d ⟶ 7 7 d ⟶ 11 12 d ⟶ 18

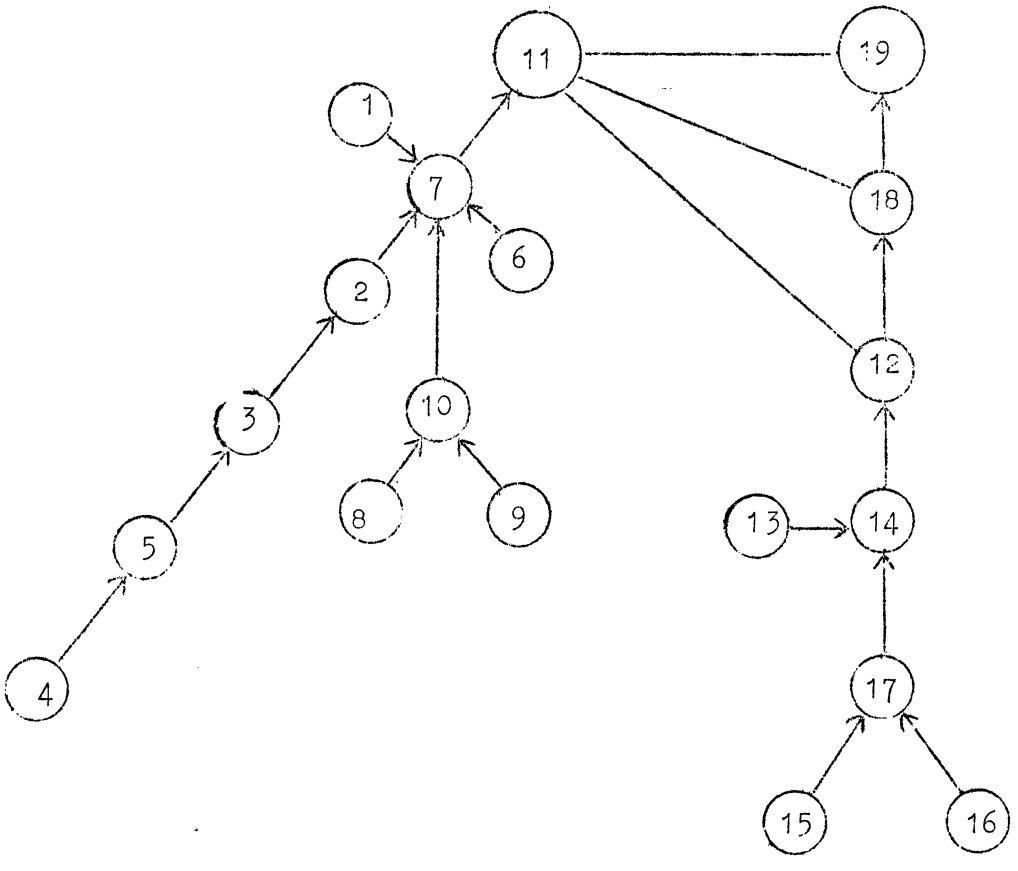
Disparaissent : les 3 antériorités correspondant aux trois dépendances trouvées, et :

10 p 11 (dérivation), 17 p 19 (dérivation), 12 p 19 (dérivation),

14 p 19 (dérivation)

Apparaissent :

11 p 18 11 p 19

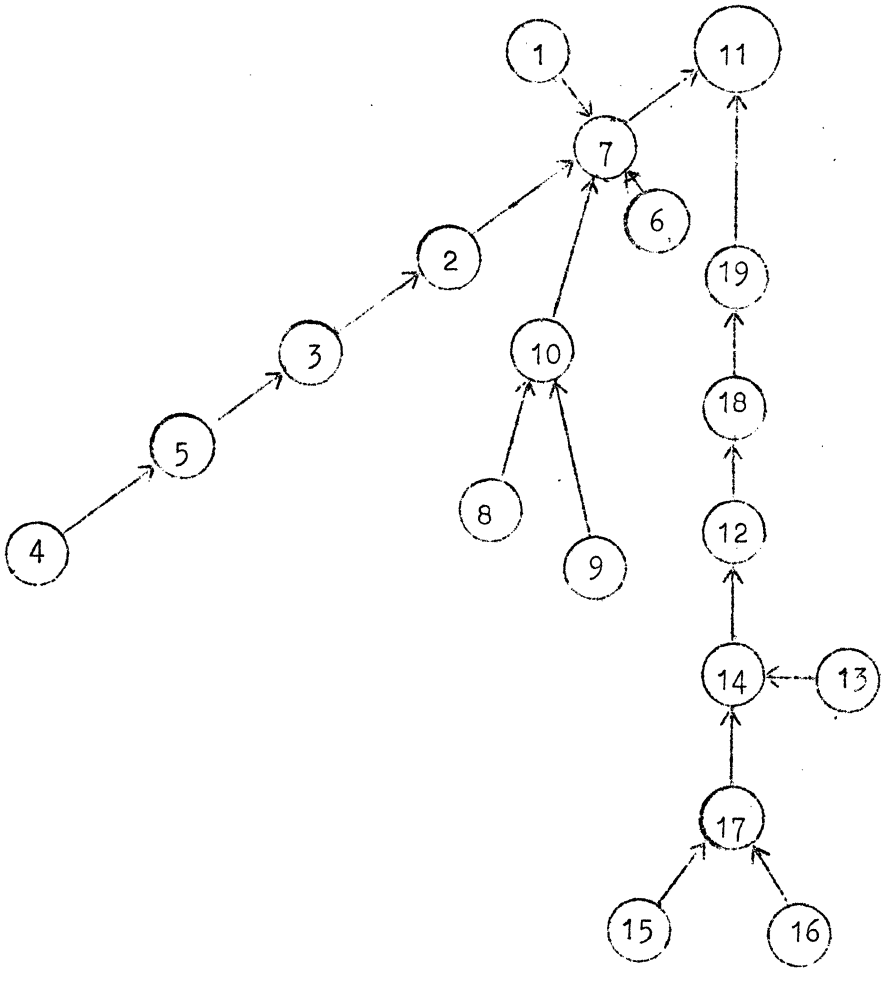


- Cinquième exploration :

On trouve :

19 d——→11, ce qui fait disparaitre 11 p 18 (dérivation) et 11 p 12 (dérivation)

D'où la structure finale :



Nous n'avons pas expliqué les dépendances que nous avons trouvées.

Signalons simplement que, dans le cas de cette phrase, elles sont surtout d'ordre morphologique ou syntactique :

ainsi, 4 d⟵5, c'est-à-dire "der" dépend de "Frequenz", est prévu en vertu de règles telles que :

- l'article "der" précède "Frequenz"
- la forme der est compatible avec le substantif, à condition toutefois que ce substantif soit au cas génitif ou au cas datif.

Autrement dit, la dépendance trouvée permet de mettre pratiquement l'article "hors-jeu", et reporte l'information intéressante sur le substantif : féminin singulier au datif ou au génitif, ce qui facilitera l'analyse ultérieure (découverte de la dépendance 5 d⟶3).

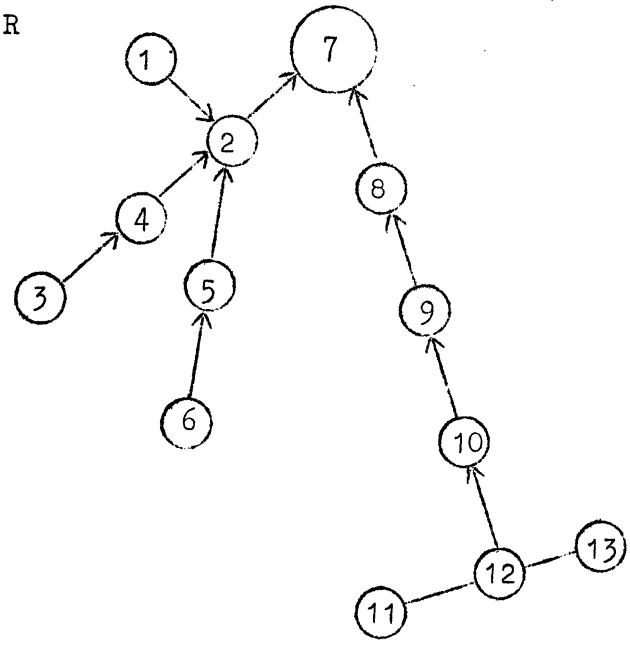
Avant de donner les diagrammes correspondant aux phrases française, anglaise et russe, remarquons la méthode d'exploitation de la machine : elle explore la phrase de gauche à droite et détermine les liaisons réelles entre les mots, en utilisant les indications du contexte proche (ordre des mots, fonction des mots, morphologie des mots) et les indications du contexte lointain (dépendances sémantiques);

elle revient ensuite en arrière, établit une nouvelle structure, plus proche de la logique, et recommence une nouvelle exploration.

