CENTRE POUR L'AUTOMATION DE L'ANALYSE LINGUISTIQUE (C.A.A.L.), GALLARATE

ROBERTO BUSA S. J., ANTONIO ZAMPOLLI

1. Le domaine de nos recherches.

1.1. Le domaine de nos recherches comprend 9 langages en 4 alphabets aussi bien »input« que »output«: les langues latine, italienne, allemande et anglaise en alphabet latin; les langues hébraïque, araméenne et nabathéenne en alphabet hébraïque, la langue grecque en alphabet grec et récemment la langue russe en alphabet cyrillique.

1.2.1. *En latin* notre programme comprend:

— les œuvres complètes de Saint Thomas d'Aquin, qui est notre travail principal: 1.700.000 lignes
— un opuscule de Saint Bernard de Clairvaux: 2.000 lignes
— la Bible, édition de la Vulgate: 100.000 lignes
— libri quattuor »Sententiarum« de Pierre Lombard: 30.000 lignes
— quelques œuvres de Boèce: 7.000 lignes
— »Liber de Causis«: 1.000 lignes
— le »Lexicon Totius Latinitatis« de Forcellini: 90.000 lemmes
— le »Thesaurus Linguae Latinae«: 150.000 lemmes
— les œuvres complètes de Sénèque pour l'Université de Padova: 50.000 lignes

1.2.2. *En italien:*

— »Testi Antichi Italiani« édition Ugolini: 40.000 mots
— les recherches lexicales et phonétiques du Dr. Antonio Zampolli sur un drame contemporain: 20.000 mots et 100.000 phonèmes.

1.2.3. *En allemand:*

— I. Kant: Prolegomena zu einer künftigen Metaphysik: 45.000 mots.
— J. W. Goethe: Farbenlehre Bd. 3, pour l'Université de Tübingen: 50.000 mots

1.2.4. *En grec:* — les œuvres complètes d'Aristote: 130.000 lignes.

1.2.5. *En anglais:*

— Nuclear Physic Abstracts, etc., pour l'Euratom: 30.000 lignes.

1.2.6. *En hébreu (en araméenne et en nabathéenne):*

— Dead Sea Scrolls: 50.000 mots
— Zorell: Lexicon Hebraicum: 9.000 lemmes

1.2.7. *En russe:*

— des articles scientifiques pour l'Euratom: à peine commencé.

1.3. S. Bernard, Testi Antichi Italiani, Kant, Goethe: sont déjà prêts entièrement.

Les textes de Qumran: presque entièrement.

La Vulgate et Forcellini: sont déjà perforés.

Pierre Lomb.: Sénèque, T. L. L., Aristote: sont seulement commencés.

Boèce et L. de Causis: ne sont pas encore commencés.

1.4. Vous serez étonnés d'apprendre que rien encore n'a été publié: ceci est dû principalement au fait que nous sommes terriblement occupés à terminer l'Index Thomisticus.

Notre travail concernant l'Index Thomisticus touche maintenant presque à sa fin, comme nous le verrons plus loin.

2. Les caractéristiques de nos recherches

2.1. Notre spécialisation couvre deux phases seulement de l'automation des recherches lingustiques:

a) la transcription de textes naturels, à partir du livre imprimé, sur bande magnétique en vue d'élaborations électroniques: et ceci pour des textes traitant n'importe quel sujet, en toute langue ou alphabet

b) le premier recensement ou inventaire intégral des facteurs linguistiques, de quelque manière qu'ils soient représentés dans les textes naturels: c'est-à-dire la compilation d'indices et de concordances de mots, de morphèmes, de graphèmes, de syntagmes, de fréquences, etc....

Il faut remarquer que ces deux fonctions sont primordiales et nécessaires pour n'importe quelle recherche automatisée de linguistique pure ou appliquée: c'est-à-dire aussi bien pour les recherches lexicales ou psychologiques, que pour l'information retrieval ou la traduction automatique.

2.2. Nous avons l'intention de définir, par l'expérience faite, la méthodologie, les implications, les temps et les coûts nécessaires pour élaborer électroniquement en tant qu'une unité, des textes de plusieurs millions de mots.

2.3. De ces textes nous recensons tout ce qui s'y trouve, sans aucune exclusion: nous retenons en effet qu'une méthode rigoureuse ne permet de porter des jugements de plus ou moins grande importance sur les faits linguistiques qu'après avoir obtenu la documentation quantitative intégrale de tout ce qui, en réalité, se trouve dans un texte.

