Trente ans d'analyse informatique de textes : Où en est-on ? Et après ?

par

R. BUSA

CAAL - GALLARATE

I. INTRODUCTION

1.1. Notre spécialisation est internationale, supra-confessionnelle, disons oecuménique. Bien plus, elle est sous l'égide de l'amitié. Je le proclamais en latin en 1974 dans l' "ALLC Bulletin", 2, 2 - Summer 1974 - p. 2. Cette amitié nous a réunis encore une fois ici aujourd'hui.

1.2. Mon propos est de faire le point : où en sommes-nous après trente ans d'analyse automatique de textes ? J'essaierai d'ébaucher une carte géographique. Et d'une carte géographique, on ne doit attendre ni des nouveautés ni des choses spectaculaires.

On réalise partout des analyses du discours humain sur ordinateur, il n'est même pas possible de se tenir au courant de tout ce qui se publie sur la question. Or, si le nombre de recherches est grand, peut-on dire la même chose de la qualité des types d'opérations ? Pour répondre, je dois rappeler trois points de linguistique :

a) types de "sémanticité" des mots,
b) six paramètres du langage,
c) distinction entre informatique documentaire et informatique linguistique.

1.3. Laissons de côté la communication du dialogue humain. Je l'appelle communication par présence, où les deux pensées se touchent et où tous les intermédiaires sont transparents. Comme nous nous occupons seulement des textes, nous nous référons à la situation du lecteur. Dans ce cas, la pensée de l'auteur du texte est absente et c'est à l'intérieur de ma pensée que je dois reconstruire ce qu'il y avait dans la sienne. En ma présence, j'ai seulement ses signifiants et les signifiés qui leur correspondent en moi.

1.4. Il y a là ce que j'appelle le triangle lexicographique, c'est-à-dire le problème herméneutique. En voici les trois angles :

1) sur la base des signes verbaux communs (je connais la langue de l'auteur);
2) avec ce qu'ils signifient dans ma tête;
3) je dois saisir ce qu'ils signifiaient dans la tête absente de l'auteur.

Par exemple, nous avons eu la surprise de nous apercevoir que les mots célèbres "ratio seminalis" -, auxquels nous avions appliqué notre méthode herméneutique -, dans la tête de St Thomas d'Aquin, qui vivait il y a sept siècles, signifiaient ce qu'aujourd'hui nous entendons par "code génétique" ou "programme génétique" : évidemment pas l'ADN, mais la fonction logicielle.

2. Parmi les mots d'un texte, nous sommes forcés de reconnaître l'existence de plusieurs degrés, ou types, ou densités de sémanticité; c'est-à-dire du rapport signifiant-signifié. J'en indique trois : chiffres, noms propres, mots communs. Je subdiviserai ce dernier groupe ultérieurement.

2.1. Les chiffres et les autres symboles mathématiques ont un contenu qui, bien que mince, est homogène et constant à l'intérieur du même texte et qui ne change pas selon la "syntaxe" du

contexte. On pourrait dire la même chose des notes de musique.

2.2. Avec l'écriture alphabétique, il y a d'abord les noms propres, qui signifient en bloc une réalité individuelle entière et complète, en tant qu'unique. Mais ils ne la qualifient pas. On peut y rattacher les acronymes, comme CNRS ou IBM.

2.3. Aux noms propres, j'oppose les mots communs avec lesquels on peut signifier plusieurs choses distinctes ou même différentes. Je les subdiviserai ultérieurement en trois classes de sémanticité.

2.3.1. Viennent d'abord les mots déictiques, avec lesquels on exprime sa connaissance en indiquant une chose, de la même façon que la mimique correspondante de "pointer son doigt vers" : celui-ci, celui-là, moi, toi, lui, etc...

2.3.2. Ensuite viennent les substantifs universaux, dont chacun signifie en bloc un objet concret, individuel, entier et complet, comme cheval, arbre, homme, pierre, ange, etc..., ou une de leurs parties physiques, comme bras, tête, roue, etc ..., ou un de leurs collectifs, comme foule, prairie, etc...