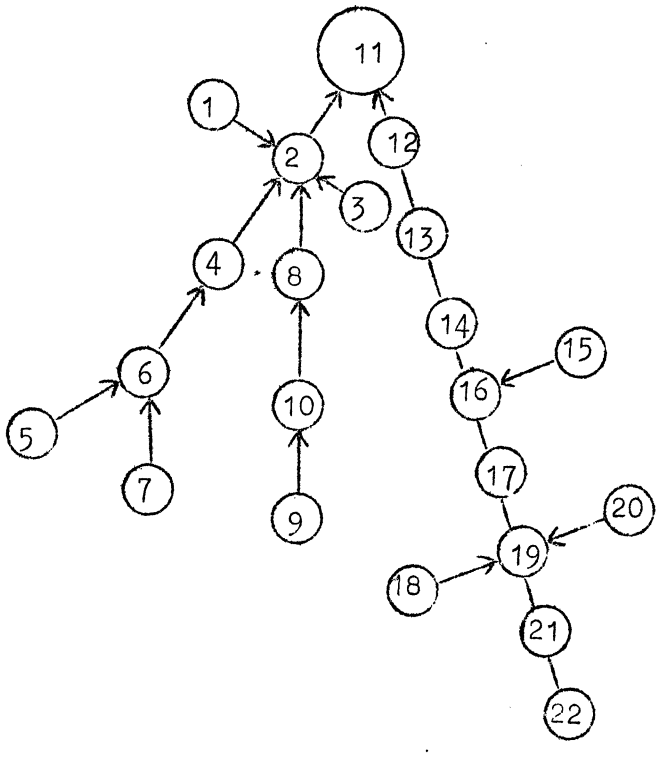
L'important ici est de voir que la machine joue simultanément sur l'ordre, la morphologie et la sémantique, et que l'analyse de la phrase ne peut se faire normalement que si toutes les dépendances "sémantiques" se font correctement (exemple de la dépendance : 2 (ou mieux 2-3) d ⟶ 7 : in Abhängigkeit (en fonction de) dépend de Zùnahme (augmentation).

En applicant la même méthode aux phrases en français, en russe et en anglais, on obtient les 4 schémas suivants :

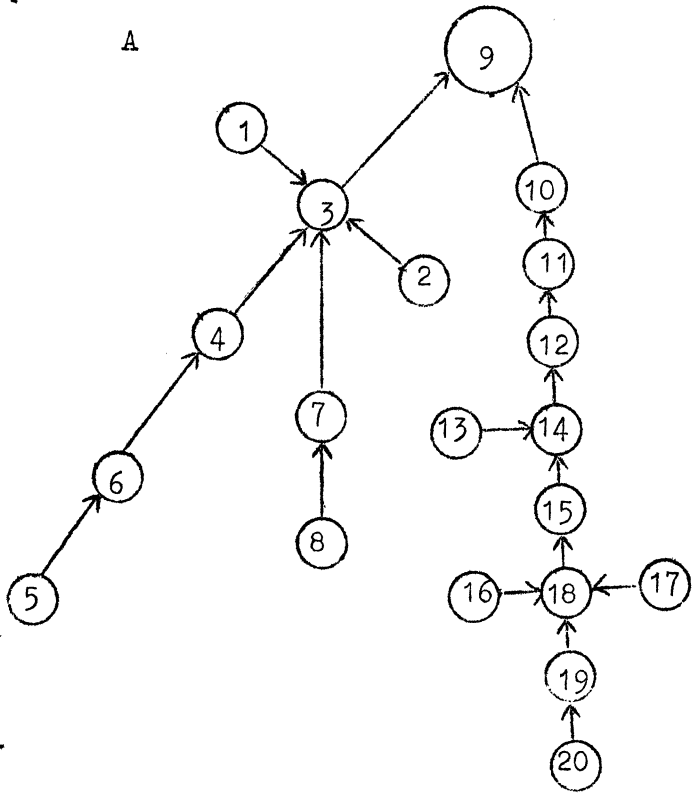
R


1
2
3
4
5
6
7
8
9
10
11
12
13

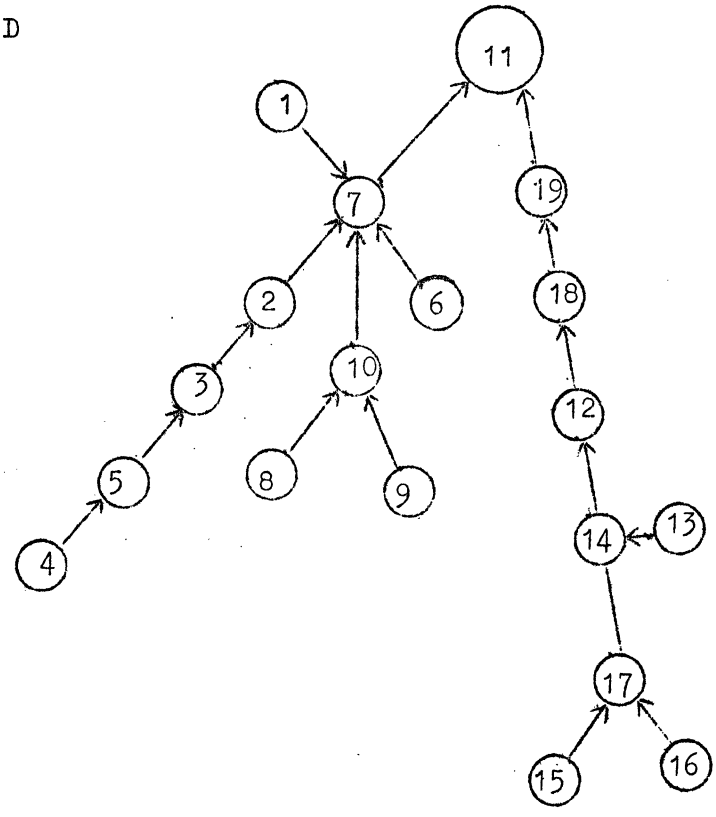
F


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

A


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

D


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19

Si on considère les 4 diagrammes obtenus, on constate :

1 - que la structure n'est plus linéaire, mais qu'elle est cependant semblable pour les phrases dans les 4 langues indo-européennes utilisées.

2 - que l'ordre des mots importants dans chaque branche est le même pour les différentes langues.

3 - que les noeuds sont identiques dans les différentes langues.

Après l'analyse de la phrase à traduire commencent les opérations de passage à la phrase traduite. Nous ne les décrirons pas, car les liens entre la traduction et la documentation automatique se situent justement à ce niveau de la structure des phrases.

Pour le montrer, nous allons rappeler le principe de l'expérience de documentation automatique effectuée à Francfort l'an dernier (9-12 juin 1959).

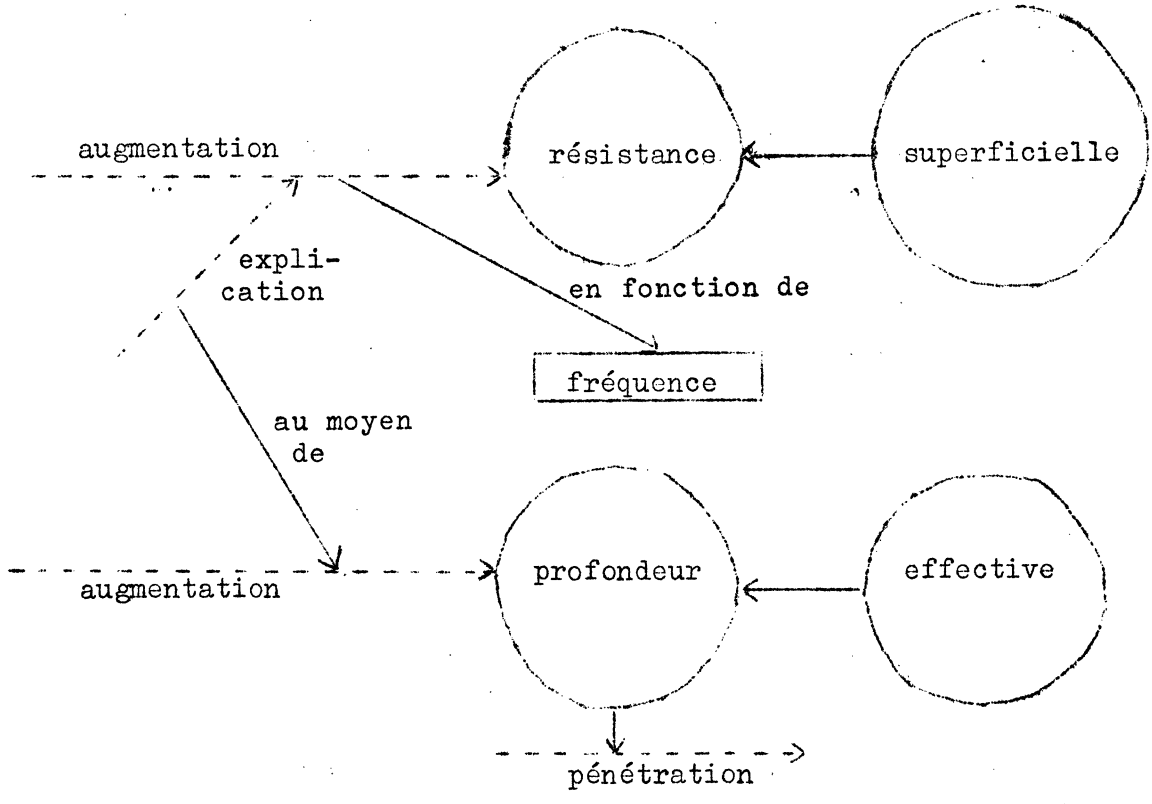
Le codage utilisé (5) employait les mots, qui étaient classés en 5 catégories distinctes :

