Recherche automatique de groupes verbaux récurrents et de formules dans les fichiers latins lemmatisés

Gérald PURNELLE

Depuis plus de trente ans, les progrès de l'informatique appliquée à l'étude des langues et des littératures ont procuré aux disciplines qui en dépendent de nouveaux instruments et de nouvelles méthodes. Dans ce domaine, une certaine évolution s'est fait jour. Si, dans les premiers temps, l'ordinateur fut d'abord utilisé en raison des possibilités qu'il offrait pour le traitement plus ou moins brutal de grandes masses de données linguistiques, ses premiers utilisateurs s'avisèrent rapidement qu'ils pouvaient par ailleurs exploiter sa capacité à résoudre différents types de problèmes et à analyser de manière plus précise ces mêmes données.

Cette évolution dans le sens de la précision et de la complexité se marque aussi dans la nature des données manipulées par la machine et dans celle des résultats qu'elle produit. Le premier objet étudié fut la forme textuelle, non lemmatisée, c'est-à-dire conservée telle qu'elle apparaît, à chaque occurrence, dans un texte donné; les produits fabriqués sur la base de ce premier élément sont la concordance de formes (simple liste alphabétique des formes d'un texte, accompagnées d'une référence et d'un contexte gauche et/ou droit) et la liste de fréquence de formes. Le principal progrès par rapport à cette situation de départ fut la mise au point de procédures de lemmatisation et d'analyse morphologique automatiques ou semi-automatiques[1]. La première a pour objet de rapporter chaque forme du texte à la forme-vedette dont elle

[1] Cfr BODSON Arthur, ÉVRARD Étienne, *Le programme d'analyse automatique du latin*, dans *RELO*, 1966, 2, pp. 17–46; DENOOZ Joseph, *Recherches sur le traitement automatique de la langue latine*, dans *RELO*, 1973, n° 1, pp. 1–89; *Le traitement des textes latins, grecs et français au Laboratoire d'Analyse Statistique des Langues Anciennes*, dans *Revista de la Universidad Complutense*, XXV, 102, mars-avril 1976, pp. 143–167.

dépend (le lemme), c'est-à-dire à la forme sous laquelle le mot apparaît dans un dictionnaire de référence. La précision de la seconde peut varier : soit un seul code représente la catégorie grammaticale d'une forme, soit plusieurs codes traduisent son analyse morphologique complète (sous-catégorie, déclinaison, conjugaison; cas, genre, nombre; voix, mode, temps, personne).

Enfin, depuis plusieurs années, à l'heure où les méthodes qui viennent d'être évoquées sont bien connues et bien rodées, on constate, dans les milieux scientifiques qui recourent à l'informatique, un souci nouveau de dépasser ce stade pour progresser dans plusieurs directions qui, toutes, ont pour point commun de déplacer le centre d'intérêt du mot (unité lexicale) vers divers éléments, à la fois plus larges et plus difficilement maîtrisables, considérés comme nouveaux objets d'étude des procédures automatiques. Pour ce qui relève de la stylistique et de l'analyse littéraire, c'est dans le sens de l'étude du contenu que s'orientent les nouvelles investigations. Depuis longtemps déjà, l'étude des listes de fréquences du vocabulaire d'une œuvre permet de rendre compte de sa thématique. Une autre méthode consiste en la recherche dans un texte des cooccurrences qu'il contient. Une cooccurrence est comprise comme la présence conjointe de deux termes au moins dans un même contexte (même unité de référence — chapitre, paragraphe ou vers —, ou même phrase). Par ailleurs, l'étude des cooccurrences répétées conduit à distinguer, dans le vocabulaire du texte, les termes qui, dans l'écriture de l'auteur, s'appellent réciproquement de manière privilégiée.

Enfin, une autre manière intéressante d'observer les rapports des unités lexicales d'un texte entre elles est la recherche et l'étude des groupes récurrents et des formules verbales qu'il contient. Dans un premier temps, sous réserve d'une définition plus nuancée, je considère que ces deux notions désignent des groupes verbaux syntaxiquement homogènes et répétés, dont les éléments constitutifs sont contigus ou peu éloignés les uns des autres dans le texte; on notera qu'il est nécessaire, pour qu'un groupe verbal entre dans cette catégorie, qu'il apparaisse plus d'une fois dans le texte ou corpus considéré[2]. Ce type de recherche revêt beaucoup d'intérêt pour l'étude des styles formulaires, des langues propres à certains genres littéraires, etc.

L'évolution dont l'historique vient d'être esquissé a donc fait passer l'intérêt de la forme au vocable (ou lemme, c'est-à-dire série de formes), puis s'est ouverte au groupe de mots.

Je propose ici une étude méthodologique de la recherche automatique, par des moyens informatiques, des groupes verbaux récurrents, dans les textes latins tels qu'ils sont lemmatisés, analysés et conservés sous forme de fichiers

[2] Pour une définition plus complète de ces notions, cfr *infra*, 2.1.

selon les méthodes mises au point par le Laboratoire d'Analyse Statistique des Langues Anciennes de l'Université de Liège. Une première partie définira l'utilité multiple que pourra présenter une telle recherche appliquée au latin. Seront ensuite analysés les problèmes et difficultés, tant théoriques que techniques, qui se posent au moment de concevoir et de choisir une méthode satisfaisante; cette partie présentera notamment l'importance que revêt la précision du codage des données textuelles tel qu'il a été mis au point par le LASLA. Enfin, après un exposé détaillé des diverses méthodes que j'ai envisagées et développées, mon travail s'achèvera par un examen philologique des résultats obtenus à partir de deux textes latins et des conclusions auxquelles ils peuvent mener. Les problèmes soulevés sont aussi ceux que pose la langue grecque, et les méthodes proposées s'adaptent facilement au traitement de cette langue. C'est par un souci de concision que j'ai limité cette présentation au latin.

1. Utilité de la recherche de groupes récurrents en latin.

Je situerai l'utilité de la recherche de groupes récurrents sur deux plans différents. En tant que purs documents mis à la disposition des philologues, les listes de contextes contenant ces récurrences me paraissent à même de leur rendre service dans deux directions au moins. La comparaison de la liste des formules de plusieurs auteurs peut conduire à dégager, parmi ces groupes récurrents, de véritables expressions idiomatiques, relevant du formulaire de la langue elle-même. Repérer par des moyens automatiques les différents types d'expressions de ce genre mais aussi leurs occurrences avec références contribue utilement à l'étude de la phraséologie. On aperçoit l'usage tant linguistique que pédagogique que l'on peut faire de tels résultats.

Par ailleurs, dans le même ordre d'idée, le repérage dans un même texte de tous les groupes verbaux récurrents, et notamment des formules, peut apporter une aide intéressante à son traducteur. Disposer de toutes les références d'une même expression lui permet, d'une part, de préciser son sens en comparant son emploi dans chaque contexte, et, d'autre part, de veiller à la traduire de manière homogène à chacune de ses occurrences.

Enfin l'étude même d'un auteur et de ses œuvres doit pouvoir se nourrir de telles recherches. La langue de cet auteur et l'usage qu'il en fait, c'est-à-dire son style, peuvent être envisagés non plus seulement en termes de vocables distincts, de fréquences de vocabulaire, de proportion des catégories grammaticales, mais aussi par le biais de données telles que le type de groupes récurrents, la proportion de formules, les rapports entretenus avec d'autres auteurs, etc. En étudiant le nombre de séquences récurrentes dans un corpus et leur fréquence, il sera possible d'identifier le style plus particulièrement formulaire de certaines œuvres et de certains genres littéraires. On songera, par exemple, à l'intérêt de

semblables relevés pour l'étude de la composition poétique[3]. En fait, l'usage qui peut être fait de ces listes de formules varie selon l'auteur envisagé et surtout selon le genre littéraire qu'il pratique. On peut attendre qu'elles contiennent, dans certains cas, les éléments d'une véritable langue de genre. Lorsqu'il porte sur la poésie latine, l'examen de ces groupes récurrents intéresse autant l'étude de la composition métrique que celle de la thématique de l'auteur ou du genre. Du côté des prosateurs, la même méthode, appliquée, par exemple, au *De Agricultura* de Caton, ouvre des perspectives relativement différentes. Cette application fera l'objet d'un prochain article en préparation.

On peut en fait considérer que, pour ces divers usages, le relevé des groupes et formules récurrents constitue un prolongement très utile de l'*index verborum*. Enrichir celui-ci par l'adjonction d'un index phraséologique ou formulaire permet de dépasser le niveau du simple vocable et peut, tout comme une concordance, diminuer l'effet d'éventuelles polysémies, invisibles dans un index.

De telles recherches ont déjà été appliquées à d'autres langues que le latin. Ainsi les chercheurs de l'Unité de Recherche Lexicologie et Textes Politiques de Saint-Cloud, qui dépend de l'Institut National de la Langue Française (INaLF), tels Pierre Lafon, André Salem et Maurice Tournier, ont depuis de nombreuses années consacré leurs travaux à l'étude de la cooccurrence dans les textes politiques et syndicaux de langue française. Un des grands domaines de recherche explorés par le Laboratoire de Saint-Cloud porte sur les "Inventaires des Segments Répétés", à savoir les listes de groupes récurrents[4].

La méthode du Laboratoire de Saint-Cloud, qui ne recourt pas à la lemmatisation, repose sur une indexation des formes (graphiques); il s'agit de "relever toutes les suites d'unités graphiques répétées au moins une fois dans le texte." Dans une seconde étape, une sélection manuelle permet de distinguer les locutions grammaticales et les segments "sémantiques".