2.3.3.a Finalement viennent tous les autres mots qui signifient n'importe quelle dimension, qualité, aspect, activité, ou forme de n'importe quel objet complet. Il y a là tous les adjectifs, les verbes, les adverbes, les prépositions etc..., mais aussi les mots substantivés qui signifient les mêmes valeurs, comme la justice, la direction, l'accélération, etc... Ces mots ne signifient pas des objets complets en soi, mais des formules de ressemblance, des aspects qu'on trouve dans les choses et par lesquels les choses diffèrent ou se ressemblent. De ces formules de ressemblance, on possède une connaissance qui en est l'idée ou concept ou définition. La définition des objets concrets est toujours un composé de ces concepts.

2.3.3.b C'est dans ce dernier groupe que l'on trouve les problèmes lexicologiques et philosophiques les plus ardus pour le classement typologique. Les taxonomies des objets dans les trois règnes de la nature ne seront jamais aussi floues que celles des mots en linguistique.

Je citerai trois classements qui sont indépendants les uns des autres; tout le monde connaît les deux premiers, quant au troisième, je travaille encore à le préciser.

En effet, on distingue les neuf parties du discours : article - nom - adjectif - pronom - verbe - adverbe - préposition - conjonction - interjection. On distingue aussi les mots-outils, ou fonctionnels ou grammaticaux, des mots pleins ou définis ou absolus. Mais il faut aussi distinguer entre mots spécifiques d'un message du discours et mots qui sont communs à tout message de n'importe quel discours.

Il est certainement très rare de trouver des mots comme "octosyllabique" ou "mildiou" dans un compte rendu de football, ou le mot "football" dans l'exégèse du Cantique biblique. Mais on trouve partout les prépositions, les articles, les pronoms, les verbes auxiliaires et les co-verbes.

Tout le monde comprend que ces mots sont l'ossature, la trame, la charpente de tout discours. Ils sont les couches instrumentales du langage, c'est-à-dire ceux "avec lesquels" on dit "ce que" l'on a à dire. Mais il faut ajouter à ce groupe quelques mots pleins, c'est-à-dire ceux qui expriment les généralités de toute réalité, en tant que telle : tout et partie, un et plusieurs, actif et passif, principe, cause, fait, conséquence, égal, semblable, différent, possible, impossible, nécessaire, type, façon, genre, espèce etc...

Depuis l'époque d'Aristote, on a isolé, parmi ces substantifs et ces adjectifs, ceux que l'on a appelés les transcendentaux, dans un sens qui n'est pas le sens moderne du mot ; la qualification de "transcendental" était appliquée aux mots :

a) qui transcendent toute partition catégorielle des êtres,
b) dont chacun équivaut aux autres.

La scolastique disait que *"res, ens, unum, bonum, aliquid, verum"* et aussi tous leurs synonymes, propres ou métaphoriques(1), *convertuntur.*

3. Six autres phénomènes et paramètres du langage interfèrent dans ce que je viens de dire, bien qu'ils ne soient pas sur le même plan. En effet, la taxonomie, la synonymie et la métaphore dépendent plutôt de la nature des choses qui ont des ressemblances et des corrélations. La vicariance, la syntaxe et le discours dépendent plutôt de la nature de la pensée. Mais tous affectent l'informatique.

3.1. En verticale, j'appelle taxonomie la hiérarchie pyramidale ou arborescente des mots, qui montent du plus particulier au plus général. Par exemple, quand je dis "les animaux", les chiens, les dauphins, les oiseaux etc... y sont compris.

3.2. La synonymie - (j'évite délibérément d'y mêler la terminologie de l'analogie) - est horizontale : une même chose ou valeur est signifiée par des mots différents. D'où les champs notionnels ou constellations des mots, qui ont la même signification ou presque, par exemple angle et coin. Il ne faudrait pas se laisser aller à les confondre avec les mots qui expriment un réseau de corrélations d'un même objet.