- objets - ou entités
- actions
- propriétés (modifiant ou précisant les objets)
- conditions (modifiant ou précisant les actions)
- relations

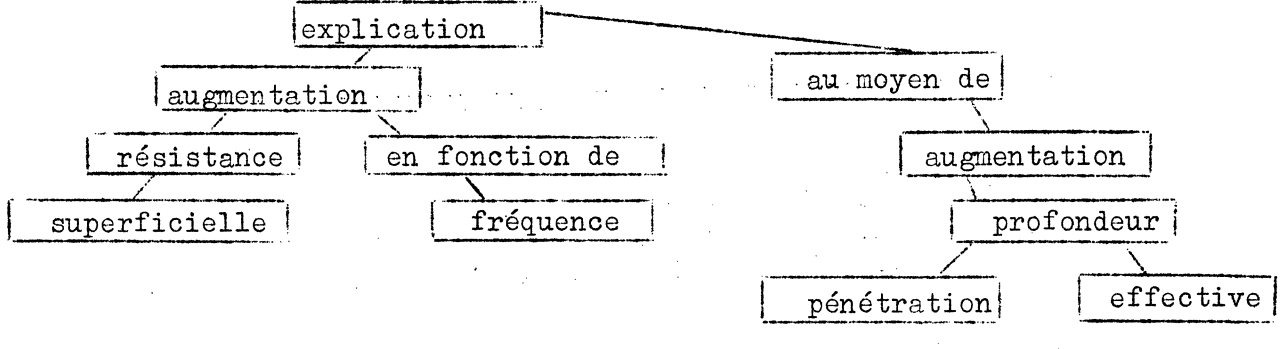
En fait, les relations, tout en pouvant être des mots, avaient surtout pour fonction principale de relier les autres mots (objets, actions, propriétés, conditions) entre eux.

La notion était définie par un "diagramme", représentation d'une "phrase-clé", où les objets se trouvaient dans des rectangles, les propriétés et les conditions dans des cercles, les actions à côté des flèches en pointillé, et les relations à côté des flèches en trait plein.

Dans ces conditions, la phrase précédente aurait été représentée par le schéma suivant :



Si nous ne tenons pas compte de la fonction des mots, ce diagramme est équivalent au suivant, qui présente la même structure que la phrase analysée par la machine :



Remarquons que le diagramme de LEROY et BRAFFORT, fondé sur des considérations sémantiques, n'était pas établi automatiquement, mais fait par des êtres humains, et que l'expérience d'automatisation portait uniquement sur le traitement de ces diagrammes par une machine électronique.

Ceci dit, ils représentent une analyse plus poussée, supposent une étape supplémentaire de l'analyse, car les fonctions des mots sont explicitées, ce qui permet de trouver une structure "ouverte" qui doit être complétée :

ainsi, pénétration (action) appelle deux objets (sujet et complément),
résistance (propriété) appelle l'objet qu'elle précise.

Ceci aurait tout aussi bien été mis en évidence par l'analyse automatique de la structure : en effet, on a admis à priori dans la méthode de la Rand qu'un mot ne pouvait dépendre que d'un autre mot : or, ce n'est pas le cas, aussi bien du point de vue grammatical (en particulier pour les pronoms relatifs qui, que, etc.) que du point de vue sémantique : les dépendances sémantiques constituent certainement le vrai problème de la traduction et de la documentation automatiques et les méthodes décrites plus haut (Cambridge et Rand) ne constituent que des ébauches, des indications de solutions.**

Dans ce qui précède nous avons voulu simplement mettre en évidence la nature identique des problèmes posés par la traduction et la documentation automatique.

Nous n'avons pas voulu, par contre, démontrer que la traduction ou la documentation automatique étaient réalisables actuellement, et encore bien moins qu'elles étaient réalisées actuellement.

Des études très longues seront encore nécessaires avant que l'on puisse discerner la direction conduisant à la résolution du problème.

** De nombreuses autres méthodes ont été utilisées pour traiter les problèmes de la traduction automatique; d'autre part, nous avons volontairement simplifié les méthodes du "Cambridge Language Research Unit" et de la "Rand Corporation". Les étudiants intéressés devront consulter les documents cités dans la bibliographie sommaire pour avoir une vue plus complète sur la traduction automatique.

Cependant, par suite de l'opinion fausse, mais très répandue, que l'obstacle à la transmission de l'information est uniquement dû à la différence des langues utilisées, ce sont les problèmes de la traduction automatique qui ont été étudiés en premier à une très grande échelle aux Etats-Unis et en U.R.S.S. C'est pourquoi, si cette tendance se poursuit, la première traduction complètement automatique précèdera peut-être la première recherche documentaire complètement automatique; mais, si ce cas se produit, on peut affirmer que la machine documentaire suivra de très près la machine à traduire.

(1) MASTERMAN M., NEEDHAM R.M., SPÄRCK HONES K. : The Analogy between mechanical translation an library retrieval.

Proceedings of the International Conference on Scientific Information.

National Academy of Sciences-National Research Council, Washington.

(2) ROGET M.P. : Thesaurus of en lish words and phrases

Longmans, Green and C° Ltd., London.

(3) CECCATO S. : Principles and classification of an operational grammar for mechanichal translation.

International Conference for Standards on a Common Language for Machine.

Searching and Translation, Cleveland, September 6-12, 1959.

(4) HARPER K.E., HAYS D.G. : Uses of machines in the construction of a grammar and computer program for structural analysis.

International Conference on Information Processing, Paris, UNESCO, 15-20 June 1959.

Preprint UNESCO/NS/ICIP/F.4.4.

(5) LEROY A., BRAFFORT P. : Notice relative à l'élaboration d'un codage par phrases-clés pour la programmation d'un système de sélection automatique des documents.

Note C.E.A. n° 278.

BIBLIOGRAPHIE SOMMAIRE S R LA TRADUCTION AUTOMATIQUE :

DELAVENAY E. : La machine à traduire,

Presses Universitaires de France, Paris, 1959.

DELAVENAY E. : Bibliographie sur la traduction automatique.

(à paraitre)

BOOTH D., BRANDWOOD L., CLEAVE J.P. : Mechanical resolution of linguistic problems.

Butterworths Scientific Publications, London, 1958.

LOCKE W.N., BOOTH A.D. : Machine translation of languages.

John Wiley and Sons, New York, 1955.

Les quatre livres cités plus haut constituent une bonne introduction. Pour se tenir au courant des progrès des travaux, il est nécessaire de consulter les publications périodiques suivantes :

Current research and development in scientific documentation
(Publication bi-annuelle de la National Science Foundation)

M.T. Mechanical Translation (publié par le M.I.T.)

Вопросы Языкознания (Издательство Академии Наук СССР, Москва)

Проблемы Кибернетики (Государственное Издательство физико-математической Литературы, Москва.)

LA DOCUMENTATION COMPLETEMENT AUTOMATIQUE

A. LEROY

Je rappellerai d'abord ce que j'entends par documentation complètement automatique. Il s'agit d'obtenir rapidement, grâce à une machine, une réponse suffisamment complète (et non pas forcément les documents permettant de répondre) à n'importe quelle question scientifique, dans la langue du demandeur, étant entendu que nous ne lui fournissons que des pages de publications scientifiques (et bien sûr les éléments de la question).

Je rappellerai aussi que je n'ai pas la prétention de dire que la chose est faite; il faudra encore des années de travail pour arriver au but. Enfin, je ne me préoccuperai pas des exigences pécuniaires, c'est-à-dire que j'envisagerai, tout en respectant la condition d'économie maximum, des moyens extrêmement puissants que les exigences pécuniaires obligeront peut-être à ramener à de plus faibles proportions.

Nous travaillons donc pour l'avenir. Il y a lieu de ne pas l'oublier sous peine d'être dérouté par les idées que je vais émettre, bien qu'elles soient extrêmement simplifiées. Je terminerais toutefois en indiquant ce que pourrait être une première étape, constituant déjà une solution partielle du problème posé.

Lecture

En premier lieu, la machine doit lire les textes qui lui sont présentés; Je n'exposerai pas ici les détails techniques des futures machines à lire et je rappellerai simplement qu'au cours de la Conférence organisée en juin 1959 par l'Unesco sur le traitement des informations, plusieurs exposés ont montré que l'on était sur la voie d'une solution acceptable.

Analyse

Une fois les mots "lus", c'est-à-dire introduits sous forme de suite de signaux dans une partie de la machine, la phase analyse doit commencer.

Que doit-être cette analyse ? Elle doit être la meilleure possible d'après le but général que nous nous sommes fixés. Cela signifie que tout le contenu scientifique des publications doit être pris en compte.

Remarquons que cette idée a déjà été émise notamment par le Russe Gutenmakker en 1956 [1] et par l'Américain Yngve en septembre 1959 [2].

Gutenmakker s'exprimait à peu près ainsi : Dans peu de temps, la vitesse d'opération des machines sera telle que l'on pourra en un temps relativement court rechercher une information scientifique dans la littérature scientifique de plusieurs années.

Les travaux pratiques sur l'établissement des diagrammes ont montré qu'il est possible de réaliser automatiquement le passage d'un texte normal au diagramme équivalent grâce d'abord à une analyse syntaxique automatique dont M. Lecerf a exposé les principes, puis à des consignes et un dictionnaire.