---

[3] Cfr PURNELLE Gérald, *Répétition de cooccurrences métriques chez Ovide*, dans *Revue. Informatique et Statistique dans les Sciences humaines*, 1987, pp. 135–166.

[4] Cfr LAFON Pierre, SALEM André, *L'inventaire des segments répétés d'un texte*, dans *Mots*, n° 6, mars 1983, pp. 161–177; LAFON Pierre, *Pour une nouvelle unité de segmentation des textes à partir de l'Inventaire des Segments Répétés. Application à la résolution générale du congrès de la CFDT en 1976*, dans *Méthodes quantitatives et informatiques dans l'étude des textes. En hommage à Charles Muller*, Colloque International CNRS, Université de Nice, 5–8 juin 1985, Paris, Slatkine, 1986, vol. II, pp. 531–540; SALEM André, *L'utilisation des Inventaires de Segments Répétés dans les études typologiques sur des corpus de textes, ibid.*, vol. II, pp. 759–769.

2. Analyse du problème et description des méthodes.

Pour illustrer à la fois les implications théoriques, les difficultés inhérentes à la langue traitée et à la nature des groupes visés, les choix opérés pour résoudre la question et l'application des méthodes envisagées, il a paru nécessaire de prendre pour base un texte où il soit possible de puiser des exemples de groupes récurrents et qui puisse facilement être soumis aux procédures développées; il devait être assez riche en groupes récurrents, mais aussi suffisamment court pour produire des résultats pouvant tenir dans les limites du volume de ce travail. J'ai choisi, comme échantillon, un ensemble de 14 chapitres extraits du *De Agricultura* de Caton, à savoir les chapitres 104 et 105, 108 à 115 et 122 à 125 (soit un total de 943 mots); ils sont thématiquement apparentés (il s'agit de recettes concernant le vin) et sont particulièrement abondants en groupes récurrents et formules (une simple lecture permet de s'en rendre compte). D'autre part, afin d'accroître mon stock d'exemples, j'ai aussi appliqué une des méthodes de recherche de formules à un discours de Cicéron, le *Pro C. Rabirio perduellionis reo ad Quirites oratio*.

J'ai, pour cette présentation, préféré laisser de côté la recherche de formules en poésie. Il m'a semblé que pour un exposé méthodologique, il était préférable de ne pas multiplier les sources de variations et de différences. Or, dans la perspective qui nous occupe, les conditions sont différentes pour la prose et pour la poésie; dans le cas de celle-ci, elles sont à la fois plus faciles et plus difficiles : le rôle que jouent la métrique et la technique de composition dans l'apparition et la position des formules en poésie[5] ajoutent à la question un aspect n'entrant pas dans la réflexion de portée générale qui convient à la prose. Quoi qu'il en soit, appliquer à la poésie les méthodes qui suivent ne serait qu'une affaire d'adaptation, de développement, de précision; d'autre part, la recherche de formules en poésie prend généralement (ou du moins peut prendre) pour unité de référence le vers (et non la phrase comme en prose, cfr *infra*, 2.2.2.), ce qui réduit considérablement un certain nombre de difficultés qui vont apparaître.

2.1. Définitions.

A un premier niveau, j'appelle *groupe récurrent* tout groupe de formes consécutives apparaissant plus d'une fois dans un texte, qu'existent ou non des liens syntaxiques ou sémantiques entre ces formes; à un second niveau, une fois identifiés et négligés les groupes de ce genre dont la rencontre est purement fortuite et n'a rien de linguistique ou formulaire, le terme *groupe récurrent* correspond au *Segment Répété* du Laboratoire de Saint-Cloud (cfr *supra*).

---

[5] Cfr PURNELLE G., *Répétition de cooccurrences métriques chez Ovide.*

Il importe de déterminer quel nombre minimal doivent atteindre les éléments d'un groupe pour que sa récurrence soit intéressante. Le nombre le plus représenté est aussi le plus bas : dans tout texte, les groupes de deux mots sont de loin majoritaires. Ceux-ci, en raison de leur brièveté, de leur fréquence et de leur faible richesse sémantique, ne présentent souvent qu'un intérêt minime ou nul, comme on peut le voir dans la partie *Analyse des résultats* (*infra*, 3.); on songera, par exemple, aux paires formées d'une préposition et d'un nom, d'un subordonnant et d'un verbe. Souvent, c'est à partir du niveau de trois mots qu'apparaissent vraiment les groupes dignes d'intérêt. On sera donc tenté de négliger les groupes binaires et de commencer toute recherche à partir du niveau trois. Ce choix reste l'entière liberté du chercheur : prendre ces groupes en compte produit à coup sûr un "bruit" important; ne pas le faire, c'est risquer d'ignorer des informations ne manquant pas d'intérêt. Pour ma part, dans les applications de ce travail, j'ai choisi de les prendre en compte, dans le but d'illustrer au mieux la méthode.

Une fois défini le groupe récurrent, on peut, à un niveau supérieur, tenter de distinguer de ce type de groupes ce qui peut plus particulièrement être appelé *formule*. On notera d'abord que la notion de formule relève, au premier chef, du domaine de la poésie, ainsi que l'indique la célèbre définition de Milmann Parry : "[une formule est] un groupe de mots qui est régulièrement employé dans les mêmes conditions métriques pour exprimer une idée essentielle donnée"[6].

En définitive, la première de ces deux notions recouvre tout groupe verbal qui présente une unité syntaxique et/ou sémantique et se trouve employé plus d'une fois dans le même corpus; à la différence de la cooccurrence répétée, le *groupe verbal récurrent*, pour être qualifié comme tel, doit présenter une certaine homogénéité : il ne s'agit plus simplement de deux termes présents dans la même phrase, quelle que soit la distance qui les sépare, mais d'une expression dont les termes, d'une part, présentent entre eux une certain lien, et, d'autre part, sont contigus dans l'ordre du texte ou, du moins, peu éloignés les uns des autres. Quant à la *formule*, il s'agit d'un groupe récurrent dont la définition est plus délicate (et plus subjective) et tient moins à la nature de ses constituants qu'à l'usage qui en est fait. Alors que, lors de l'identification d'un groupe récurrent, quasiment aucune considération d'ordre littéraire, stylistique ou sémantique n'intervient, il faut, pour pouvoir dénommer "formule" un tel groupe, que certaines conditions soient satisfaites dans l'un de ces sens. En fait, dans l'acception où je le prends dans les lignes qui suivent, le terme désigne un groupe plus homogène encore qui constitue une véritable expression figée,

[6] "... a group of words which is regularly employed under the same metrical conditions to express a given essential idea." (PARRY M., *Studies in the Epic Technique of the Oral Versemaking*, dans *Harvard Studies in Classical Philology*, 1930, 41, pp. 73–147.

sinon dans toute la langue, du moins dans celle de l'auteur ou du genre qui y recourt et fonctionnant comme une seule entité sémantique.

2.2. Contraintes d'ordre linguistique et formel.

2.2.1. La variation morphologique.

Le latin est une langue flexionnelle; la plupart des mots (substantifs, verbes, adjectifs, pronoms) varient formellement selon la fonction exercée dans la phrase, qui est marquée par des suffixes différents. Cela signifie que, graphiquement, un même mot revêt généralement dans un texte de multiples formes, parfois très dissemblables (*sum*, *erat*, *fui*; *bonus*, *melior*, *optimus*). Entreprendre une recherche de cooccurrences ou de groupes récurrents sur la seule base des unités graphiques expose donc à ne repérer qu'une faible partie de ces phénomèmes. Ainsi, pour prendre un exemple dans l'échantillon tiré de Caton, les mots AQVA et VINVM sont cooccurrents dans les contextes suivants :

Si uoles scire in uinum aqua addita sit necne, ... (111)
Vinum Coum si uoles facere, aquam ex alto marinam sumito ... (112,1)
De eo uino cyatum sumito et misceto aqua ... (114,2)

mais présentent chacun deux formes différentes. Si chacune de ces formes n'a pas été préalablement rapportée à son lemme, il est impossible de repérer automatiquement la cooccurrence répétée.

De la même manière, les groupes récurrents sont eux aussi constitués d'éléments qui peuvent, d'une occurrence à l'autre, varier morphologiquement et graphiquement. Deux exemples tirés de Caton :

aceti acris (104,1) / *acetum acerrimum* (104,2)
odorem deteriorem demere / *demptus erit odor deterior* (110,1)

Quelle que soit la méthode utilisée pour les repérer, de tels groupes, qui pourtant sont bien homogènes, échapperaient à toute investigation fondée sur l'identification des seules formes graphiques.

Pour une langue flexionnelle, une recherche efficace des groupes récurrents est donc quasiment impossible à réaliser sans lemmatisation préalable du texte étudié; elle a en tout cas toutes les chances d'être grandement incomplète et donc assez peu utile. On constate cependant que nombre de chercheurs se contentent d'une situation aussi peu confortable et renoncent à consacrer du temps à la lemmatisation de leurs corpus. C'est notamment le cas du Laboratoire de Saint-Cloud, où la recherche des cooccurrences, comme l'établissement des Inventaires de Segments Répétés, ne recourent à aucune procédure de ce genre et restent tributaires des variations morphologiques et graphiques des

termes étudiés[7]. Ainsi, pour ces chercheurs, un groupe de mots au singulier forme un segment différent de ce même groupe au pluriel.