3.3. L'emploi métaphorique est opposé à l'emploi propre d'un mot. Un même mot signifie d'abord une chose, mais ensuite des choses bien différentes à cause de quelque ressemblance proche ou lointaine : les fleurs du jardin et les Fleurs du Mal.

3.4. La vicariance est un autre aspect horizontal du langage. Les pronoms semblent être les principaux mots vicaires. Quand on chante : *"Adjutorium nostrum in nomine Domini, qui fecit coelum et terram"*, le mot *"dominus"* est présent deux fois, en personne dans la première proposition, par procuration dans la deuxième. Les mots sous-entendus se rattachent à cette catégorie, car ils sont "présents par leur absence".

3.5. Par syntaxe, j'entends ici la structuration des mots en propositions et en phrases. Nous parlons par propositions. La proposition est le niveau le plus bas d'unité d'ensemble de nos expressions. Les mots individuels sont seulement des éléments et des composants (sauf bien entendu les interjections, les exclamations, les titres, les listes etc...).

3.5.1. La proposition a trois fonctions :

- elle ôte l'ambiguïté aux mots. Ainsi : elle porte et la porte;
- elle "ajoute" les fonctions grammaticales et les fonctions logiques : substantifs, adjectifs, verbes, etc... sujet, prédicat, objet, compléments, etc...
- elle "ajoute" aussi toutes les précisions qui définissent les contours du contenu sémantique de chaque mot, à la suite de son incorporation dans l'ensemble. Et c'est pourquoi toute recherche sémantique a besoin de concordance.

3.5.2. Cela implique que chaque mot acquiert dans la proposition des couches de significations qu'il ne possède pas encore quand il est isolé de tout contexte, c'est-à-dire quand il est passible d'emploi, comme les briques gardées dans un entrepôt.

La différenciation entre emploi-substantif et emploi-adjectif, et même emploi verbal et emploi nominal est syntaxique. Elle ne tient pas -scientifiquement- quand on essaye de l'introduire dans un thesaurus de mots individuels passibles d'emploi. On la trouve, bien sûr, dans les dictionnaires courants, mais comme dérivée, ou d'un inventaire de son usage, ou des contraintes sémantiques, relevées ou imaginées. Monsieur Jacques de Chabannes, Seigneur de la Palisse serait tout à fait d'accord pour dire que les parties du discours existent seulement dans le discours. Elles ne sont pas des catégories adéquates pour les mots atomisés dans un dictionnaire.

En tout cas, il est absolument nécessaire de distinguer dans chaque mot et pour chaque mot, les niveaux de sens possibles qu'il a comme forme et qu'il garde toujours et partout, et les niveaux de sens qui lui sont ajoutés par la syntaxe et qui changent selon le contexte.

3.6. J'appelle discours tous les ensembles progressifs dans lesquels les propositions peuvent être groupées : phrases, alinéas, paragraphes, chapitres, etc... articles, résumés, etc...
Chaque proposition individuelle est déjà un ensemble; un tout en soi, un petit système structuré complet. Je parle de système, quand plusieurs choses différentes sont rassemblées en unités opératives dans un même but. Il est vrai que chaque mot individuel est aussi un petit système en soi -(disaient = plusieurs + dans le passé + dire + à + quelqu'un)-, mais pour ne pas se noyer dans la philosophie de l'un et du multiple, il suffira d'établir une équation mots/proposition = atomes/molécule.

3.6.1. Dans un système, le tout contient plus que la somme de ses parties, car il contient la formule de leur organisation. Avec les mêmes mots, des auteurs divers écriraient des textes divers, de même qu'avec les mêmes briques, des architectes divers feraient des maisons diverses.