Polysémie-synonymie

Néanmoins cela ne sera pas toujours aussi simple. Nous savons en effet qu'il faut tenir compte des problèmes de polysémie et de synonymie. Nous avons eu l'occasion de dire qu'ils étaient résolus à partir des considérations de contexte. En effet, si à la place de chaque mot on considère son contexte, on obtient autant de "notions" qu'il y a de contextes différents. Par contre, des contextes identiques ne désignent qu'une seule notion même si celle-ci possède ordinairement plusieurs formes écrites. Il faut donc un dictionnaire qui donne tous les contextes possibles; on voit le vague de cette assertion; qu'est-ce exactement que le contexte en question ? Comment peut-on être sûr d'avoir tous les contextes possibles ?

Diagramme général

Le diagramme général dont il a déjà été question permet de répondre. En effet, puisqu'il contient l'ensemble des contenus scientifiques, il contient l'ensemble des contextes interessants et, en définitive, l'ensemble des définitions. C'est donc le meilleur dictionnaire qui soit, c'est un dictionnaire continuellement mis à jour.

Que sont en effet les définitions ?

Il y en a plusieurs types qui ont été mis en évidence par M.M. Prot [3]. Nous lui emprunterons d'ailleurs les exemples qui suivent :

La définition spécifique ex. L'hexagone est un polygone qui a six côtés.

La définition générique ex. Le triangle, le losange, le trapèze, l'hexagone sont des polygones.

La définition descriptive ou figurative ex. "L'hypothénuse est, dans un triangle-rectangle, le côté opposé à l'angle droit.

La définition génitive associe à ce que l'on veut définir un passé, une origine, une histoire, une fabrication en un mot : une genèse.

La définition destinative S'il s'agit d'un objet, la définition destinative indique par exemple les transformations auxquelles il se prête, les emplois qu'on peut en faire, la façon dont il change avec le temps.

La définition équivalente ou synonymique ex. harpiste = personne qui joue de la harpe.

etc. (M. Prot en distingue 9 sortes)

Considérons maintenant ce qui se trouve en face du mot benzine dans un dictionnaire ordinaire :

liquide volatif, incolore, aromatique, extrait des goudrons de houille, employé pour détacher les étoffes et comme carburant.

On reconnaît dans cette définition complexe une succession de définitions élémentaires : spécifique, génitive, destinative

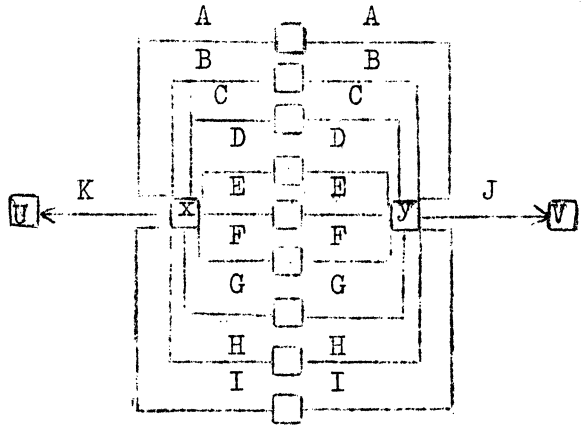
↓	↓	↓
liquide volatil	extrait de	utilisé pour

Et finalement, si l'on envisage tous les types de définitions possibles, on arrive à tout ce qui peut être dit sur un mot donc au diagramme général qui est ainsi le meilleur des dictionnaires possibles.

Supposons donc le diagramme général dans un état suffisant d'avancement (on peut envisager qu'une première étape sera manuelle, puis qu'une deuxième sera semi-automatique introduisant une aide par le jeu de bons dictionnaires existants que l'on incorporera dans la machine) pour qu'on puisse considérer que la plupart des mots qui y figurent y possèdent une partie suffisante des définitions connues au moment de l'expérience.

La machine continue de lire les mots des textes qui lui sont présentés et se reporte au diagramme général pour savoir quel est le sens véritable de chaque mot lu, en comparant le contexte de ces mots avec celui du mot correspondant dans le diagramme général. La machine est donc capable de distinguer entre les formes qui, étant identiques, expriment des choses différentes. Comment peut-elle alors résoudre le second problème, celui de la synonymie ? Ici, des formes différentes expriment la même chose. Sur le diagramme général, cela peut par exemple [illegible] entendu):

se représenter de la manière suivante (en simplifiant, bien entendu) :



A B C etc... sont par exemple des "actions".
Les carrés représentent des "objets".

Les expressions x et y apparaissent en première approximation synonymes à 90%.

N'attribuons pas à ce chiffre plus de signification qu'il n'en a. Reconnaissons du moins qu'il est une indication sur la parenté des deux notions
La difficulté est évidemment de trouver le "seuil" à partir duquel ils peuvent être déclarés synonymes. De plus, il peut se faire que l'une des deux expressions à comparer, bien que possédant le même sens que l'autre, ne soit pas aussi souvent employée; son contexte général est alors moins fourni.

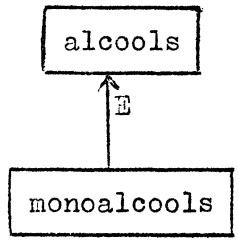
Mais remarquons encore que même si la machine "juge" que deux expressions ne sont pas suffisamment synonymes pour pouvoir être confondues, rien ne risque d'être perdu au moment de la sélection puisque, suivant le principe déjà exposé, la machine fournit d'abord la réponse correspondant directement à la question (au point de vue de la forme) mais aussi des indications sur la possibilité de rechercher des notions voisines et en tout premier lieu les notions présentant un certain degré de synonymité avec celles qui figurent dans la question.

Le diagramme général pourra donc se compléter automatiquement et le diagramme de chaque phrase se construire de la même façon à partir du diagramme général dans les normes voulues, à condition que l'on ait donné des consignes générales pour l'adoption de la forme qui sert à exprimer la signification commune aux synonymes.

Superposition

Les diagrammes correspondant à chaque phrase ne sont pas tous indépendants, il faut réaliser leur superposition. Pour pouvoir superposer deux diagrammes élémentaires, il faut que l'on soit sûr que les expressions que l'on confond soient exactement les mêmes; il faut donc là encore considérer le contexte.

Il peut se faire aussi que des liaisons hiérarchiques puissent s'établir; c'est un peu ce que l'on demandait de faire avec les phrases de chimie dans lesquelles figuraient les mots alcools et monoalcools. Il y avait évidemment une relation d'appartenance qui apparaissait



Or cette relation n'était pas construite automatiquement à partir des tableaux correspondant aux phrases proposées, pour la simple raison qu'il n'y a pas lieu de le faire. En effet, il est bien improbable que dans l'ensemble des publications déjà analysées on ait pas signifié cette liaison; par exemple : les alcools comprennent les monoalcools etc... Enfin n'oublions pas que nous pouvons à tout moment renseigner nous-mêmes la machine en lui communiquant tout ce dont nous sommes sûrs.

Halo sémantique

Le diagramme général permet donc de faire beaucoup de choses. Peut-il aller loin dans cette direction et, en particulier, peut-il permettre de franchir l'obstacle du halo sémantique dont M. Gardin a parlé ?

Et pourquoi pas ? Pourquoi la machine ne pourrait-elle pas "apprendre" que le fait d'emmurer un architecte constitue dans certaines conditions, et seulement dans ces conditions, une condamnation à mort ? Pourquoi ne pourrait-elle pas apprendre qu'il intervient alors la notion de responsabilité professionnelle? Il suffit pour cela que des textes aient existé,qui aient indiqué ces relations de dépendance entre emmurage et condamnation à mort et entre accident provoquant la mort et l'emmurage et responsabilité professionnelle.

Enfin, remarquons encore une fois que la question ne se pose pas pour nous de la même façon que pour M. Gardin. Nous nous intéressons à des collections de millions de documents et nous ne demandons absolument pas que tous les documents traitant de l'écroulement d'un palais etc... (pour nous il s'agirait plutôt de l'explosion d'un réacteur) apparaissent toutes les fois que l'on s'intéresse à la notion de responsabilité professionnelle. Et ceci, même dans le cas où nous désirons voir sortir des

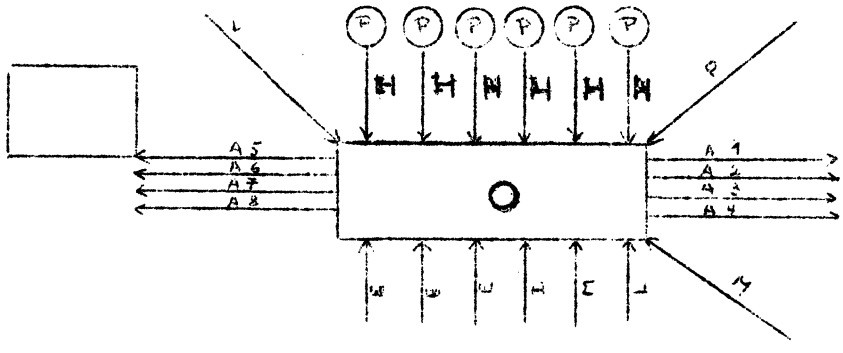
documents non pas une réponse directe. En tout premier lieu ne doivent sortir que les documents traitant de celle-ci sur un plan général; il est d'ailleurs probable que de très nombreux documents sortiront déjà. Et c'est seulement si cela n'est pas satisfaisant que nous pourrons consulter la liste des indications fournies par la machine, ex : des documents pourront vous être fournis qui correspondent à des notions voisines, notions partiellement synonymes d'abord telles que: ..... puis hiérarchiques..... puis latérales etc...