Fort heureusement, pour le latin, la situation est plus favorable, puisqu'on dispose depuis vingt-cinq ans de procédures semi-automatiques de lemmatisation et d'analyse morphologique, développées par le LASLA de Liège. Il est donc possible, pour le latin, de respecter cette première contrainte.

2.2.2. La notion de phrase; la ponctuation.

Il est bien connu que la ponctuation qui figure dans les éditions des œuvres latines n'est pas originale, mais est le fait des éditeurs modernes, qui identifient dans le texte des séquences logiques et ponctuent selon leur sentiment; elle peut donc varier d'une édition à l'autre. Dans le système actuel du LASLA, seules sont prises en considération les ponctuations fortes, c'est-à-dire les points (y compris points d'interrogation et d'exclamation); les virgules et les points-virgules sont donc négligés. Ceci signifie que si l'on veut amener l'ordinateur à délimiter chaque phrase du texte, il ne peut prendre appui que sur les ponctuations fortes et identifiera comme parties d'une même phrase des segments qui, dans l'édition, sont séparés par des points-virgules. Cet état de fait présente notamment un inconvénient pour notre recherche. En effet, dès lors que deux mots se suivent plus d'une fois dans le texte, ils seront chaque fois considérés comme groupe récurrent même si, dans un cas au moins, une virgule ou un point-virgule les sépare, interdisant de poser leur homogénéité. On tiendra donc compte de cet inconvénient et des imperfections (information incorrecte et excédentaire) qu'il entraînera.

Cependant, on l'a vu, c'est bien la phrase qui doit servir de base de référence aux recherches envisagées, étant donné que, généralement, elle est plus courte que le paragraphe. Les enregistrements des fichiers du LASLA ne comportant pas l'indication du numéro de phrase, il a fallu recoder chaque fichier pour faire figurer cette information en regard de chaque forme.

2.2.3. L'ordre des mots.

L'ordre des mots en latin pose lui aussi problème. Les éléments constitutifs d'un segment homogène peuvent en effet, d'une occurrence à l'autre, présenter un ordre différent. A ce point de vue, le latin offre une grande différence avec les langues européennes modernes, où l'ordre des mots varie beaucoup moins; en anglais ou en allemand, par exemple, la place de l'épithète est fixe; en français, elle est certes variable (devant ou derrière le nom) et régie par

[7] Cfr P. LAFON, p. 534.

de multiples règles, mais elle n'est pas assez lâche pour écarter l'adjectif de son régissant. Cette différence tient essentiellement au fait que le latin, étant langue flexionnelle, marque la fonction des mots dans la phrase au moyen de désinences casuelles, tandis que, dans les langues modernes, c'est la place respective des syntagmes qui détermine leur fonction; même l'allemand, qui possède un système casuel, présente un ordre des mots moins souple.

Cette spécificité du latin a une première conséquence: un groupe récurrent, quelle que soit son étendue, voit l'ordre de ses constituants varier beaucoup plus en latin qu'en français. Soit le segment *pouvoir d'achat*, repéré dans les textes syndicaux par la Laboratoire de Saint-Cloud[8]; si nombreuses que soient ses occurrences, il n'y a aucune chance que l'ordre de ses trois éléments soit modifié. Par contre, dans les deux textes latins qui ont été étudiés (l'échantillon du *De Agricultura* de Caton et le *Pro Rabirio* de Cicéron) on trouve plusieurs cas de permutation. Exemples:

in pila contundito / contundito in pila (CAT.)
facere uoles / uoles facere (CAT.)
corpus omnium ciuium Romanorum / omnium ciuium Romanorum corpus (CIC.)
pro salute communi / pro communi salute (CIC.)
sceleris ac parricidi nefarii mortuum condemnabimus / nefarii sceleris ac parricidi mortuum condemnabimus (CIC.)

L'implication de cette différence est manifeste: il est plus difficile de repérer toutes les occurrences d'un groupe en latin qu'en français. Il ne suffit pas d'identifier la succession immédiate des mêmes termes et dans un même ordre. On verra qu'elle oblige à renoncer aux méthodes les plus simples et à développer des procédures plus longues.

La nature particulière de l'ordre des mots en latin comporte, dans notre perspective, un second inconvénient : dans la définition d'un groupe syntaxiquement homogène, la contiguïté de ses éléments n'est en aucune manière un critère absolu; deux ou plusieurs termes contigus peuvent n'être rapprochés que fortuitement (et en tout cas ne pas appartenir à la même formule ou groupe syntaxique); inversement, les éléments d'une même formule peuvent être dissociés, soit par la division du groupe et l'insertion en son sein d'un terme étranger au groupe, soit par insertion d'un nouvel élément dépendant de la formule.

En conséquence, dans l'état actuel du système d'analyse et de codage utilisé par le LASLA on ne peut trouver tous les groupes sémantiquement et syntaxiquement homogènes, dès lors que certains ont toutes les chances d'être

[8] Cfr P. LAFON, A. SALEM, p. 165.

éclatés dans la phrase. En effet, à l'heure actuelle, en partant de manière purement automatique des fichiers tels qu'ils existent, la méthode ne peut se fonder que sur le concept de contiguïté ou de proximité. Il est donc impossible de dépasser ce principe tant que l'on ne dispose pas, pour chaque forme, d'une analyse syntaxique traduisant exactement sa fonction dans la phrase (ou, du moins, la forme dont elle dépend). De tels développements du système sont à l'étude (cfr *infra*, 4.2); lorsqu'on disposera de textes latins enrichis de cette manière, on pourra envisager la mise au point de nouvelles procédures de recherche automatique de groupes récurrents syntaxiquement homogènes, quelle que soit la distance qui sépare leurs éléments.

2.3. Le système de codage du LASLA.

Les procédures auxquelles sont soumis les textes latins présentent deux aspects complémentaires : la lemmatisation et l'analyse morphologique. Elles se développent selon plusieurs étapes, qu'il n'est pas nécessaire de passer en revue ici[9]. Je me contenterai d'en rappeler le résultat final. Lorsqu'un texte latin a été lemmatisé et analysé, il est conservé dans un fichier, dont chaque enregistrement contient, sur 80 positions, une forme du texte, son lemme, sa référence complète (code d'œuvre; livre; chapitre; paragraphe, vers ou ligne), son analyse grammaticale, son numéro d'ordre dans le texte et divers autres codes.

A l'intérieur d'un fichier lemmatisé et analysé, les formes sont généralement disposées dans l'ordre du texte, mais on peut facilement trier ce fichier en prenant pour critère de tri le lemme de chaque enregistrement. On obtient alors un index alphabétique de tous les mots du texte, les formes de chaque lemme pouvant être rangées, selon les besoins, soit dans un ordre grammatical (l'ordre des déclinaisons ou des conjugaisons dans le cas du latin), soit, plus simplement, dans l'ordre croissant de leur référence (c'est-à-dire dans l'ordre de leur apparition dans le texte). Dans l'exposé des méthodes qui suit, j'utiliserai tantôt l'ordre du texte, tantôt l'ordre alphabétique des lemmes, avec ordre des références pour les formes; j'appellerai ce second état *ordre d'index*.

Dans l'un et l'autre cas, il s'est avéré nécessaire de préparer les fichiers qui doivent être soumis aux procédures de recherche, et ceci dans deux directions; il faut : 1. numéroter les phrases et reporter dans l'enregistrement de chaque forme le numéro de la phrase où elle apparaît; 2. renuméroter les formes; les raisons de cette seconde modification seront exposées plus loin (cfr 2.4.1.). Ces deux opérations sont accomplies par un même programme.

---

[9] Pour une description détaillée de la procédure et du codage, cfr DENOOZ Joseph, *Techniques et méthodes*, dans *RELO*, 1978, 1, pp. 1–36.

2.4. La recherche de groupes récurrents.

2.4.1. Le choix des formes à rechercher.

Dans la recherche de groupes récurrents, un bon principe est de ne pas chercher à limiter *a priori*, dans une quelconque direction, la nature et la qualité des résultats produits par l'ordinateur et de le laisser identifier toutes les occurrences qu'il trouve, quitte à éliminer *a posteriori* l'information non pertinente, non souhaitée ou inutile, après examen. De cette manière, on évite de passer à côté de cas intéressants et on laisse le champ ouvert à toute découverte inattendue. Cependant, laisser totalement libre la recherche de toutes les séquences d'au moins deux formes contiguës a pour effet de produire un "bruit" important, c'est-à-dire une grand nombre de résultats sans aucun intérêt, dont l'apparition est purement fortuite et dont les éléments constitutifs ne participent pas au sens. Il est dès lors nécessaire, en fin de traitement, d'épurer les résultats en éliminant manuellement les groupes de ce type. Quelques exemples, tirés de l'échantillon :

id est ad aluum crudam et ad *lateris dolorem* et ad *cœliacum* (125,1)

hoc uinum deterius non erit quam *Coum* (105,2) / *nolito implere quadrantalibus quinque minus* sit quam *plenum* (112,1)

uinum ad *aluum mouendam concinnare uites* (115,2) / uinum ad *isciacos sic facito de iunipiro* (123,1)

diem postea commisceto cum eo uino in *dolio et oblinito die LX* (109,1) / *si uoles scire* in uinum *aqua addita sit nec ne uasculum facito* (111,1)

uoles de lacu quam primum uinum in *dolia indito sinito dies XV* (113,1) / in uinum *mustum ueratri atri manipulum coicito* (115,1)

Il y a cependant moyen, sans contrevenir au principe d'ouverture énoncé plus haut, d'éviter en cours de traitement et de manière automatique la production de bon nombre de ces résultats ne présentant manifestement aucun intérêt (c'est-à-dire n'ayant aucune chance d'être formulaires), et ce malgré l'absence d'analyse syntaxique et de codage de dépendance.