3.6.3. Dans l'arc opératif de l'expression qui va de la pensée au texte, l'ensemble global du sens se trouve aux extrêmes dans deux situations très différentes.

a) Dans le texte écrit, le discours est séquentiel et linéaire, mot après mot, comme les notes dans la musique ou les photogrammes sur l'écran du cinéma. Au contraire, la pensée est multidimensionnelle, où tout tend à être co-présent à tout, lié à tous les autres et où il n'y a pas un seul centre, car chaque élément peut être choisi comme centre du tout.

b) Le système du discours est fermé et généré, celui de la pensée est ouvert et génératif ... De plus, dans la pensée, tout est co-présent, par conséquent dans les discours, il y a les "présuppositions"...

4. Il faut aussi se rendre compte qu'il y a une différence entre l'informatique documentaire et l'informatique linguistique proprement dite.

4.1. J'appelle documentaire toute élaboration qui traite les mots comme mots-clés. Les catalogues et les banques de données en sont un exemple. Le paquet logiciel STAIRS-IBM est de ce type. On pourrait dire que les mots-clés sont traités comme s'ils étaient des noms propres ou des mots-étiquettes : on ne s'y occupe pas de la syntaxe ni des variations de sens qui en dépendent, on s'occupe encore moins du texte en tant que tel. On se limite à rassembler tous les morceaux qui contiennent un mot donné ou un groupe de mots donnés.

4.2. Au contraire, l'élaboration linguistique s'occupe des mots, mais en visant le discours et, par le discours, le message conceptuel dont il est l'expression.

II. OU EN EST-ON ?

5. Faisons donc le point de l'informatique en suivant les trois degrés : chiffres, noms propres et mots communs.

5.1. C'est dans le traitement scientifique ou administratif des chiffres que l'informatique a eu son triomphe. Cette informatique a donné l'essor à l'explosion technologique qui multiplie nos espoirs et nos terreurs. Le traitement d'autres symboles, comme les notes musicales par exemple (bien qu'il y en ait très peu comparativement) n'offre pas plus de difficultés que le calcul numérique.

5.2. L'informatique des noms propres, individus ou sociétés, par exemple auprès des organismes nationaux de nos "chers" taxes et impôts, n'a pas rencontré plus de difficultés de principe que la composition informatique des annuaires téléphoniques. Ils ont dû tout simplement pousser jusqu'au fond des exigences informatiques, la systématisation des noms, prénoms, titres, adresses, etc..., à peu près comme avaient déjà fait les bibliothécaires avec les données bibliographiques.

5.3. Mais pour les mots communs, même l'informatique documentaire s'est déjà heurtée à de gros problèmes. Je les réduis à deux problèmes qui dépendent de paramètres taxonomiques et

synonymiques du lexique. Je vais les expliquer par deux exemples.

5.3.1. Dans une banque de données, pour chercher tout ce qui se réfère au chien, je dois établir un lien entre le mot "chien" d'un côté et le mot "bulldog" et "pékinois", etc... Pour chercher tout ce qui se réfère au mammifère, j'en dois lier le mot avec les mots chat, chien, cheval, vache, etc...

5.3.2. Dans la même banque de données, si je voulais me documenter sur les lois du domicile, je ne pourrais pas me limiter à la recherche du mot "domicile", car son champ notionnel contient d'autres mots, comme logement, logis, habitation, résidence, appartement, etc...

5.3.3. Dans les deux exemples, il y a en plus le problème du repérage des endroits où le mot-clé est représenté par un mot vicaire; ce qui est très fréquent avec l'emploi des pronoms.

5.4. Finalement pour préciser à quel point l'informatique linguistique proprement dite est arrivée, je vais d'abord ébaucher un schéma de ses étapes, rangées en trois niveaux :

- traitement des mots individuels;
- traitement des mots associés;
- traitement des ensembles de mots.

Bien entendu, je prends le mot "mot" dans sa valeur informatique d'une séquence de graphèmes séparée des autres -(dans nos langues)- par des espaces ou des ponctuations.

La saisie du texte demande d'abord la définition de son système graphémique : les lettres, les pro-lettres, tous les signes que j'appelle accidents, car ils tombent sur le même espace que la lettre, comme les accents, la cédille, les voyelles hébraïques et arabes, etc..., les ponctuations, les codes typologiques et toute autre particularité graphique, comme les majuscules, les caractères gras, italiques, soulignés, etc...