Disposition en machine

Bien entendu l'obstacle que l'on peut voir à la réalisation de telles considérations c'est qu'elles font intervenir un trop grand nombre d'éléments. Nous avons jusqu'ici supposé qu'il n'y avait aucune limite matérielle, ce qui est une considération évidemment erronée. Mais n'oublions pas que nous oeuvrons pour l'avenir et que si nous laissons de côté les questions pécuniaires rien ne s'oppose à ce que nous fassions intervenir des mémoires de très grande capacité et des éléments logiques opérant à une très grande vitesse.

La seule question que nous nous posons est donc de savoir comment un tel diagramme peut être introduit en machine.

Considérons donc un morceau de diagramme.



Exemple de représentation machine (1) :

Notions classées par ordre alphabétique	Numéros des documents + Numéros des phrases	Actions ou Relations liées à 0	Notion située à l'autre extrêmité de l'action ou de la relation
0	17888-4	E1	objet 1
			objet 2
			objet 3
	173739 - 1	E2	objet 4
			objet 5
		Action 1	objet 6
		etc...	

(1)Le nombre de colonnes pourrait être réduit facilement à deux seulement,une colonne contenant toutes les informations classées ici en 4 colonnes, et l'autre servant seulement à indiquer dans quelle colonne du système représenté ici se trouverait une information donnée.

Modification du diagramme général

Le diagramme général est continuellement modifié, soit parce qu'il est complété, soit parce que des informations nouvelles annulent des renseignements enregistrés antérieurement; avant de rendre cette annulation effective, il faudra en attendre la confirmation - De toute façon, les renseignements périmés devront être conservés à part, ne serait-ce que pour pouvoir répondre aux questions concernant les historiques.

Cas ou la phrase introduite est incorrecte :

C'est un peu le cas pour la première phrase des travaux pratiques : Le bleu, le rouge, le vert sont trois couleurs fondamentales dont le mélange fournit pratiquement l'ensemble des couleurs possibles.

Il est certain qu'il manque une précision à propos du mélange. Mais si cette phrase apparaît dans un texte, elle doit être portée dans le diagramme général et c'est celui-ci qui permet de voir que le mélange en question se fait de telle et telle façons, parce que cela a déjà été écrit auparavant. Au cas où une telle phrase concernerait une découverte, le diagramme général s'en enrichirait jusqu'à ce que le contraire soit prouvé ou qu'elle soit modifiée.

Degré de nouveauté d'un texte

Lorsque le diagramme général sera suffisamment développé, certaines parties des textes qui seront présentés à la machine s'avèreront figurer déjà dans le diagramme, du moins quant au contenu réellement exprimé. A la limite, certains textes apparaîtront comme n'apportant rien de nouveau; ils auront déjà été "écrits" dans diverses publications; ils n'apporteront alors qu'une confirmation.

Synthèses

On voit apparaître la notion de synthèse qu'il est ainsi possible de réaliser à l'aide d'une telle machine. Pour réaliser une synthèse dans un domaine donné il suffit qu'elle fournisse la partie du diagramme général qui correspond au domaine en question. Nous pouvons même dire qu'elle est capable d'opérer des déductions.

En effet, prenons l'exemple de déduction par déplacement hiérarchique simple; même si cela n'est dit dans aucun texte, il est facile de faire apprendre à la machine qu'un objet appartenant à une certaine classe possède toutes les propriétés de la classe. Ex : un triangle isocèle possède les propriétés générales des triangles, par exemple la propriété d'avoir trois côtés.

En fait la machine effectuait déjà une opération analogue à une déduction lorsqu'elle était capable de résoudre les problèmes de polysémie et de synonymie, de réaliser la superposition des diagrammes et d'établir des liens hiérarchiques.

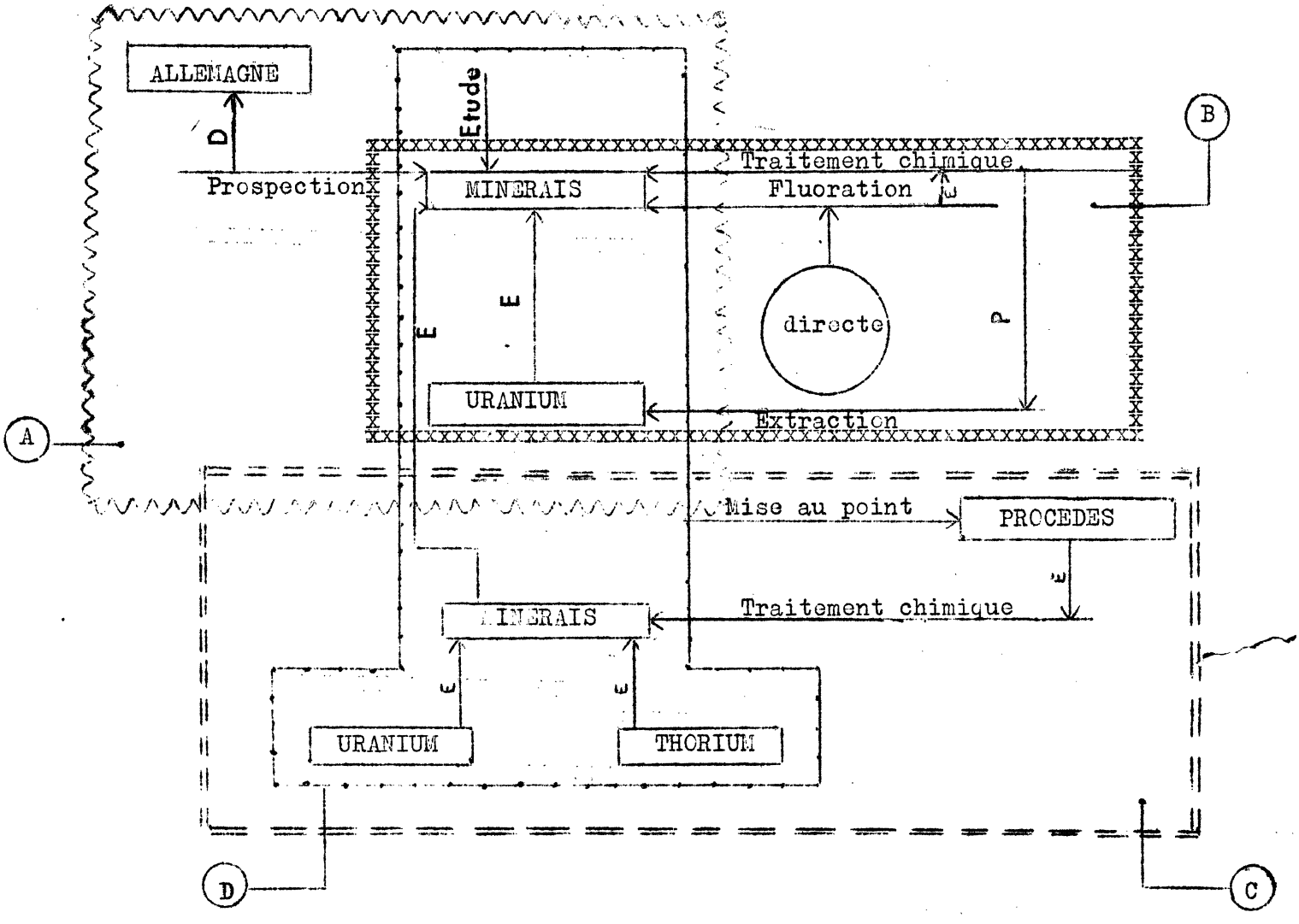
Cas des textes avec figures

Il peut se faire que certains dessins ou diagrammes qui figurent dans les textes doivent être transcrits sous peine de perdre une partie de l'information. Sans doute, des méthodes utilisées par M. Gardin pour les ornements abstraits par exemple pourront-elles alors être d'un grand secours.

Sélection

Elle se fait par comparaison, c'est-à-dire que les éléments de la question posée sont comparés avec le contenu de la mémoire. Et comme nous nous servons ici d'un appareil électronique qui est supposé capable de suivre rapidement n'importe quel "circuit" à l'intérieur du diagramme général, il est possible de prévoir des schémas aussi complexes qu'on le veut.

Reprenons notre exemple du diagramme général



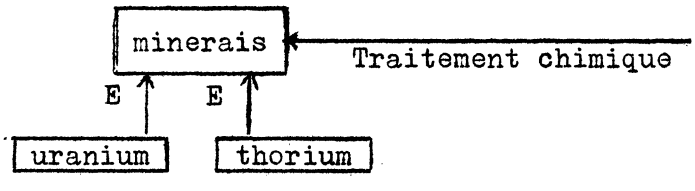
Nous savons maintenant que la représentation en machine peut par exemple être la suivante :

Minerais			
	B.	E2	Uranium
	B.	Traitement chimique 2	
	B.	Fluoration 2	
	D.	Etude 2	
	D.	E2	Uranium
	D.	E2	Minerais
	A.	E2	Uranium
	A.	Prospection 2	
	A.	Etude 2	
Minerais	C.	Traitement chimique 2	
	C.	E2	Uranium
	C.	E2	Thorium
Traitement chimique etc...	B.	E2	Fluoration

Comment se fait la sélection ?