Il est exclu d'éliminer systématiquement et massivement les mots grammaticaux et le verbe "être" dans la recherche des formules et groupes homogènes. La raison en est qu'ils participent souvent de ces groupes, même s'ils n'en constituent pas la substance significative. Dans une formule de trois ou quatre mots, peut très bien entrer un subordonnant ou une préposition.

Exemples :

in uase aheneo uel in plumbeo deferuefacito ubi refrixerit in lagonam (122,1) / *inferuefacito cum congio uini ueteris* ubi refrixerit in lagonam (123,1)
si uoles facere (109,1; 112,1)

Exemple pour le verbe "être" :

id uinum erit lene et suaue *et bono colore et bene odoratum* (109,1) / *uinium asperum quod* erit lene et suaue *si uoles facere sic facito* (109,1)

Renoncer à prendre en compte un de ces mots grammaticaux ou le verbe "être" revient à s'interdire de repérer les groupes pertinents auxquels ils appartiennent.

Seuls, semblent pouvoir être éliminés au départ :

1. les conjonctions de coordination, car même si elles entrent dans un groupe récurrent, les éléments qu'elles coordonnent suffisent à identifier le groupe (cfr l'exemple ci-dessus). Par ailleurs, si on maintient ces conjonctions, on risque de voir apparaître dans les listes de résultats de prétendus groupes récurrents constitués d'un mot simplement précédé ou suivi d'une conjonction; tout terme apparaissant plus d'une fois en coordination, avec quelque autre terme que ce soit, sera donc inutilement repéré; exemple :

erit in tecto in cratibus componito et *si qua acina corrupta erunt depurgato* (112,2) / *post quadriennium in cuneum* componito et *instipato* (113,2)

2. les emplois du verbe *esse* comme auxiliaire dans les formes périphrastiques du passif latin. Les conventions de codage du LASLA imposent de ménager deux enregistrements pour toute forme périphrastique, l'un pour le participe et l'autre pour la forme de l'auxiliaire (lemmatisée SVM avec l'indice 2); il est donc non seulement utile mais nécessaire de supprimer cette forme dans la recherche de groupes, sous peine de duplication de l'information; par ailleurs, si l'on maintenait séparées les deux parties de la périphrase, on s'exposerait à identifier systématiquement celle-ci comme groupe récurrent, ce qui serait abusif, la périphrase constituant sémantiquement un seul signe.

Mais, une des bases des méthodes développées étant le numéro d'ordre de chaque mot dans le texte, si on néglige certains mots de manière à faire en sorte qu'ils n'existent pas *pour la recherche des groupes récurrents*, le principe de la méthode étant la contiguïté, on risque de ne pas repérer des cas tels que le groupe *erit lene et suaue* (cfr *supra*). Il est dès lors nécessaire de faire aussi abstraction de la place qu'occupent ces mots dans la numérotation continue des formes du texte et de "passer" au dessus d'eux, donc de renuméroter

pour les besoins de la méthode les formes dans leurs enregistrements respectifs (cfr *supra*, 2.3.).

Hormis ces deux cas, il est donc exclu d'éliminer *au départ* les mots grammaticaux et le verbe "être", même si leur prise en compte par les premières étapes des méthodes est fastidieuse. Cependant, afin d'alléger la masse des résultats finals produits, on peut envisager d'éliminer automatiquement certains des groupes repérés, dans lesquels ils rentrent, et qui, en raison de la nature de leurs éléments constitutifs, n'ont aucune chance d'être intéressants. Pour ce faire, il est permis d'établir un certain nombre de règles d'élimination, basées sur des critères purement morphologiques, en l'absence de codage syntaxique. Voici la liste des règles que j'ai appliquées; elle ne couvre sans doute pas toutes les possibilités, mais, en l'occurrence, il convient de rester prudent et de se souvenir qu'il vaut mieux obtenir un résultat inutile en plus plutôt que de perdre un information intéressante en raison d'une règle trop sévère.

On élimine :

- tout groupe ne contenant ni substantif, ni verbe, ni adjectif, ni adverbe non grammatical; exemples dans Caton :

  id est ad aluum crudam et ad *lateris dolorem* et ad *cœliacum* (125,1)

  hoc uinum durabit tibi usque ad *solstitium* (104,2) / *ita relinquito* usque ad *uindemiam* (112,2)

- un groupe de deux mots dont l'un est une préposition, si l'autre n'est pas un substantif; exemples :

  id uinum seruato ad *aluum mouendam* (115,1) / *umbra ubi iam passa erit* seruato ad *uindemiam in urnam musti contundito* (125,1)

  de eo *uino cyatum sumito et misceto* (114,2) / *per uindemiam* de iis *uitibus quod delegeris seorsum seruato* (115,2)

- un groupe de deux mots si l'un est SVM ou NON; exemples :

  et calamum in pila contundito quod siet *sextarium unum eodem in dolium* (105,2) / *uinum asperum* quod erit *lene et suaue si uoles* (109,1)

  hoc uinum deterius non erit *quam Coum* (105,2) / *sumito mari tranquillo cum uentus* non erit *dies LXX ante uindemiam quo* (112,1) / *triduum sub dio si pluuiae* non erunt *si pluuia erit in tecto* (112,2) / *in sole ponito ubi herba* non siet *et amphoras operito ne aqua* (113,2)

- un groupe contenant un subordonnant (conjonction de subordination, pronom relatif ou interrogatif) mais pas de verbe; exemple :

indito et uini sextarium de eo uino quod *uoles experiri eodem infundito* (108,1) / *uasculum facito de materia hederacia* uinum id quod *putabis aquam habere eo demittito* (111,1)

- un groupe de deux mots dont l'un est un pronom (sauf si ce pronom est un subordonnant et l'autre mot un verbe); exemple :

hoc uinum *durabit tibi usque ad solstitium* (104,2) / hoc uinum *deterius non erit quam Coum* (105,2) / hoc uinum *seorsum legito si uoles seruare* (114,2)

On peut toujours choisir d'aller plus loin dans la sélection et d'éliminer des associations échappant aux critères d'exclusion qui viennent d'être définis, afin d'alléger les listes de résultats. On peut, par exemple, décider de négliger les groupes de deux mots constitués d'une préposition et d'un substantif; ce genre de construction est en effet à ce point banal que, dans la plupart des cas, il n'a que peu de chance d'accéder au statut de formule. La prudence s'impose cependant. Ainsi, si l'exemple suivant est effectivement peu intéressant,

si uoles scire in uinum *aqua addita sit necne* (111,1) / in uinum *mustum ueratri atri manipulum coicito* (115,1)

d'autres le sont davantage :

percolato polentam abicito uinum ponito sub dio (108,2) / *ponito in sole biduum aut triduum* sub dio (112,2)

On verra dans la section suivante à quel moment, selon la méthode choisie, intervient l'application des critères énoncés ci-dessus dans la sélection des groupes pertinents.

2.4.2. Description des différentes méthodes.

Précisons le but à atteindre : il s'agit de choisir une méthode permettant de produire de manière entièrement automatique les groupes contigus récurrents d'un texte; elle doit les repérer tous, quelle que soit leur longueur et tenir compte de la possible variation de l'ordre des éléments à l'intérieur de ces groupes; il faut y appliquer les critères de sélection des groupes qui n'ont aucune chance d'être pertinents; enfin, on doit pouvoir identifier les groupes participant à d'autres plus larges et les mettre en relation avec ceux-ci.

A. *Le choix de la méthode*

La première méthode qui se puisse envisager est fréquemment utilisée, notamment en français; c'est à elle que recourent les chercheurs du Laboratoire de Saint-Cloud :

> "Pour une forme donnée, et si le texte que l'on étudie n'est pas trop long, on peut repérer sans trop de mal les séquences répétées dans lesquelles elle fonctionne, à l'aide des outils traditionnels du lexicométricien, que sont la concordance et l'index alphabétique. En effet, en se reportant à l'entrée correspondante d'une concordance munie d'un contexte suffisamment étendu et dont les lignes sont triées sur la partie droite du contexte par ordre alphabétique, on peut dresser la liste des expressions figées qui contiennent cette forme." (P. LAFON, A. SALEM, p. 162)

Cette méthode, facile à programmer, est sans doute suffisante pour le français, pour les raisons qui ont été indiquées plus haut : l'ordre des mots y est assez régulier pour permettre de trouver de cette façon tous les groupes récurrents. Cependant, des raisons du même ordre interdisent de l'appliquer au latin; on a vu en effet que dans plusieurs groupes récurrents les termes peuvent apparaître dans un ordre variant à chaque occurrence. (cfr *supra*, 2.2.3.).