2.4. Enfin notre intention n'est pas de présenter comme résultat de notre travail un fichier où chaque mot serait porté sur une fiche, avec un

large contexte — (nous avons en effet abandonné ce projet initial, après l'avoir expérimenté sur, environ, 800.000 mots et autant de fiches) —; mais notre intention est de publier en volumes un système d'indices et de concordances, qui sera:

a) un document objectif et complet du panorama linguistique d'un texte ou d'un auteur,

b) un instrument facile à manier pour des recherches ultérieures sur ce texte ou cet auteur.

2.5. De ces quatre caractéristiques fondamentales, en dérivant cinq autres:

a) la nécessité d'une rédaction précédant la perforation,

b) la nécessité de faire la perforation avec le plus grand soin,

c) la „lemmatisation" grâce à un „dictionnaire de machine",

d) le traitement des mots, dont la fréquence et très élevée,

e) la sélection automatique des contextes dans les concordances.

2.6. Il s'ensuit enfin qu'il nous est absolument nécessaire d'utiliser des ordinateurs électroniques proprement dits et de conserver les machines à cartes perforées seulement comme auxiliaires et pour un usage marginal.

3. Les phases de nos élaborations

Pour illustrer plus clairement ces données, je résumerai d'abord les différentes phases de la préparation de l'Index Thomisticus; puis nous verrons comment nous en avons organisé la présentation finale.

3.1. Mais je dois tout d'abord exposer dans quel sens nous utilisons les termes de forme, mot et lemme:

— la forme est, pour nous, un type spécifique de séquence de symboles graphiques, délimitée par des espaces ou par la ponctuation;

— les mots sont les occurrences individuelles de chaque forme dans le texte;

— le lemme est ce qui, dans les lexiques, représente toutes les formes réunies dans un même paradigme, parce qu'elles sont les différentes flexions d'une même unité graphico-sémantique;

— il va de soi que pour les indéclinables, lemme et forme coïncident;

— »Lemmatiser« signifie donc, pour nous, attribuer à une forme les codes, en vertu desquels l'ordinateur pourra la réunir, ou pourra en exprimer l'appartenance soit à son lemme, s'il s'agit d'une forme univoque, soit à ses lemmes, s'il s'agit d'une forme homographe.

3.2. Les phases de l'élaboration de l'Index Thomisticus peuvent se résumer de la façon suivante:

(H. signifie travail seulement de l'homme; M. seulement de la machine; H.M. travail alterné de l'homme et de la machine)

H. 1. pré-édition,

H.M. 2. perforation,
M. 3. transcription sur bande magnétique,
M. 4. »lemmatisation«; tabulation de 40 % des mots,
M. 5. recherche du contexte pour les 60 % restants,
M. 6. triage alphabétique des mots avec contexte,
M. 7. lemmatisation de ces mots grâce au dictionnaire de machine,
M. 8. — liste et concordances des formes que ne comporte pas le dictionnaire de machine,
— concordances des homographes à sélectionner,
H. 9. lemmatisation de ces formes et sélection de ces homographes,
H.M. 10. perforation de ces formes et de ces homographes, révision et transcription sur bande magnétique,
M. 11. lemmatisation des mots respectifs,
M. 12. réorganisation de tous les mots dans l'ordre du texte,
M. 13. distribution de ces mots dans les différentes parties de la section: »concordances«, œuvre par œuvre,
M. 14. fusion de ces mots dans une section générale unique: »concordances des œuvres complètes«.

3.3. Dans le précédente liste des phases de l'élaboration, nous supposons déjà préparés le premier dictionnaire de machine et les programmes de l'ordinateur.

3.4. Pour la première moitié des œuvres authentiques, les phases de 8 à 11 furent effectuées, œuvre par œuvre, par cycles d'environ 300.000 mots de texte toutes les deux semaines.
Mais pour la seconde moitié, notre intention est de:
— rechercher le contexte non plus de 60 % des mots, mais seulement des homographes à sélectionner, et des »formes nouvelles«;
— effectuer les phases 8—11 non plus œuvre par œuvre, mais en un seul cycle sur tout l'ensemble du reste des œuvres authentiques, puis sur tous l'ensemble des œuvres apocryphes.

3.5. Au cours des phases 4, 7 et 11, l'ordinateur accumule sur des bandes à part toutes les formes avec leurs codes de lemmatisation et leurs totaux de fréquence, compose ensuite et imprime à partir de ces bandes les différentes parties de la section-indices.

3.6. Il ne faut pas plus oublier le caractère inévitable et capital de la correction des erreurs qui apparaissent au fur et à mesure: erreurs de mots, erreurs de lemmatisation et par conséquent erreurs dans les totaux de fréquence.