Question demandant une réponse en documents : Quels sont les documents qui se rapportent aux traitements chimiques des minerais d'uranium et de thorium ?

Diagramme de la question :



représentation machine :

Minerais		
	Traitement chimique 2 E 2 E 2	 Uranium Thorium
Uranium	E 1	Minerais
Thorium	E 1	Minerais

Comparaison : on voit que dans le diagramme général on a :

Minerais	C. C. C.	Traitement chimique 2 E 2 E 2	 Uranium Thorium

Les deux autres conditions figureraient également si la représentation machine était complète et une simple comparaison montre que le document C répond à la question.

De plus, des indications supplémentaires sont données grâce à la relation E 1 qui permet de renvoyer aux minerais d'uranium, montrant que si les documents fournis ne suffisent pas, il est possible d'obtenir également des documents se rapportant au traitement des minerais d'uranium en général.

Minerais	D	E 1	Minerais

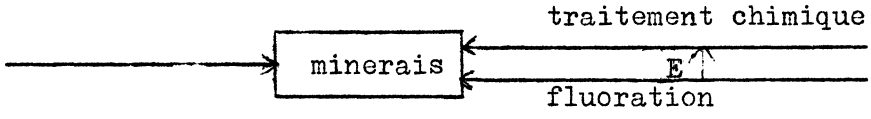
Minerais	D	E 2	Uranium

Il y a aussi des indications montrant qu'il est également possible de répondre à la question posée à l'aide du groupement de plusieurs documents.

Considérons par exemple la question suivante :

Quels sont les documents se rapportant à la prospection de minerais qui peuvent donner lieu à un traitement par fluoration ?

Question :



Minerais	Prospection 2 Traitement chimique 2 Fluoration 2	
Traitement chimique	E2	Fluoration

Comparaison : On voit que l'on trouve :

Minerais	B. B. A.	Traitement chimique 2 (1) Fluoration 2 Prospection 2	
Traitement chimique	B.	E2	Fluoration

(1) C. ne figure pas, car il ne s'agit pas des mêmes minerais.

On voit que (A + B) répond à la question

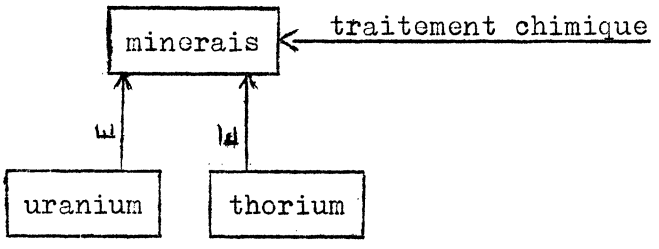
A concerne la prospection des minerais

B concerne le traitement chimique de ces mêmes minerais et notamment par fluoration.

Dans l'autre catégorie de minerais, on ne trouve pas les indications nécessaires.

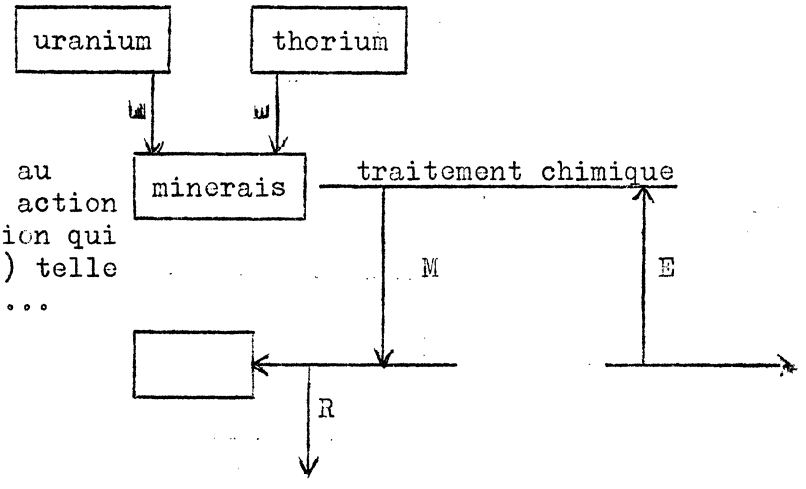
Question demandant une réponse directe

ex.: Comment s'effectuent les traitements chimiques des minerais d'uranium et de thorium ?



La partie du diagramme général que nous avons n'est pas assez complète. Le véritable diagramme général permettrait de répondre à la question par la considération des relations d'ordre E et M par exemple :

Le traitement chimique des minerais d'uranium et de thorium s'effectue au moyen de (M) telle action sur tel objet, action qui a pour résultat (R) telle transformation etc...



Traduction automatique

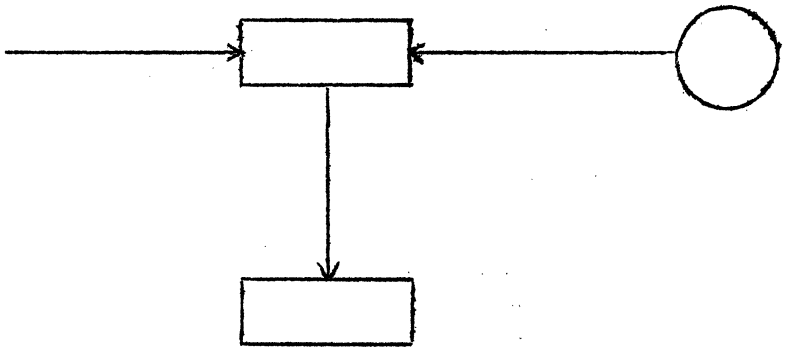
Il semble donc que la machine que nous avons conçue est maintenant capable de lire les documents, de les analyser et de répondre à n'importe quelle question scientifique. Mais nous n'avons raisonné que sur une seule langue naturelle. Or, notre but est de répondre dans la langue du demandeur. D'autre part, pour ne pas perdre d'information, pour pouvoir répondre réellement à n'importe quelle question scientifique, il faut travailler sur l'ensemble des publications scientifiques qui sont écrites dans un grand nombre de langues.

Il est bien évident que tous les langages doivent donner lieu au traitement que nous avons considéré jusqu'ici comme s'appliquant à l'analyse des textes français.

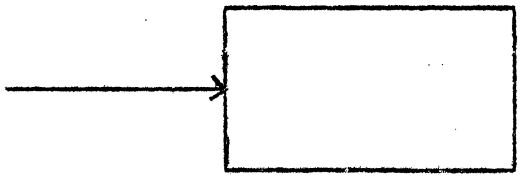
On partira du principe que les choses que l'on veut exprimer dans les langages qui nous intéressent sont les mêmes.

Rappelons-nous maintenant que les diagrammes sont censés représenter la signification réelle des textes et ne pas tenir compte de la forme signifiante. Sans doute il peut se faire qu'une disposition particulière à une langue donnée semble s'opposer à cette conception; mais ce n'est, la plupart du temps, qu'une apparence. En allemand par exemple on aura tendance à grouper des mots plus souvent qu'en français; cela donne lieu à la représentation suivante :

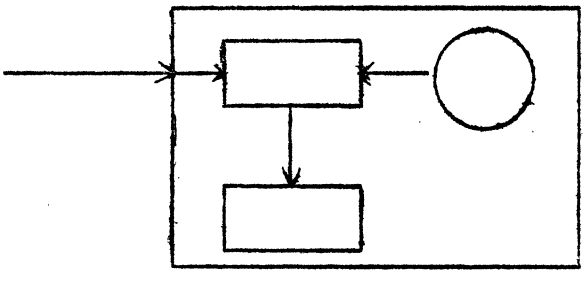
disposition française :



disposition allemande :



Mais un programme de décomposition fera apparaître la même structure que pour le français.

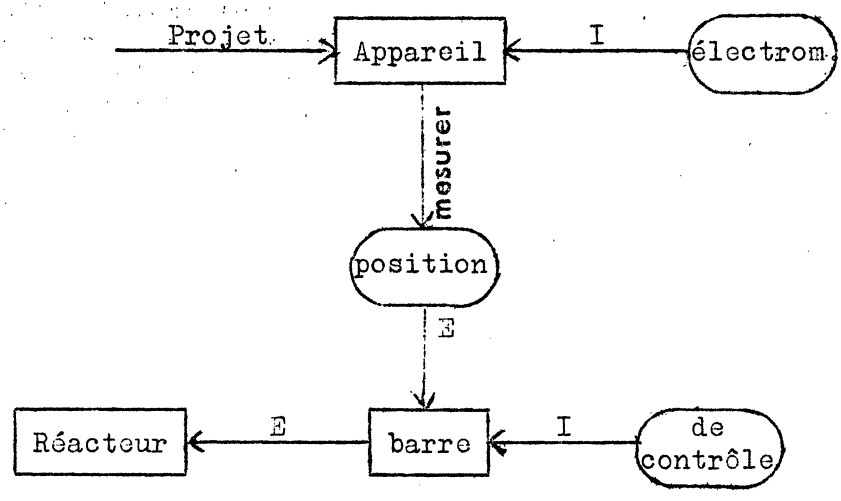


On voit ainsi que le degré de décomposition nécessaire pour chaque langue sera fixé par la considération des exigences de l'ensemble des autres langues.