Le méthode à utiliser doit donc être conçue de manière à traiter tout groupe de mots sans tenir compte de l'ordre de ses éléments. Le seul moyen d'y parvenir est de baser la recherche sur des groupes de mots classés en ordre alphabétique. Ce principe étant posé, on peut imaginer deux voies opposées pour entreprendre le repérage des groupes récurrents; on peut prendre pour base le fichier en ordre du texte ou son index. Comme on va le voir, chacune de ces deux méthodes, radicalement différentes, présente ses avantages et ses inconvénients, à plusieurs points de vue. Elles ont pour point commun d'être fondées sur la production combinatoire de paires d'éléments que l'on compare ensuite pour y découvrir les récurrences.

B. *La recherche en ordre du texte*

Dans ses grandes lignes, cette méthode est elle aussi assez habituelle; elle est notamment utilisée dans la recherche des formules métriques[10]. Elle revient à lire le texte phrase par phrase et à produire des séquences d'un nombre égal de mots, constituées progressivement dans le texte. Soit une phrase constituée de huit mots A B C D E F G H; la procédure produit sept paires de mots AB BC CD DE EF FG GH, six triplets ABC BCD CDE DEF FGH, cinq quadruplets, etc. On voit que le nombre de séries produites à chaque niveau est facilement calculable : il est égal au nombre de mots de la phrase plus un, moins le nombre d'éléments de la séquence. Une fois que sont constituées toutes les séries d'un même niveau, il

[10] Cfr VIKIS-FREIBERGS Vaira, FREIBERGS Imants, *Formulaic Analysis of the Computer-Accessible Corpus of Latvian Sun Songs*, dans *Computers and the Humanities*, 12 (1978), p. 333.

suffit de les trier et de les comparer pour identifier celles qui sont récurrentes; encore faut-il que les éléments qui entrent dans une série présente plusieurs fois soient à chaque fois dans le même ordre; il est donc nécessaire, lors de la lecture du fichier et de la création des séries, d'opérer sur leurs éléments un tri alphabétique. Pour une phrase D B F A E C, par exemple, on créera les paires BD, BF, AF, AE, CE.

Cette méthode permet donc en théorie de repérer tous les groupes récurrents de toute longueur, par niveau de longueur. Elle nécessite de soumettre le texte à la procédure pour chaque niveau de longueur. On ne peut en effet envisager de repérer les paires récurrentes pour tenter ensuite d'en inférer (grâce aux numéros d'ordre par exemple) les groupes de longueur supérieure; ce n'est théoriquement pas possible pour toutes les occurrences de toutes les formules, parce qu'une série récurrente de quatre éléments peut, en raison de la variation de l'ordre des mots, présenter des paires incompatibles : la récurrence ABCD/BDAC est impossible à repérer en partant des segments de deux mots (AB BC CD / BD AD AC). La seule solution est donc de produire toutes les séquences possibles à chaque niveau, puis de les comparer.

Il demeure cependant deux difficultés. La première réside dans le niveau maximal à atteindre. Au moment de traiter un texte, on ignore quel sera le nombre d'éléments des groupes récurrents les plus longs. Il est indubitablement préférable, à ce point de vue, de commencer l'application de la procédure par les groupes de deux mots, puis de poursuivre vers le haut jusqu'au moment où le niveau traité ne révèle plus aucun groupe récurrent, plutôt que de tâtonner en commençant par le haut pour redescendre vers le niveau le plus bas.

La seconde difficulté a plus d'importance. Dès lors qu'existe un groupe récurrent de trois mots, si, dans deux de ses occurrences au moins, ses éléments se présentent dans le même ordre (ABC), il est inévitable que lors de la recherche des groupes de deux mots (au niveau précédent) ces éléments aient été identifiés comme constituant deux groupes différents de deux mots: si, dans deux phrases, on trouve le groupe ABC, les groupes AB et BC auront été préalablement repérés et conservés dans la liste des groupes binaires; ceci peut se produire à tous les niveaux : un groupe de quatre mots apparaissant dans le même ordre produit deux groupes de trois mots et trois groupes de deux. On voit la difficulté que ce phénomène entraîne pour la recherche : la liste des groupes du niveau inférieur est inutilement gonflée par des informations parasites. Il est donc nécessaire de supprimer cet excédent en soustrayant de la liste, à chaque niveau, les séries incluses dans un des niveaux supérieurs. Une solution, pour y parvenir, serait, après avoir constitué la liste des groupes de trois mots récurrents, de produire à partir de cette liste la série des paires récurrentes qu'ils contiennent et de la soustraire des résultats rassemblés au niveau inférieur; cette solution, qui

devrait être appliquée à chaque niveau, se révélerait assez fastidieuse : elle comporterait en effet quatre opérations par niveau. Comme on le verra dans la description des procédures, il est possible de faire l'économie de cette opération de production en sens inverse, en organisant l'information de telle manière que l'on puisse localiser facilement l'appartenance d'un groupe de deux mots à un groupe de trois (en se basant sur le numéro d'ordre des mots impliqués).

La recherche des séries récurrentes à chaque niveau nécessite quatre opérations différentes; pour la soustraction des séries redondantes, une seule suffit par niveau. Lorsque le niveau maximal est atteint, on constitue une concordance des groupes récurrents repérés et l'on procède à la mise en relation des groupes apparentés.

a. On commence par la recherche des paires. Le premier programme lit le fichier contenant le texte (extension .TXT) et crée le fichier .201 (c'est-à-dire 1er fichier du niveau 2); après avoir isolé les formes d'une phrase, il apparie le lemme de chacune à celui de sa suivante (sauf le dernier) et, pour autant que la série ainsi constituée satisfasse aux critères de sélection des groupes pertinents énoncés plus haut (2.4.1.), il la conserve dans un enregistrement du nouveau fichier, en plaçant les lemmes qui la constituent en ordre alphabétique. L'enregistrement contient une autre information : la liste croissante des numéros d'ordre (dans le texte) des formes associées. Pour exemple, de la phrase n° 7,

> *Ubi muria facta erit, eodem in dolium infundito; schœnum et calamum in pila contundito, quod siet sextarium unum; eodem in dolium infundito, ut odoratum siet.* (105,2)

le programme produit les paires suivantes :

```
FACIO MVRIA /00114 00115
EODEM FACIO /00115 00116
DOLIVM IN /00117 00118
DOLIVM INFVNDO~2 /00118 00119
INFVNDO~2 SCHOENVS /00119 00120
CALAMVS SCHOENVS /00120 00121
CALAMVS IN /00121 00122
IN PILA~2 /00122 00123
CONTVNDO PILA~2 /00123 00124
CONTVNDO QVI~1 /00124 00125
SEXTARIVS VNVS /00127 00128
EODEM VNVS /00128 00129
DOLIVM IN /00130 00131
DOLIVM INFVNDO~2 /00131 00132
INFVNDO~2 VT~4 /00132 00133
```

Les séries suivantes ont été supprimées par le filtre du test de pertinence : MVRIA VBI, QVI~1 SVM~1, ODORATVS VT~4, ODORATVS SVM~1.

b. Le fichier .201 est trié dans l'ordre alphabétique des séries binaires (fichier .202); deux extraits :

```
DO RADIX /00657 00658
DOLIVM IMPLEO /00497 00498
DOLIVM IMPONO /00045 00046
DOLIVM IN /00010 00011
DOLIVM IN /00044 00045
DOLIVM IN /00095 00096
DOLIVM IN /00117 00118
DOLIVM IN /00130 00131
DOLIVM IN /00210 00211
DOLIVM IN /00255 00256
DOLIVM IN /00290 00291
DOLIVM IN /00375 00376
```

```
DOLIVM IN /00467 00468
DOLIVM IN /00516 00517
DOLIVM IN /00542 00543
DOLIVM IN /00563 00564
DOLIVM INDO /00011 00012
DOLIVM INDO /00564 00565
DOLIVM INFIMVM /00291 00292
DOLIVM INFVNDO~2 /00118 00119
DOLIVM INFVNDO~2 /00131 00132
DOLIVM INFVNDO~2 /00376 00377
DOLIVM ITEM /00411 00412
```

```
IN LAGENA /00799 00800
IN LAGENA /00828 00829
IN MARE /00374 00375
IN OPERCVLVM /00043 00044
IN PILA~2 /00122 00123
IN PILA~2 /00654 00655
IN PILA~2 /00782 00783
IN QVADRIENNIVM /00613 00614
IN QVADRIENNIVM /00618 00619
```

```
IN SCOPIO /00482 00483
IN SOL /00147 00148
IN SOL /00437 00438
IN SOL /00599 00600
IN SOL /00614 00615
IN TECTVM /00152 00153
IN TECTVM /00450 00451
IN TECTVM /00451 00452
```

c. Le deuxième programme sélectionne les séries récurrentes de niveau 2 (marquées ci-dessus) et les conserve (fichier .203) dans des enregistrements d'un nouveau format contenant : les numéros d'ordre de l'occurrence; un code (2) indiquant le niveau de combinaison; un numéro (croissant) attribué à chaque groupe récurrent; le groupe lui-même. En opérant cette sélection, ce programme doit effectuer un test touchant un cas particulier : il peut arriver qu'une forme, dans le texte, soit précédée et suivie du même mot; arrivé à cette forme, le premier programme produit donc deux paires identiques fondées sur trois occurrences seulement : soit le texte A B A aux numéros d'ordre 100, 101 et 102, les paires A–B/100–101 et A–B/101–102 sont créées; de même, au niveau 3, si le texte contient A B C A, on obtient les séries A–B–C/100–101–102 et A–B–C/101–102–103. Or, si, par hasard, cette association n'apparaît nulle part ailleurs dans le texte, elle risque d'être considérée à tort comme pertinente par le second programme; c'est pourquoi ce programme, pour déterminer qu'un ensemble de séries de même composition (mêmes lemmes) est pertinent, vérifie que les premiers numéros d'ordre d'au moins deux de ces paires sont à une distance suffisante l'un de l'autre (c'est le cas pour IN TECTVM ci-dessous); cette distance (différence) doit être égale ou supérieure à la valeur du niveau traité : pour le niveau 2, elle vaudra 2. Dans les deux exemples théoriques envisagés (A–B, A–B–C), les numéros d'ordre pris en compte sont 100 et 101; si, respectivement, A–B et A–B–C ne présentent pas d'autre occurrence, ces associations seront éliminées.