3.7. Le Corpus Thomisticum est divisé en deux groupes:
a) les œuvres authentiques: environ 8.500.000 mots
b) les œuvres (peut-être) apocryphes: environ 1.500.000 mots.

4. La pré-édition

4.1. Elle consiste pour nous, à lire le texte mot par mot et à y ajouter à la main les signes ou les symboles qui devront être perforés avec les lettres et la ponctuation.

4.2. Elle comprend:

a) corriger les erreurs d'impression du texte

b) préciser la référence chaque fois qu'elle change

c) caractériser certains symboles et certaines situations graphiques: par exemple, marquer le point qui n'est pas un point final, mais un point d'abréviation d'un mot, ou bien qui représente l'un et l'autre; marquer le tiret qui n'est pas un signe de ponctuation, mais un trait-d'union etc....

d) spécifier certains types de phrases ou certains types de mots que l'on veut mettre en évidence. Nous appelons »spécificatifs« ces signes spéciaux et, jusqu'à maintenant, nous en avons trois systèmes:

— dans l'Index Thomisticus, pour distinguer les phrases principalement selon »l'autenzia«, c'est-à-dire selon l'attribution de paternité: citations littérales, citations selon le sens, références à des titres d'œuvres, etc...

— dans les textes de Qumran, principalement selon l'état paléographique: lecture incertaine ou alternative, particularité due au scribe, mot effacé etc...

— pour les concordances de Sénèque, également pour le rapport entre les mots du texte et leurs variantes dans l'apparat, afin que variantes et mots du texte figurent ensuite, et naturellement en corrélation, dans les concordances finales

e) distinguer les »formules« des »mots«: c'est-à-dire les symboles et les expressions, par exemple arithmétiques ou géométriques, qui ne sont pas des »mots«: ainsi MI et DII qui étaient les nombres romains 1001 et 502, et non le vocatif de *meus* ni le nominatif pluriel de *deus*, ou bien AB qui signifiait le segment AB et non la préposition *ab*.

5. La perforation

5.1. Bien que nous continuions à perforer sur cartes, nous aussi, sommes convaincus qu'il est préférable pour ce genre de travaux de perforer sur bandes de papier perforé.

5.2. Mais notre préoccupation essentielle est de perforer avec le plus grand soin: comment réduire les erreurs à une quantité minime et négligeable, puisque sur une grande quantité l'absence totale et complète d'erreurs est une limite aussi irréalisable que souhaitable?
Je ferai sur ce sujet un exposé à part.

Je me bornerai à dire que, une fois le texte perforé, nous le vérifions d'abord à la machine, puis nous l'imprimons (avec l'ordinateur) deux fois de suite, et le confrontons chaque fois avec le document original, en le lisant mot par mot.

5.3. La transcription de toutes les œuvres de Saint Thomas sur des cartes perforées à reporter ensuite sur bande magnétique, a nécessité leur lecture au moins six fois mot par mot: deux fois pour la pré-édition, une fois pour la perforation, une fois pour la vérification, deux fois pour le contrôle par la lecture. Lire six fois 1.700.000 lignes de texte — puisque tel en est le nombre dans l'œuvre de Saint Thomas — équivaut donc à lire une fois 10 millions de lignes.

6. La lemmatisation

6.1. La lemmatisation, comme je disais, consiste à reconnaître et à codifier ce qui, en elle, est l'unité graphique et en même temps significative.

6.2. La lemmatisation peut être au niveau du mot ou au niveau de la forme. Pour rester dans les limites de la terminologie Aristotelicienne-Thomiste, je dirai que la lemmatisation au niveau-forme considère les mots comme »termes« de la résolution d'une proposition dans ses éléments.

Par contre la lemmatisation au niveau du mot considère des mêmes termes la »suppositio« c'est-à-dire la fonction représentative signifiante de ce terme précisément dans cette proposition.

La lemmatisation au niveau du mot ne requiert pas un »dictionnaire de machine« contrairement à la lemmatisation au niveau de la forme.

6.3. Il est évident que la lemmatisation au niveau du mot demande beaucoup plus de temps que la lemmatisation au niveau de la forme. C'est pourquoi l'analyse linguistique de textes de grande étendue doit être faite automatiquement grâce au dictionnaire de machine. Ou peut peut-être en conclure que la lemmatisation au niveau du mot peut-être prise en considération seulement comme une des méthodes pour parvenir à constituer un dictionnaire de machine suffisant. Je me permets de proposer qu'un des points de discussion de ce colloque soit la comparaison entre ces deux lemmatisations.