La saisie ouvre une alternative. On peut considérer cette séquence de graphèmes comme une série physique d'une poussière de petits éléments : cela équivaut pour ainsi dire à la traiter comme un alphabet inconnu ou comme un message chiffré. Mais on peut aussi décider d'élaborer ces signes comme ayant un sens, c'est-à-dire comme mots d'une langue connue.

Dans le premier cas, on peut y appliquer plusieurs types de calculs mathématiques et statistiques, sans faire encore de la linguistique, c'est-à-dire sans s'occuper de leur valeur sémantique. On pourra aboutir peut-être à la conclusion qu'il s'agit d'un texte et d'une langue à confier à un examen sémantique.

Dans le deuxième cas, on est déjà plongé dans la linguistique, car on s'occupe des signifiés et des signifiants. Cette dernière hypothèse est la plus courante parmi nous.
Voici donc ses niveaux et ses étapes.

5.4.1.a La détermination du paquet de sens que chaque mot contient en soi indépendamment du contexte est le résultat de ce que nous appelons lemmatisation et qui, par la nature des choses mêmes, ne peut être que "morphologique". Elle devrait déterminer quatre éléments :

- à quel lemme appartient chaque forme univoque ?
- quels sont les lemmes auxquels appartient chaque forme homographe ?
- à quelle catégorie flexive du même lemme appartient une forme qui peut en représenter plusieurs ? Comme par exemple, en latin, *"veni"* peut être impératif ou indicatif parfait.
- quelles sont les catégories morphologiques de chaque flexion ?

5.4.1.b Après, il faut déterminer les niveaux de sens que chaque mot reçoit par sa structuration dans la proposition, notamment :

- le genre littéraire du contexte;
- la typologie des différentes parties, disons, rhétoriques d'un texte : titre, sommaire, introduction, discours connectif, etc...;
- la typologie des multiples possibilités de ces corps étrangers que sont les citations;
- la précision de la fonction grammaticale : en effet certains mots dans le même discours, tels des acteurs de la même compagnie, opèrent comme substantifs dans une pièce, et comme adjectifs dans une autre;
- la précision de la fonction logique : le même mot ici est le sujet, là c'est le prédicat, ailleurs c'est le complément, etc...

Tout cela, en lisant, nous l'apercevons sans même réfléchir, car, ce que nous arrivons à comprendre d'abord, c'est le sens d'ensemble des propositions. Mais pour pouvoir l'élaborer à la machine, il faut l'avoir codifié avec des signes informatiques. Ces codes sémantiques des mots peuvent, avant tout, être ajoutés à la main et cela demande quatre types d'intervention humaine :

- une rédaction préalable du texte;
- la lemmatisation des formes des mots;
- le tri des homographes;
- l'analyse des contextes dans la concordance ou du texte ligne après ligne.

L'ampleur de ce travail humain a obligé, dès le début, à se poser la question suivante : est-ce que l'on pourrait reconnaître toutes ces catégories sémantiques avec des programmes d'ordinateurs ?

5.4.2. Par traitement des mots associés (en tant que sous-ensembles des propositions), je n'entends pas les groupements physiques ou topographiques, c'est-à-dire les séries des mots qui sont l'un à côté de l'autre, mais j'entends les mots associés sémantiquement, contigus ou non.

Il y a plusieurs degrés de "syntagmes" :

- des mots associés sont à considérer comme un seul mot au même niveau et au même titre que les mots individuels : par exemple Charles de Gaulle; chien-loup et surtout les formes verbales composées;

- des locutions idiomatiques ont un sens qui ne correspond pas au sens des mots composants, comme en italien "mamma mia", ou en français "une chienne de vie". D'autres locutions gardent les sens propres de leurs composants, comme l'expression "par exemple" et "grâce à";
- les corrélations grammaticales ou syntaxiques sont aussi des syntagmes; comme par exemple entre le substantif et son adjectif, entre le verbe et son sujet, ou lorsque plusieurs mots sont coordonnés dans la même proposition;
- finalement, les corrélations entre les mots vicaires et les mots qu'ils représentent, y compris les corrélations des mots sous-entendus.