Ci-dessous un échantillon du fichier produit, correspondant aux exemples précédents :

```
00010 00011                2 27DOLIVM IN
00044 00045                2 27DOLIVM IN
00095 00096                2 27DOLIVM IN
00117 00118                2 27DOLIVM IN
00130 00131                2 27DOLIVM IN
00210 00211                2 27DOLIVM IN
00255 00256                2 27DOLIVM IN
00290 00291                2 27DOLIVM IN
00375 00376                2 27DOLIVM IN
00467 00468                2 27DOLIVM IN
00516 00517                2 27DOLIVM IN
00542 00543                2 27DOLIVM IN
00563 00564                2 27DOLIVM IN
00011 00012                2 28DOLIVM INDO
00564 00565                2 28DOLIVM INDO
00118 00119                2 29DOLIVM INFVNDO^2
00131 00132                2 29DOLIVM INFVNDO^2
00376 00377                2 29DOLIVM INFVNDO^2
00799 00800                2 39IN LAGENA
00828 00829                2 39IN LAGENA
00122 00123                2 40IN PILA^2
00654 00655                2 40IN PILA^2
00782 00783                2 40IN PILA^2
00613 00614                2 41IN QVADRIENNIVM
00618 00619                2 41IN QVADRIENNIVM
00147 00148                2 42IN SOL
00437 00438                2 42IN SOL
00599 00600                2 42IN SOL
00614 00615                2 42IN SOL
00152 00153                2 43IN TECTVM
00450 00451                2 43IN TECTVM
00451 00452                2 43IN TECTVM
```

d. Le fichier .203 est trié (.204) sur les numéros d'ordre (positions 1 à 5) et conservé jusqu'à la fin de la production des séries récurrentes du niveau supérieur. Un extrait :

```
00095 00096                2 27DOLIVM IN
00100 00101                2 44IN VAS^2
00102 00103                2 10AQVA DVLCIS
00103 00104                2 31DVLCIS QVADRANTAL
00117 00118                2 27DOLIVM IN
00118 00119                2 29DOLIVM INFVNDO^2
00122 00123                2 40IN PILA^2
00127 00128                2 56SEXTARIVS VNVS
00130 00131                2 27DOLIVM IN
```

```
00131 00132                    2 29DOLIVM INFVNDO~2
00136 00137                    2 24DIES POST~2
```

e. Les quatre opérations sont ensuite renouvelées pour le niveau 3, jusqu'à production des séries récurrentes dans le fichier .304. Échantillons des fichiers (.301 à 304) sur base de l'exemple initial :

```
FACIO MVRIA VBI~3 /00113 00114 00115
EODEM FACIO MVRIA /00114 00115 00116
DOLIVM EODEM IN /00116 00117 00118
DOLIVM IN INFVNDO~2 /00117 00118 00119
DOLIVM INFVNDO~2 SCHOENVS /00118 00119 00120
CALAMVS INFVNDO~2 SCHOENVS /00119 00120 00121
CALAMVS IN SCHOENVS /00120 00121 00122
CALAMVS IN PILA~2 /00121 00122 00123
CONTVNDO IN PILA~2 /00122 00123 00124
CONTVNDO PILA~2 QVI~1 /00123 00124 00125
CONTVNDO QVI~1 SVM~1 /00124 00125 00126
QVI~1 SEXTARIVS SVM~1 /00125 00126 00127
SEXTARIVS SVM~1 VNVS /00126 00127 00128
EODEM SEXTARIVS VNVS /00127 00128 00129
DOLIVM EODEM IN /00129 00130 00131
DOLIVM IN INFVNDO~2 /00130 00131 00132
DOLIVM INFVNDO~2 VT~4 /00131 00132 00133
INFVNDO~2 ODORATVS~2 VT~4 /00132 00133 00134
ODORATVS~2 SVM~1 VT~4 /00133 00134 00135
```

(Le test de pertinence n'a condamné aucune série.)

```
DOLIVM IMPONO IN /00044 00045 00046
DOLIVM IMPONO OBLINO /00045 00046 00047
DOLIVM IN INDO /00010 00011 00012
DOLIVM IN INDO /00563 00564 00565
DOLIVM IN INFIMVM /00290 00291 00292
DOLIVM IN INFVNDO~2 /00117 00118 00119
DOLIVM IN INFVNDO~2 /00130 00131 00132
DOLIVM IN INFVNDO~2 /00375 00376 00377
DOLIVM IN LAVO~2 /00516 00517 00518
DOLIVM IN MARE /00374 00375 00376
DOLIVM IN OBLINO /00255 00256 00257
DOLIVM IN OPERCVLVM /00043 00044 00045
DOLIVM IN OPERIO /00542 00543 00544
DOLIVM IN PONO /00541 00542 00543
DOLIVM IN QVADRAGENARIVS /00095 00096 00097
DOLIVM IN QVINQVAGENARIVS /00467 00468 00469
DOLIVM IN REFRIGESCO /00094 00095 00096
DOLIVM IN SCRIBO /00466 00467 00468
DOLIVM IN SVM~1 /00210 00211 00212
DOLIVM IN VINVM /00254 00255 00256
DOLIVM IN VINVM /00562 00563 00564
```

```
DOLIVM INDO SINO /00564 00565 00566
00116 00117 00118              3 10DOLIVM EODEM IN
00129 00130 00131              3 10DOLIVM EODEM IN
00010 00011 00012              3 11DOLIVM IN INDO
00563 00564 00565              3 11DOLIVM IN INDO
00117 00118 00119              3 12DOLIVM IN INFVNDO~2
00130 00131 00132              3 12DOLIVM IN INFVNDO~2
00375 00376 00377              3 12DOLIVM IN INFVNDO~2
00254 00255 00256              3 13DOLIVM IN VINVM
00562 00563 00564              3 13DOLIVM IN VINVM
00102 00103 00104              3  4AQVA DVLCIS QVADRANTAL
00116 00117 00118              3 10DOLIVM EODEM IN
00117 00118 00119              3 12DOLIVM IN INFVNDO~2
00122 00123 00124              3  7CONTVNDO IN PILA~2
00129 00130 00131              3 10DOLIVM EODEM IN
00130 00131 00132              3 12DOLIVM IN INFVNDO~2
```

f. Un fois que l'on dispose des listes de séries récurrentes de deux niveaux consécutifs (fichiers .204 et .304), un troisième programme procède au repérage des séries du niveau inférieur qui, incluses dans des séries du niveau supérieur, sont redondantes. Pour ce faire, il compare les deux fichiers ligne à ligne, en recherchant dans les séries de numéros d'ordre du second celles des enregistrements du premier. Tout enregistrement de niveau inférieur identifié de cette manière est éliminé (marqué d'un astérisque ci-dessous). Tous les autres enregistrements sont envoyés dans un nouveau fichier (.205), qui constitue la liste complète des groupes récurrents de ce niveau.

```
00103 00104              2 31DVLCIS QVADRANTAL *
00117 00118              2 27DOLIVM IN *
00118 00119              2 29DOLIVM INFVNDO~2 *
00122 00123              2 40IN PILA~2 *
00127 00128              2 56SEXTARIVS VNVS
00130 00131              2 27DOLIVM IN *
00131 00132              2 29DOLIVM INFVNDO~2 *
00136 00137              2 24DIES POST~2
```

g. Les cinq étapes décrites ci-dessus sont répétées pour chaque niveau (production des séries récurrentes et soustraction des paires redondantes au niveau inférieur) jusqu'au moment où on atteint le niveau supérieur à la longueur maximale des groupes récurrents du texte, la procédure ne livrant plus aucun résultat. Pour l'échantillon tiré de Caton, le niveau maximal est 4, avec trois groupes de cette longueur seulement (fichiers .401 et .404) :

```
EODEM FACIO MVRIA VBI~3 /00113 00114 00115 00116
EODEM FACIO IN MVRIA /00114 00115 00116 00117
DOLIVM EODEM FACIO IN /00115 00116 00117 00118
DOLIVM EODEM IN INFVNDO~2 /00116 00117 00118 00119
```

```
DOLIVM IN INFVNDO~2 SCHOENVS /00117 00118 00119 00120
CALAMVS DOLIVM INFVNDO~2 SCHOENVS /00118 00119 00120 00121
CALAMVS IN INFVNDO~2 SCHOENVS /00119 00120 00121 00122
CALAMVS IN PILA~2 SCHOENVS /00120 00121 00122 00123
CALAMVS CONTVNDO IN PILA~2 /00121 00122 00123 00124
CONTVNDO IN PILA~2 QVI~1 /00122 00123 00124 00125
CONTVNDO PILA~2 QVI~1 SVM~1 /00123 00124 00125 00126
CONTVNDO QVI~1 SEXTARIVS SVM~1 /00124 00125 00126 00127
QVI~1 SEXTARIVS SVM~1 VNVS /00125 00126 00127 00128
EODEM SEXTARIVS SVM~1 VNVS /00126 00127 00128 00129
EODEM IN SEXTARIVS VNVS /00127 00128 00129 00130
DOLIVM EODEM IN VNVS /00128 00129 00130 00131
DOLIVM EODEM IN INFVNDO~2 /00129 00130 00131 00132
DOLIVM IN INFVNDO~2 VT~4 /00130 00131 00132 00133
DOLIVM INFVNDO~2 ODORATVS~2 VT~4 /00131 00132 00133 00134
INFVNDO~2 ODORATVS~2 SVM~1 VT~4 /00132 00133 00134 00135

00116 00117 00118 00119        4  2DOLIVM EODEM IN INFVNDO~2
00129 00130 00131 00132        4  2DOLIVM EODEM IN INFVNDO~2
00683 00684 00685 00686        4  1AD ALVVS MOVEO SERVO
00724 00725 00726 00727        4  1AD ALVVS MOVEO SERVO
00797 00798 00799 00800        4  3IN LAGENA REFRIGESCO VBI~3
00826 00827 00828 00829        4  3IN LAGENA REFRIGESCO VBI~3
```