Pour ce qui nous concerne, je renvoie à mon exposé sur notre dictionnaire latin de machine, qui contient, en ce moment, 80.000 formes de mots latins, et sur les problèmes, dont il nous impose la solution.

6.4. Je propose aussi que soit inclus dans les points à discuter la comparaison entre concordances non lemmatisées et concordances lemmatisées.

7. Le choix automatique du contexte

7.1. Dès le début, une des principales objections contre la compilation

automatique des concordances fut la suivante: l'ordinateur ne peut pas choisir intelligemment les mots du contexte, car ce travail requiert l'habilité spécifique de l'homme qui compile; et, par conséquent, les contextes découpés automatiquement sur la base de schémas fixes ne sont vraisemblablement pas suffisants.

7.2. Je tiens à remarquer que, dans une concordance, on peut faire l'étude d'une forme ou d'un lemme sur des plans différents: par exemple le but peut-être simplement de reconnaître morphologiquement ou de définir lexicalement une forme ou un lemme; il peut être aussi la recherche doctrinale des concepts dont ce mot est le soutien.

7.3. En tout cas, je soutiens que, pour un mot qui n'est pas un apax, des contextes d'environ 100 positions, c'est-à-dire de dix-douze mots, sont pratiquement et dans leur ensemble toujours suffisants pour une définition lexicale de ce mot, si l'ordinateur a été programmé pour reconnaître les limites des phrases en fonction de la hiérarchie des signes de ponctuation ou des autres situations graphiques.

7.4. Nos programmes, en effet, organisent la délimitation du contexte pour chaque mot de la façon suivante:

— l'ordinateur cherche jusqu'à 50 positions — plus ou moins, selon les désirs — avant le mot et 50 positions après;

— dans le cas où la cinquantième position coupe un mot, il cherche à quelle extrémité on peut inclure un mot entier, si l'on ajoute les positions gagnées à l'extrémité opposée en renonçant au mot que la cinquantième position couperait en deux;

— en outre, si, avant la cinquantième position, il rencontre un point final ou un autre signe équivalent, il ne vas pas au-delà dans ce sens, mais ajoute à l'extrémité opposé les positions qui n'ont pas été utilisées à la première extrémité;

— mais si le mot se trouve à l'intérieur d'une citation littérale, l'ordinateur franchit le point final, sans cependant, aller au-delà de la citation, lorsque celle-ci est suffisamment longue pour fournir un contexte entier;

— par contre, dans le cas où le mot se trouve entre deux points rapprochés, l'ordinateur franchit le point final ou de droite ou de gauche, si la présence ou l'absence de signes spécificatifs, qu'a le mot, continue seulement à droite ou seulement à gauche; par contre il franchit le point toujours à droit, si la situation est la même des deux côtés, etc....

Nous avons avec nous des exemples de concordances obtenus avec ce programme, afin que la personne intéressée puisse les examiner.

7.5. Enfin la cohérence et la constance de son caractère systématique plaident en faveur de cette délimitation automatique des contextes: on offre

ainsi une documentation parfaitement objective et entièrement à l'abri des fluctuations et de la subjectivité inévitable du choix de l'homme: quand, par exemple, on employait vingt étudiants pour choisir les contextes pour la concordance du même auteur, comment pouvait-on garantir que leurs critères n'aient pas été différents et n'aient pas subi de variations?

8. La structure de nos indices et concordances finals

8.1. Les résultats de nos travaux, pour chacune des deux parties (œuvres authentiques et apocryphes), seront publiés, répartis en deux grosses sections: la section-indices et la section-concordances.

La section-indices contient la documentation relative aux formes et aux lemmes: elle est donc récapitulative.

La section-concordances contient la documentation relative a tous les mots: elle est donc analytique.

8.2. Avant tout on présentera les indices de chaque œuvre particulière (le Corpus Thomisticum en contient 126). Ce sont:

8.2.1. Le *laterculum formarum* (Fig. 53,54; p. 264—5) qui donnera, par ordre alphabétique:
- — toutes les formes rencontrées dans l'œuvre, lemmatisées et codifiées morphologiquement;
- — toutes leurs homographes possibles selon Forcellini;
- — les fréquences absolues de chacune, soit totale, soit détaillée par signes spécificatifs;
- — les fréquences proportionnelles;

8.2.2. Le *conspectus lemmatum* (Fig. 58; p. 269) qui donnera également et de la même façon, en les divisant en 3 groupes, pour tous les lemmes rappelés par les mots de l'œuvre:
- — les lemmes dont on aura rencontré au moins une présence effective;
- — les lemmes représentés seulement par des formes homographes dont la valeur est absente de tout contexte examiné;
- — les lemmes représentés seulement par des formes homographes sans occurrences, mais dont on n'aura pas même examiné la présence éventuelle, parce que très peu probable;

8.2.3. Le *lemmatum formae* qui présente, pour chaque lemme, la liste des formes qui le rappelle dans l'œuvre; chacune de ces formes est suivie seulement de la fréquence absolue.