4.3. Dans le discours, les ensembles sont progressifs, à partir de l'élément minimal qu'est la proposition jusqu'à l'ensemble du texte entier. Dans le texte entier, il y a deux aspects à bien distinguer : le style en tant que forme et formule globale et complexive de l'expression, et le contenu ou message du texte. Le style a comme squelette d'une part le choix des mots, et de l'autre leurs quantités, leurs proportions et leurs distributions. La totalité du message est abrégée par le sommaire, par le résumé, par la liste des mots-clés, parfois par le titre.

5. La réponse à la question que nous nous sommes posée comme thème est que, aujourd'hui, la linguistique informatique :

- possède déjà les outils nécessaires pour élaborer les mots individuels selon leur lemme et leur morphologie, aboutissant à la description exhaustive du premier niveau du système lexicologique;
- a commencé à essayer de reconnaître avec des programmes soit les précisions des mots ambigus qui dépendent de la syntaxe, soit les niveaux syntaxiques du sens de chaque mot avec ses fonctions grammaticales et logiques;
- a commencé à programmer des outils pour détecter automatiquement des associations sémantiques entre mots individuels;
- piétine devant le mur de la formalisation et de la reconnaissance automatique du sens global des phrases et des textes.

Quelqu'un a déjà dit que ce n'est pas beaucoup en principe et que dans le royaume de l'informatique, la linguistique est encore dans la situation de Cendrillon. D'autres ont dit que les Indices et les Concordances, même lemmatisées et même accompagnées par le système lexicologique complet, sont en principe des réalisations pauvres, initiales et trop faciles. En tout cas, il est vrai que nos programmes ont à peine commencé à déborder les niveaux morphologiques des mots individuels.

Si vous êtes d'accord sur cette idée qu'il y aurait une véritable explosion d'une "industrie de l'information", le jour où nous pourrions pratiquer l'indexation et le résumé automatiques, vous serez aussi d'accord que ce jour est encore caché dans un avenir lointain.

6.1. Il est évident que la raison d'une telle situation est la nature intrinsèque du langage. En effet, d'un côté, les signifiés dans la pensée sont une force générative et créative de type artistique, multi-dimensionnelle, co-présente, pluricentrique, capable d'investir la fonction expressive des

choses (marques d'encre, bruits, ou bits) qui ne sont ni les pensées des interlocuteurs ni les signifiés; cette force est capable d'organiser avec un contrôle unitaire finalisé n'importe quelle multiplicité dans des ensembles toujours nouveaux et renouvelés.

D'un autre côté, tout texte est un grand nombre, une poussière de réalités physiques chargées de la fonction symbolisante, non par nature, mais par l'impératif et le choix de la pensée. Et ces réalités physiques devenues signes, sont en ordre séquentiel. On pourrait aussi dénoncer que probablement, il y a là quelque contribution de la faiblesse humaine, comme vient de le remarquer le mathématicien-physicien américain Clifford A. TRUESDELL (John Hopkins Univ.)(2).

Bref, il reproche ce que j'appellerais le "mythe" de l'ordinateur, qui pousse à s'en servir sans assez d'intelligence.

Il faut avouer que cela est vrai, parfois même chez nous. Par exemple, trop souvent on réduit la "machine readable form" à la saisie des graphèmes des mots et des ponctuations. Mais, quand c'est nous qui lisons, l'intuition des sens globaux des phrases nous apporte un tas d'informations qui ne sont pas représentées par des signes graphiques spéciaux. Par conséquent, un texte n'est vraiment en "machine readable form" que lorsque l'on en a codifié la typologie du discours, lemmatisé les mots, et défini le système morpho-lexicologique des mots individuels. Comparons donc tout ce qui a été mis sur bandes magnétiques avec le peu de systèmes lexicologiques intégraux qu'on a publiés et nous reconnaîtrons alors que, même chez nous, s'est insinué le mythe d'employer l'ordinateur pour réduire l'emploi de notre intelligence.