h. Arrivé à ce stade, on dispose d'un certain nombre de fichiers (en l'occurrence, les fichiers .205, .305 et .404) contenant tous les groupes récurrents du texte *sans recouvrement ni duplication.* Ils sont triés dans l'ordre alphabétique des séries de lemmes en un seul fichier (.T01). Deux extraits :

```
00645 00646                    2  1ABLAQVEO~1 SIGNO
00735 00736                    2  1ABLAQVEO~1 SIGNO
00634 00635                    2  2ABLAQVEO~1 VITIS
00644 00645                    2  2ABLAQVEO~1 VITIS
00013 00014                    2  3ACER~2 ACETVM
00062 00063                    2  3ACER~2 ACETVM
00880 00881                    2  4AD ALVVS
00729 00730 00731              3  1AD ALVVS MOVEO
00683 00684 00685 00686        4  1AD ALVVS MOVEO SERVO
00724 00725 00726 00727        4  1AD ALVVS MOVEO SERVO
00417 00418                    2  5AD VINDEMIA
00862 00863                    2  5AD VINDEMIA
00728 00729                    2  6AD VINVM
00808 00809                    2  6AD VINVM
00036 00037                    2  7ADDO AQVA
00318 00319                    2  7ADDO AQVA
00396 00397 00398              3  2ALTER DOLIVM IN
00409 00410 00411              3  2ALTER DOLIVM IN
00704 00705                    2  8ALVVS MOVEO
00767 00768                    2  8ALVVS MOVEO
00143 00144 00145              3  3AMPHORA DIFFVNDO~2 IN
```

```
00581 00582 00583            3  3AMPHORA DIFFVNDO~2 IN
00598 00699                  2  9AMPHORA IN
00713 00714                  2  9AMPHORA IN
00201 00202                  2 26DIVVM SVB
00441 00442                  2 26DIVVM SVB
00116 00117 00118 00119      4  2DOLIVM EODEM IN INFVNDO~2
00129 00130 00131 00132      4  2DOLIVM EODEM IN INFVNDO~2
00044 00045                  2 27DOLIVM IN
00095 00096                  2 27DOLIVM IN
00210 00211                  2 27DOLIVM IN
00290 00291                  2 27DOLIVM IN
00467 00468                  2 27DOLIVM IN
00516 00517                  2 27DOLIVM IN
00542 00543                  2 27DOLIVM IN
00010 00011 00012            3 11DOLIVM IN INDO
00563 00564 00565            3 11DOLIVM IN INDO
00375 00376 00377            3 12DOLIVM IN INFVNDO~2
00254 00255 00256            3 13DOLIVM IN VINVM
00562 00563 00564            3 13DOLIVM IN VINVM
00139 00140                  2 30DOLIVM OBLINO
00256 00257                  2 30DOLIVM OBLINO
00296 00297                  2 30DOLIVM OBLINO
```

Ce fichier est lu par un quatrième programme qui isole toutes les séries se rapportant à un même groupe et attribue à celui-ci un numéro; il crée deux fichiers, l'un recevant les groupes ainsi distingués (.FOR) et le second les séries de numéros d'ordre de leurs occurrences (.T02, trié en .T03 sur les numéros d'ordre), chacune étant précédée du numéro de son groupe. Il reste alors à les utiliser pour deux opérations qui seront décrites plus loin : la constitution d'une concordance rassemblant les contextes de toutes les occurrences des groupes récurrents répérés; la mise en relation des groupes apparentés (cfr *infra*, D. et E.).

B. *La recherche en ordre d'index.*

La seconde méthode, qui prend pour base le fichier trié en ordre d'index (ordre alphabétique des lemmes et ordre croissant des références) est plus courte : elle comprend six étapes, soit trois programmes et trois tris, et ne nécessite qu'une seule application. Elle permet donc d'identifier tous les groupes récurrents, de toutes longueurs, en une seule fois.

a. Lisant le fichier .NDX, le premier programme isole, successivement pour chaque lemme, tous les enregistrements pourvus d'un numéro d'ordre (c'est-à-dire ne correspondant ni à une conjonction de coordination, ni à SVM auxiliaire); il apparie chaque occurrence à chacune de celles qui la suivent dans la liste et

crée pour chaque association un enregistrement dans un nouveau fichier (.001), comprenant : a. les deux numéros de phrases des deux occurrences du lemme; b. les deux numéros d'ordre dans le texte; c. les deux premiers codes de l'analyse du lemme (catégorie et sous-catégorie); d. le lemme. Exemple, pour trois lemmes consécutifs :

```
     a          b
  1    2     1    2    c d
┌────┬────┬────┬────┬──┬─┐
00002000130003600240 53ADDO
00002000170003600319 53ADDO
00002000250003600584 53ADDO
00013000170024000319 53ADDO
00013000250024000584 53ADDO
00017000250031900584 53ADDO
00020000200039700410 48ALTER
00027000300062800684 12ALVVS
00027000310062800704 12ALVVS
00027000330062800726 12ALVVS
00027000340062800730 12ALVVS
00027000360062800767 12ALVVS
00027000420062800881 12ALVVS
00030000310068400704 12ALVVS
00030000330068400726 12ALVVS
00030000340068400730 12ALVVS
00030000360068400767 12ALVVS
00030000420068400881 12ALVVS
00031000330070400726 12ALVVS
00031000340070400730 12ALVVS
00031000360070400767 12ALVVS
00031000420070400881 12ALVVS
00033000340072600730 12ALVVS
00033000360072600767 12ALVVS
00033000420072600881 12ALVVS
00034000360073000767 12ALVVS
00034000420073000881 12ALVVS
00036000420076700881 12ALVVS
```

b. Le fichier .001 est trié dans l'ordre croissant des quinze premières positions (les deux numéros de phrases et le premier numéro d'ordre); on obtient de la sorte un fichier rassemblant en séquence tous les lemmes apparaissant dans les deux mêmes phrases. Exemple, pour les phrases 8 et 24 :

```
00008000240013600578 70POST 2
00008000240013700579 15DIES
00008000240014300581 53DIFFVNDO 2
00008000240014400582 70IN
00008000240014400585 70IN
00008000240014400599 70IN
00008000240014400614 70IN
00008000240014500583 11AMPHORA
00008000240014500587 11AMPHORA
```

```
00008000240014500591 11AMPHORA
00008000240014500598 11AMPHORA
00008000240014500606 11AMPHORA
00008000240014700582 70IN
00008000240014700585 70IN
00008000240014700599 70IN
00008000240014700614 70IN
00008000240014800600 13SOL
00008000240014800615 13SOL
00008000240014900616 53SINO
00008000240015000606 15CPONO
00008000240015200582 70IN
00008000240015200585 70IN
00008000240015200599 70IN
00008000240015200614 70IN
```

c. Le deuxième programme isole chaque série d'enregistrements présentant la même paire de numéros de phrases, classés par l'opération précédente dans l'ordre croissant des numéros d'ordre de la première phrase (3e colonne). Il parcourt cette liste de haut en bas pour repérer des séries séquentielles (marquées ci-dessus) de numéros d'ordre en troisième colonne (première phrase), sans intervalle (la différence doit toujours être 1); dès qu'il a repéré le début et la fin d'une telle séquence, il la trie sur la quatrième colonne (numéro d'ordre de la deuxième phrase). Soit, pour l'exemple précédent, les séquences suivantes :

```
00008000240013600578 70POST*2
00008000240013700579 15DIES
00008000240014300581 53DIFFVNDO*2
00008000240014400582 70IN
00008000240014500583 11AMPHORA
00008000240014400585 70IN
00008000240014500587 11AMPHORA
00008000240014500591 11AMPHORA
00008000240014500598 11AMPHORA
00008000240014400599 70IN
00008000240014500606 11AMPHORA
00008000240014400614 70IN
00008000240014700582 70IN
00008000240014700585 70IN
00008000240014700599 70IN
00008000240014800600 13SOL
00008000240015000606 15CPONO
00008000240014700614 70IN
00008000240014800615 13SOL
00008000240014900616 53SINO
```