Le langage des diagrammes apparaît ainsi comme un langage pouvant jouer le rôle de langage intermédiaire dans la phase traduction. En réalité, pour que nous puissions parler d'un langage intermédiaire, il faut encore que nous indiquions à quel moment il y a équivalence entre deux diagrammes correspondant à deux langues différentes.

Le gros travail qui a été effectué lors de l'analyse nous permet donc d'atteindre une partie de notre but. Il n'y a bien sûr pas lieu de se dire que toutes les difficultés sont résolues, il s'en faut de beaucoup, ne serait-ce que parce que la réalisation de l'analyse automatique elle-même demandera encore de gros efforts. Néanmoins la direction semble prometteuse.

Passage à une forme littéraire



Il est entendu qu'il faut en réalité considérer la forme équivalente en machine.

Convenons :

1. de remplacer E par la préposition "de"
   barre de réacteur
   position de barre

2. de lier directement un objet et une propriété reliés par I
   appareil électromagnétique
   barre "de contrôle"

3. de traduire une action suivant un objet par le pronom "qui" suivi de l'action (rappelons le cas des transformations linguistiques).

   On obtient :

   Projet appareil électromagnétique qui mesure position de barre d contrôle de réacteur.

Cette phrase . est assez proche d'un français correct; il n'y manque guère que les articles. Il est facile d'imaginer des procédés analogues pour les autres langues. Là encore il est tout à fait possible de raffiner la méthode, même avec des moyens relativement simples.

Conclusion

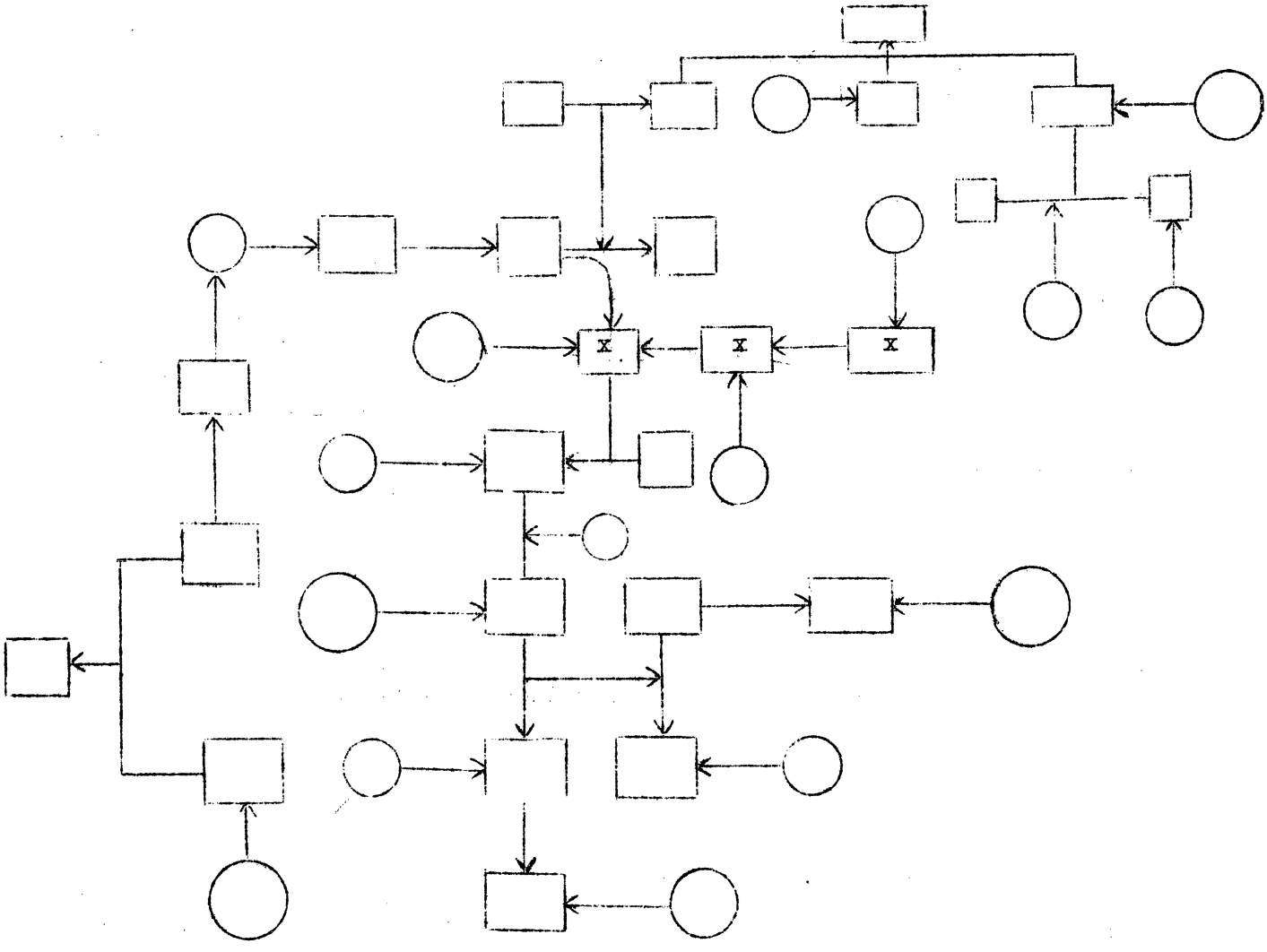
On se rend bien compte qu'il faudra des années pour mener à bien une telle entreprise d'autant plus que nous n'en avons donné que les lignes générales, et pendant ce temps des systèmes du genre de celui que M. Cardin a détaillé, et qui donnent réellement satisfaction lorsque les collections de documents ne sont pas trop élevées, seront les seuls utilisables. Cela ne veut pas dire que notre système n'est pas réalisable, bien au contraire. Nous dirons même qu'il est possible de concevoir assez rapidement un système basé sur le même principe, mais mettant en jeu des moyens beaucoup moins importants. Une des difficultés qu'il faut combattre pour cela relève de l'énorme capacité de mémoire nécessaire. Une manière radicale de réduire nos exigences sur ce point c'est de réaliser des résumés au lieu de stocker les textes entiers dans la mémoire. Bien entendu, il faut que cette réalisation soit automatique et nous en dirons quelques mots.

Résumé automatique

Nous savons déjà, par MM.Pietsch et de Grolier, que l'Américain H.P. Luhn a pu faire établir des résumés automatiquement en faisant repérer par une machine les phrases présentant les mots statistiquement les plus fréquents, et en faisant réunir ces phrases en un auto-résumé. Nous savons aussi que ce procédé est imparfait et cela se comprend car il est difficile de penser que les phrases existant dans le texte soient réellement les plus significatives que l'on puisse construire.

Cherchons donc à améliorer le procédé en nous aidant du diagramme associé au texte et en supposant que ce diagramme ait pu être construit automatiquement

Par un procédé du genre de celui de H.P. Luhn, il est facile d'obtenir les termes les plus significatifs et de les marquer en tant que tels sur le diagramme - (Ici, nous les marquerons d'une croix)



Convenons maintenant de "sortir" du diagramme les notions les plus significatives accompagnées de leur voisinage immédiat (par exemple, s'il s'agit d'un objet, devront être dégagées, en plus de celui-ci, les propriétés qui y sont directement associées, et les actions qui en partent ou qui y aboutissent, accompagnées de leurs qualificatifs et de l'objet se trouvant à l'autre extrêmité).

Il y a tout lieu de penser que les phrases ainsi construites seraient réellement les plus significatives et qu'elles n'existeraient pas forcément dans le texte initial.

Mais nous avons supposé au départ que le diagramme avait pu être construit automatiquement; or le diagramme général, indispensable pour cette construction, ne serait plus lui-même qu'un "diagramme r[illegible] du véritable diagramme général. Au moment de l'analyse, la machine [illegible] alors appel plus souvent à l'homme pour l'aider dans son travail, m[illegible]ment à chaque fois qu'elle trouverait un "contexte" ne se trouvant pas encore en mémoire.

La documentation deviendrait ainsi moins automatique, et la sélection moins complète; mais on se rend bien compte que, même dans ces conditions, une telle machine présenterait encore des qualités inestimables. Il est donc tout à fait raisonnable d'en envisager la construction, d'autant plus que rien n'empêche de l'améliorer continuellement au fur et à mesure des progrès techniques.

Cette machine pourra d'ailleurs être considérée comme le prototype de machines spécialisées dans un domaine scientifique donné, et qui pourraient convenir à de grandes institutions nationales, machines qui pourraient de plus être connectées à une machine générale, ou entre elles, pour le cas où elles ne suffiraient pas à répondre à des questions qui sortiraient de leur domaine propre.

BIBLIOGRAPHIE

(1) Gutenmakker L.I. Machines statistiques et d'information d'un type nouveau

Vesth.Akad.Naouk1956, 10, 13-21.

(2) Yngve In defense of English.International Conference for Standards on a common language for machine searching and translation.Cleveland.Sept.1959

(3) Prot M. Langage et Logique - Hermann - 1949

Je résumerai tam-

CONCLUSION GENERALE

par P. BRAFFORT

Je voudrais maintenant dessiner à grands traits les perspectives qui s'offrent à nous. Ces perspectives s'insèrent naturellement dans l'automatisation progressive des fonctions documentaires, telle que M. Leroy l'a définie. En analysant d'une façon complètement automatique les documents divers que rédigent les auteurs de rapports scientifiques et techniques, nous maitrisons complètement les connaissances élaborées par l'homme et rédigées sous forme de textes écrits. Il sera évidemment très important de pouvoir manipuler cette masse qui devient de plus en plus considérable avec la certitude de ne rien laisser échapper, puisque toutes les possibilités seront examinées par un système automatique et ceci avec une précision plus ou moins grande suivant le niveau de la population documentaire à examiner.