Après avoir publié mes 12.000 pages qui contiennent les cinquante tables du système lexicologique de mon Index Thomisticus, c'est avec reconnaissance et admiration que j'ai vu publier par exemple, le Dictionnaire fréquentiel du L.A.S.L.A. et toutes les tables que Paul Tombeur a déjà publiées ou est en train de publier.

Je l'ai ressenti comme une approbation flatteuse de mes efforts et une confirmation de ce que je viens de dire : un texte n'est vraiment en "machine readable form" que quand on en a défini le système lexicologique.

C'est avec une extrême curiosité que j'ai commencé à comparer le pourcentage de fréquences du L.A.S.L.A. avec mes propres fréquences.

III. ET APRES ?

7. Je ne parlerai pas des interprétations linguistiques ou philologiques pures, mais des interactions recherche-ordinateur, c'est-à-dire des niveaux et étapes informatiques de ces recherches.

Je vais grouper les types qualitatifs des futures recherches de notre spécialisation en trois grandes orientations.

- la première orientation est la statistique globale des textes;
- la deuxième est de poursuivre, dans la progression qui cherche à atteindre les couches syntaxiques de la sémanticité des mots individuels, les associations et vicariances sémantiques des mots, les sens globaux des ensembles;
- la troisième est une systématisation lexicologique du lexique général, qui en détache les mots à lier en taxonomies arborescentes et les mots à lier en champs notionnels.

7.1. Sur le sujet de la statistique globale d'un texte ou d'une oeuvre, on a tenu un workshop ("atelier") les 5, 6, 7 juin derniers, chez moi à Gallarate. Par statistique globale, j'entends une analyse statistique :

- de tous les éléments dont on peut mesurer les quantités : formes, lemmes, associations, longueurs, etc ...;
- qui, d'abord, relève chaque élément dans chaque tranche homogène du texte;
- qui, ensuite, ira relever, progressivement, le même élément dans les ensembles progressifs du texte;
- jusqu'à le relever dans l'ensemble final de deux façons : l'une donnera la formule de l'ensemble en gros, brute, l'autre la formule totale, spécifiée selon les relevés de toutes les tranches;
- et qui finalement, rassemble et fond tous ces pourcentages des mots et des éléments individuels dans des totaux par types et par catégories de mots;
- et ces totaux dans une formule statistique, générale et complexive de tous les mots et de tous les éléments.

Le point de départ de ce parcours est la description des premiers niveaux du système lexicologique dont je viens de parler, mais nous n'en sommes encore nulle part.

Une telle recherche est encore une forêt vierge, elle est encore toute à faire, il y a là un grand nombre de thèses possibles pour des doctorats en linguistique computationnelle...
En effet, elle correspond à l'effort d'une description statistique globale du style de l'ensemble.
Elle pourrait rendre au moins trois grands services :

- pour la connaissance pure du style;
- pour les recherches d'authenticité, de chronologie et d'empreintes personnelles de l'expression;
- mais surtout pour le relevé de standards de normalité dans l'emploi docimastique dans tous les domaines où l'on doit évaluer le comportement humain.

7.2. Comparons le nombre de recherches de programmes d'analyse syntaxique (je donne à ces mots le sens le plus général possible) au nombre de textes qu'on a déjà sur bande. On verra que le rapport est de "trop peu a beaucoup trop". La digestion linguistique informatique des textes est lente et laborieuse.