8.3. Pour l'ensemble de toutes les œuvres les indices seront divisés en différents groupes:

8.3.1. Tout d'abord des listes analogues à celle du 8.2 qui représenteront un panorama complet de la terminologie thomiste.

8.3.2. Puis la liste des formes et des lemmes selon les fréquences; des formes par désinences (c'est-à-dire un »index a contrario«) et par codes mor-

phologiques; et la liste de l'homographie des formes et des désinences.

8.3.3. Enfin deux *tabulae vocum*, une pour les formes, une pour les lemmes; à coté de la colonne réservée aux formes et aux lemmes, on y trouvera autant de colonnes que d'œuvres comprises dans cette partie du »Corpus Thomisticum«; et dans chaque colonne on trouvera le chiffre total des présences de cette forme ou de ce lemme pour chaque œuvre, ainsi que son pourcentage par rapport au total des mots de l'œuvre. Ceci permettera de voir, d'un seul coup d'œil, la répartition des présences de chaque mot dans la succession des différentes œuvres.

8.4. La sectioconcordances contiendra pour chaque forme de mots au moins l'indication de tous les endroits où elle se trouve de la première œuvre à la dernière.

Nous aurons donc une concordance unique pour les œuvres complètes, et non autant de cycles et de séries de concordances qu'il y a d'œuvres: nous ne donnerons pas même pour les œuvres plus importantes une concordance à part.

Ainsi la forme *connaturaliter*, par exemple, sera suivie de tous les passages qui la contiennent, d'abord dans les trois Sommes (in Sent., c. Gentiles, Theologiae), puis dans les Quaest. Disput., puis dans les Commentaires (aristotéliens, autres et bibliques), enfin dans les Opuscules.

8.4.1. La section concordance sera divisée en cinq groupes de telle sorte que les mots énumérés dans l'un ne se retrouvent dans aucun des autres.

8.4.2. Les mots se trouvant dans des phrases qui sont des références (par ex. ut dicit Philosophus in 5 Metaph.) seront d'abord réunis dans un index locorum, qui renverra à la seconde partie de ce premier groupe; cette seconde partie renferme, dans un ordre naturellement discontinu, toutes les lignes où se trouvent des références de ce type.

Cette documentation servira a celui qui voudrait étudier quels auteurs et quelles œuvres cita Saint Thomas, et avec quelle phraséologie.

8.4.3. Viendront ensuite, par ordre alphabétique et accompagnées de l'index locorum, toutes les phrases où l'A. se cite lui-même (par exemple: ut supra habitum est, ut infra dicetur).

8.4.4. Puis tous les débuts de période que l'A. cite en référence aux diverses parties du texte commenté, comme »incipits«.

8.4.5. Ces trois premières subdivisions de la section concordances ne dépasseront pas, dans la mesure où nous pouvons prévoir, 6 ou 7 pour cent de l'ensemble des mots.

8.4.6. Le groupe suivant par contre contiendra environ 60 $^{0}/_{00}$ des mots restants. Comme je le montrerai dans un exposé à part, un petit nombre de formes — 600, c'est-à-dire bien moins d'un pour cent — ont des

fréquences si élévées qu'elles totalisent 60 % des mots de l'œuvre. Le souci d'offrir une documentation à la fois intégrale et maniable, nous a menés à réunir ces mots en petits groupes, que nous appelons syntagmes, qui sont présentés par ordre alphabétique et qui sont suivis de l'index locorum, dont j'aurai l'occasion de parler plus en détail.

8.4.7. Le reste des formes, plus de 99 %, qui cependant représentent seulement un peu moins de 40 % du texte, sera présenté selon la manière traditionelle des concordances.

Tous les contextes (d'environ 100 caractères, c'est-à-dire de 13 à 15 mots, découpés automatiquement par l'ordinateur, comme nous avons dit plus haut) qui contiennent la même forme seront réunis dans l'ordre du texte. Les formes seront ordonnées à l'intérieur de leur lemme, selon la séquence des codes morphologiques.

Nous aurons donc une concordance sur forme et non pas directement sur lemme. Nous mettrons en évidence, après la référence de chaque contexte, le code spécifique que le mot en question possédait; de cette manière, si cette phrase fait partie d'une citation littérale ou d'une citation de sens, l'usager en sera informé.