Mais nous pouvons déjà aller plus loin et apercevoir dans l'évolution de notre propre langage quelques indications sur ce que pourraît être demain. Nous ressentons tous que le langage de la vie quotidienne et, en particulier, le langage écrit est en train d'évoluer très considérablement. En fait, avec l'apparition du téléphone, de la radio, la quantité de langage écrit effectivement utilisé diminue relativement. Bien plus, le langage écrit change d'allure; il suffit d'ouvrir les yeux pour voir que le langage linéaire, les suites de mots, font place peu à peu aux graphiques, aux diagrammes. C'est ainsi qu'on nous indique par un simple schéma qu'il est interdit de tourner à gauche; rédigée en toutes lettres, cette interdiction demanderait beaucoup trop de temps pour être comprise, que ne permet la vitesse de nos véhicules.

Nous revenons ainsi à une forme nouvelle de hiéroglyphes qui n'est plus purement symbolique, mais pictographique. Toutefois, au lieu d'être comme les pictographies primitives un simple dessin des apparences des choses, nos pictogrammes actuels ne calquent pas les apparences mais les structures. Nous nous efforçons instinctivement de créer un langage écrit qui soit analogique quant aux structures. Et déjà, ce travail que la société effectue spontanément, pour des raisons d'économie de pensée, se poursuivra parallèlement avec un autre effort qui déjà va plus loin, au delà de ce langage idéographique.

Car on peut imaginer que l'étape dans laquelle on rédige des documents scientifiques après avoir fait des expériences sur la nature peut un jour disparaître. On peut envisager que le dialogue entre l'homme et la nature puisse se faire, dans un symbolisme imposé par la nature elle-même. On peut envisager d'écouter la nature "généticienne" en utilisant le langage des chromosomes, en considérant le système des gênes comme un système symbolique (il y a déjà quelques travaux dans ce domaine). On peut écouter la nature "astrophysicienne" en utilisant le langage des spectres, etc...

Dans nos centres nucléaires, à côté des grands appareils de physique que sont les accélérateurs de particules, on voit des appareils de mesure qui ne sont pas directement lus par les hommes, mais suivis eux-mêmes de systèmes électroniques qui classent les informations et se livrent à une certaine interprétation. Vous savez que pour dépouiller les clichés de chambres à bulles qui se trouvent près des grands appareils de physique on a maintenant des machines électroniques qui observent automatiquement les clichés et qui suivent les traces laissées par les particules ionisantes dans l'émulsion photographique, qui comptent ces particules ionisantes etc... et on est proche du jour où les calculateurs branchés sur ces appareils de lecture feront des calculs de courbures de trajectoires et où la machine pourra dire, sans intervention de l'individu humain : "voici un évènement qui n'est pas interprétable à l'aide des particules déjà connues!"

Ceci n'est pas un avenir extrêmement lointain, mais un avenir de quelques années ou dizaines d'années. Plus loin encore, nous apercevons la possibilité de construire des systèmes de décision complexes qui mettent la nature à la question et qui interprètent dans un langage interne les réponses que la nature donne à ces questions, pour en déduire, dans un langage accessible à l'homme, quelles décisions il faut prendre, quelles nouvelles expériences entreprendre, quelles nouvelles conceptions élaborer.

Nous avons nous-mêmes en cours de réalisation un projet que nous appellons ETNA : c'est l'Expérience d'un Théoricien Nucléaire Automatique. Quel est l'objet de cette expérience ? Il s'agit d'accumuler d'un côté les données expérimentales relatives à la physique des noyaux légers et moyens, d'autre part les formes théoriques de la théorie nucléaire des couches, et comparer les résultats expérimentaux avec les déductions tirées de la théorie, de façon à voir, dans un premier stade, jusqu'à quel domaine s'applique exactement la théorie des couches. Il s'agit de tester une théorie par des faits, résoudre les équations qui nous sont proposées dans la théorie des couches et comparer avec l'expérience, et ceci d'une façon entièrement mécanisée. Mais notre intention est d'aller plus loin, de munir le système en question d'une boucle de contre-réaction supplémentaire, de façon à permettre au système automatique constatant que la théorie des couches ne s'applique plus à partir d'un certain domaine, de corriger le système formel destiné à décrire les résultats expérimentaux et de créer un autre langage

théorique avec un critère d'optimalité pour l'adéquation du système formel et des faits expérimentaux. Les études que nous avons déjà faites nous montrent que de tels efforts ne peuvent être poursuivis qu'avec un matériel de calcul automatique extrêmement puissant, et ce n'est donc pas avant un ou deux ans que nous pourrons effectivement réaliser sur machine de telles expériences. Bien entendu ces recherches aux perspectives lointaines sont poursuivies côte à côte avec des travaux beaucoup plus terre-à-terre et qui donnent un plein emploi à notre matériel de calcul électronique : calculs de réacteurs, calculs de grandes installations chimiques, justifiant l'utilisation de ces machines du point de vue financier.

C'est dire que nos ambitions, pour grandes qu'elles soient, s'accompagnent du souci de garder le contact avec des réalisations immédiates. Ceci me permet de revenir enfin sur ces précautions, qui ne sont pas seulement des précautions de langage, sur lesquelles nous avons insisté à plusieurs reprises au cours de ces leçons. Encore une fois ce séminaire n'a pas été conçu comme un ensemble de dogmes que nous distribuons ex cathédra. Ce n'est pas du tout la bonne parole que nous voulons déverser à notre auditoire, c'est seulement un dialogue que nous voulons entamer en indiquant quelles sont nos méthodes de travail et nos perspectives et en nous efforçant d'obtenir de partout les suggestions, les nouvelles idées, les nouvelles voies à développer aussi. Et dès maintenant je crois que vous êtes tous persuadés que ces nouvelles voies sont réellement diverses et doivent le demeurer.

Qu'il n'y a pas qu'une solution, mais tout un arc-en-ciel de solutions qui correspond à l'arc-en-ciel des situations en ce qui concerne l'information scientifique. Mais nous pensons qu'il y a une possibilité d'unifier toutes ces études partielles et que l'information scientifique entièrement automatique est un espoir qui est réalisable dans quelques années, que la fusion du calcul numérique et du traitement des informations non numériques se fera comme le laisse prévoir le développement du calcul littéral automatique (quelques travaux, notamment aux Etats-Unis, sont déjà fort avancés dans ce domaine). Nous pensons aussi que les systèmes de mécanisation partielle ou à petite échelle, qu'il est indispensable de continuer de développer et à étudier pour nos laboratoires, pour nos institutions d'importance moyenne, peuvent être intégrés, peuvent être rendu compatibles et cohérents avec un système plus général, de façon à prévoir un dispatching des recherches d'informations.

Nous croyons donc qu'une intégration de ce genre est possible et ce que nous voudrions faire maintenant c'est, périodiquement. rassembler ceux d'entre vous qui sont désireux de participer à notr effort commun, mais en nous limitant cette fois, puisque le dépar pris, à des sujets ou à des techniques plus particulières. C'est dire que nous ne pensons pas dans l'immédiat refaire un séminair général et aussi vaste, mais prendre quelques questions les unes

les autres, par exemple le problème de la transformation de l'information linéaire du langage en information "diagramme", et la programmation des informations bidimensionnelles. C'est là un problème très important dans la perspective que nous évoquions tout à l'heure de voir se développer le langage écrit sous une forme graphique, bidimensionnelle.

Nous avons encore un problème qui pourrait faire l'objet d'un séminaire particulier, ce serait celui de l'analyse automatique des textes dans les langues naturelles. Il y a enfin les problèmes plus mathématiques de l'estimation des stratégies de sélection. On nous propose souvent des procédés documentaires merveilleux, tout au moins pour leurs auteurs, en donnant des arguments purement "philosophiques". Il est indispensable, lorsqu'on propose des investissements, de posséder une méthode d'évaluation objective de ces systèmes. D'où la nécessite de développer certains aspects de la théorie des jeux et de la recherche opérationnelle pour estimer ces problèmes, pour exprimer ces stratégies de recherche. Lorsqu'on regarde les choses un peu plus techniquement, on s'aperçoit qu'on ne se trouve pas exactement dans le cas d'une théorie des jeux à deux personnes, mais dans le cas d'une théorie ne respectant pas la hiérarchie des types de Russell, l'un des joueurs se situant à un certain niveau et l'autre étant une "métapersonne". Cela pose des problèmes techniques qui n'ont pas encore été résolus.

Le traitement automatique de l'information sous toutes ses formes est, on le voit, une entreprise gigantesque qui met en oeuvre toutes les disciplines scientifiques et toutes les techniques modernes.

Pour rassembler tous ces moyens, il faut aussi rassembler toutes les compétences, et c'est bien là la vocation d'Euratom qui prépare, à l'échelle de nos six pays, une collaboration dont l'aboutissement se situe, n'en doutons pas, à l'échelle du monde entier.

Séparé aussi il est

/61 f