Bien que ce ne soit pas mon domaine spécifique, je pense cependant qu'il n'y aura pas beaucoup à ajouter à ce qu'on peut lire dans les revues suivantes :

- American Journal of Computational Linguistics,

- SIGART Newsletter (Special Interest Group on Artificial Intelligence of the Assoc. for Computing Machinery),
- Artificial Intelligence (North-Holland Publ.),
- Cognitive Science (Ablex Publ., New-Jersey),
- Newsletter del Gruppo di Lavoro su Intelligenza Artificiale della Assoc. Ital. Calcolo Automatico (1981),
- IEEE Transactions on Man-Machine Communication.

A mon avis et à moins d'une explication meilleure, une des raisons de cette lenteur est le fait que ces recherches demandent beaucoup de temps de travail humain et de travail organisé en équipes. Il est vrai qu'il y a de bonnes raisons qui forcent trop de chercheurs à publier vite, mais il faut bien qu'on trouve les moyens de se dégager des Fourches Caudines qui, à présent, nous laissent passer au compte-gouttes.

Mais, à propos de ces recherches, j'ai déjà publié ailleurs l'observation suivante : elles semblent être un kilomètre de formalisations et de calculs à la verticale sur un centimètre de base. Je pense qu'il vaudrait mieux gagner centimètre par centimètre sur un kilomètre de base. En d'autres termes, à mon avis, la quantité de calculs est disproportionnée à l'information inductive qui fournit les données de fait qui en sont la base.

Tout langage étant une activité du type de la création artistique, pour aboutir à des probabilités appréciables dans la pratique, il me semble qu'il est nécessaire de recenser d'abord à la main plusieurs textes, différents et très étendus, de plusieurs millions de mots et d'y relever, à la main, les quantités des différentes structurations syntaxiques. A mon avis, cela est nécessaire avant d'écrire des programmes qui, par après, à la vitesse des machines, exécuteraient la même analyse.

Le problème de la reconnaissance du sens global d'un ensemble a déjà été formulé comme recherche de l'indexation et du résumé automatique. On pourrait même la considérer comme indépendante des recherches d'analyse syntaxique et non comme un de leurs résultats. Mais je pense de nouveau qu'une réponse, si elle est possible, sortira d'un recensement inductif préalable d'un nombre énorme de textes, recensement qui ne peut être effectué que par un travail long, diligent, systématique et progressif, c'est-à-dire échelonné.

Tout simplement, on pourrait envisager des comparaisons entre de longs textes de forme courante et les mêmes textes réduits en forme télégraphique...

7.3. Les tâches, dont je viens d'ébaucher quelques aspects sont bien lourdes et difficiles. J'espère que la troisième direction est un peu plus facile. En effet, on a déjà partout, soit des dictionnaires de synonymes, qui peuvent donner une base aux recherches des champs notionnels, soit des taxonomies scientifiques des choses, et, par conséquent, de leurs noms, qui sont déjà développées dans maints domaines.

Bien sûr, il faudra travailler à perfectionner ces documents, parfois empiriques, en leur apportant la stystématisation exigée par l'emploi informatique.

IV. CONCULSIONS

A la veille de fixer mes yeux dans les yeux de Dieu, comme toute science est un service, je viens de vous montrer que notre spécialisation a encore beaucoup de services à rendre. Je conclus par un voeu.

Mon voeu est que nos associations forment et choisissent un groupe de spécialistes qui auraient pour but d'étudier et de proposer des plans et des stratégies pour nos recherches dans un avenir proche. Cela nous permettrait de rendre un service plus organisé et moins dispersé.

NOTES

(1) Par exemple, dans St. Thomas, voir : Summa Theologiae, *tertia pars,* 10-3-C3 et 17-2-C0; *in Metaph.* livre 11, lect. 3 et 8.

(2) Clifford A. TRUESDELL (John Hopkins Univ.), "Il calcolatore : rovina della scienza e minaccia per il genere umano". (pp. 37-65 du "La nuova ragione"), P. Rossi ed., Il Mulino, Bologna (1981) mm. 150 x 212 pp. 254, Actes du Congrès Intern. "La cultura scientifica nel mondo contemporaneo", Milan 6-8 Février 1980.