Il est ainsi possible de distinguer, à l'intérieur d'une séquence de la première phrase, des séries séquentielles plus courtes encore dans la seconde phrase. Chaque groupe ainsi identifié est soumis aux tests de pertinence (c'est pour cette raison que les deux codes d'analyse figurent dans l'enregistrement) et, s'il y satisfait, il est conservé dans un enregistrement (fichier .003) comprenant :

a. la série d'éléments qui le constituent, placés en ordre alphabétique; b. les deux numéros de phrases; c. les numéros d'ordre de la première occurrence et d. ceux de la seconde. Exemple pour les groupes de l'échantillon ci-dessus :

```
a           b          c          d
DIES POST*2 /0000800024 00136 00137/00578 00579
AMPHORA DIFFVNDO*2 IN/0000800024 00143 00144 00145/ 00581 00582
    00583
AMPHORA IN/0000800024 00144 00145/ 00598 00599
IN SOL /0000800024 00147 00148/ 00599 00600
IN SINO SOL /0000800024 00147 00148 00149/ 00614 00615 00616
```

Le programme qui vient d'être décrit appelle quelques remarques. Tout d'abord, et sans entrer dans les détails, on notera qu'il a été nécessaire d'y introduire plusieurs procédures de sécurité, destinées à éviter la prise en compte de séries non valides dans des cas de figure bien précis. Ensuite, on a vu que dans une série de lemmes appartenant aux deux mêmes phrases, le programme repère les numéros d'ordre se suivant sans intervalle; dans ce cas, le pas de la différence autorisée entre deux numéros d'ordre est de 1; la procédure permet de choisir un pas plus large et, partant, de sélectionner des séries récurrentes non séquentielles. Les résultats produits sont plus nombreux et plus riches; cependant, en examinant quelques exemples (en 3.2.), on verra que cet enrichissement ne va pas sans accroître le "bruit" mêlé aux résultats pertinents.

d. Le fichier ainsi constitué (.003) est trié dans l'ordre alphabétique des groupes. Voici le début du nouveau fichier (.004) :

```
ABLAQVEO*1 SIGNO /0002600033 00645 00646/ 00735 00736
ABLAQVEO*1 VITIS /0002600026 00634 00635/ 00644 00645
ACER*2 ACETVM /0000100004 00013 00014/ 00062 00063
AD ALVVS /0002900041 00683 00684/ 00880 00881
AD ALVVS /0003200041 00725 00726/ 00880 00881
AD ALVVS /0003300041 00729 00730/ 00880 00881
AD ALVVS MOVEO /0002900033 00683 00684 00685/ 00729 00730 00731
AD ALVVS MOVEO /0003200033 00725 00726 00727/ 00729 00730 00731
AD ALVVS MOVEO SERVO /0002900032 00683 00684 00685 00686/ 00724
    00725 00726 00727
AD VINDEMIA /0001900040 00417 00418/ 00862 00863
AD VINVM /0003300038 00728 00729/ 00808 00809
ADDO AQVA /0000200016 00036 00037/ 00318 00319
ALTER DOLIVM IN /0001900019 00396 00397 00398/ 00409 00410 00411
ALVVS MOVEO /0002900030 00684 00685/ 00704 00705
ALVVS MOVEO /0002900035 00684 00685/ 00767 00768
ALVVS MOVEO /0003000032 00704 00705/ 00726 00727
ALVVS MOVEO /0003000033 00704 00705/ 00730 00731
ALVVS MOVEO /0003000035 00704 00705/ 00767 00768
ALVVS MOVEO /0003200035 00726 00727/ 00767 00768
ALVVS MOVEO /0003300035 00730 00731/ 00767 00768
```

e. Le dernier programme lit le fichier .004, et effectue les mêmes opérations que le quatrième programme dans la méthode en ordre du texte : il distingue chaque groupe récurrent, le numérote, le stocke dans un fichier .FOR et envoie dans le fichier .005 ses différentes séries de numéros d'ordre. Ce dernier fichier est trié dans l'ordre des numéros (.006). Tout comme dans l'autre méthode, c'est à ce stade-ci que l'on passe aux deux dernières opérations. Avant de les décrire (D. et E.), il convient d'évaluer les avantages et inconvénients respectifs des deux méthodes qui viennent d'être décrites.

C. *Comparaison des deux méthodes*

Le principal inconvénient de la méthode en ordre d'index réside dans la croissance exponentielle du volume de la première sortie produite, à savoir la création des paires de références (fichier .001). Si l'on compare les volumes atteints pour les principaux fichiers par les deux méthodes, on voit que la première est nettement plus économique : pour un corpus-échantillon de 887 mots (sans les conjonctions de coordination et les auxiliaires), soit 335 lemmes différents, on obtient les volumes suivants :

méthode	fichiers	nombre de Kb
ordre d'index	TXT, NDX	73
	001, 002	112
	003, 004	10
	005, 006	3
ordre du texte	201, 202	22
	203, 204, 205	10
	301, 302	34
	303, 304, 305	4
	401, 402	42
	403, 404	1
	T01	14

Cette multiplication exponentielle de l'information tient à sa production combinatoire : dès lors que l'on associe chaque occurrence d'un lemme avec chaque autre, on doit attendre que les lemmes les plus fréquents, comme nombre de mots grammaticaux, se revèlent particulièrement prolifiques.

On a déjà noté les deux avantages de la méthode en ordre d'index : elle permet de repérer automatiquement en une seule séquence d'opérations tous les groupes récurrents, quelle que soit leur longueur, alors que l'autre

méthode nécessite un nombre indéterminé d'applications, selon le nombre de niveaux de combinaisons à explorer; enfin elle autorise, de manière très souple, à tenir compte d'intervalles entre les éléments d'un groupe récurrent et peut donc ainsi se libérer du principe de contiguïté qui s'était imposé. Dans l'autre méthode, une telle option, qui *a priori* n'est pas impossible, poserait davantage de problèmes pratiques.

L'avantage principal de la méthode en ordre du texte est son faible coût en mémoire de masse : le nombre d'enregistrements créés par la procédure d'association est, à tous les niveaux, toujours inférieur au nombre d'enregistrements du fichier contenant le texte.

On donnera donc, d'une manière générale, la préférence à la première méthode; pour les corpus les plus volumineux, c'est elle qui s'imposera; on pourra cependant recourir à la seconde si l'on traite de petits fichiers ou si l'on souhaite élargir la recherche des groupes au-delà de la contiguïté.

D. *Constitution de la concordance.*

La meilleure façon d'organiser les résultats est à mon sens de produire une concordance présentant les groupes classés dans l'ordre alphabétique et suivis de leurs occurrences avec contexte et références. Un programme constitue ce fichier automatiquement, à partir du fichier en ordre du texte et du dernier fichier issu de l'une des deux méthodes et contenant toutes les séries de numéros d'ordre, affectées du numéro de leur groupe récurrent (fichier .T03 ou .006). Chaque contexte est affecté du numéro de son groupe récurrent, ce qui permet de l'en rapprocher en triant en un seul ensemble le fichier produit et la liste des groupes. On trouvera le début de cette concordance dans l'annexe et plusieurs extraits dans la section 3.

E. *Comparaisons des groupes récurrents.*

Quelle que soit la méthode employée, une fois le traitement effectué, on dispose d'une liste close de tous les groupes récurrents de toutes longueurs. Un simple coup d'œil sur le début de la concordance permet de remarquer qu'un grand nombre de groupes s'inscrivent dans d'autres, de longueur supérieure. Ainsi AD ALVVS et ALVVS MOVEO sont inclus dans AD ALVVS MOVEO, lui-même compris dans AD ALVVS MOVEO SERVO. Il me paraît important de pouvoir établir la liste complète de ces apparentements, partiels ou complets, qui rapprochent les groupes entre eux, et de mettre en relation de la manière la plus pratique possible les séquences apparentées.

Il est cependant impossible de produire une concordance finale hiérarchisée où chaque groupe récurrent viendrait se ranger avec ses occurrences sous le

groupe de niveau supérieur dont, le cas échéant, il est issu, lui-même subordonné à un autre, plus large encore; la raison en est que tout groupe d'un niveau donné peut entrer dans la composition de plus d'un groupe d'un niveau supérieur; ainsi DOLIVM IN entre dans DOLIVM IN INDO, mais aussi dans DOLIVM IN INFVNDO et dans DOLIVM IN VINVM; une structure arborescente à progression binaire a donc peu de chance d'exister dans la réalité. Pour pallier cette impossibilité, j'ai choisi de mentionner, dans la concordance, sous chaque groupe récurrent et avant ses occurrences, la liste de ses parentés; on voit dans l'exemple ci-dessous que ces relations sont de trois types :

```
37    DOLIVM IN INFVNDO~2
37        -->  DOLIVM EODEM IN INFVNDO~2  = 34    3/3 -> 4
37        <--  DOLIVM IN  = 35    2/2 -> 3
37        <->  ALTER DOLIVM IN  = 10    2/3 -> 3
37        <->  DOLIVM IN INDO  = 36    2/3 -> 3
37        <->  DOLIVM IN VINVM  = 38    2/3 -> 3
37 112,01 ubi hauseris de mari *in dolium infundito*
          nolito implere quadrantalibus quinque minus
```

- l'*inclusion* : un groupe constitue une partie d'un autre plus long (flèche vers la droite -->).
- la *production* : un groupe plus long contient tous les éléments d'un autre (flèche vers la gauche <--).
- le *partage* : deux groupes ont en commun un nombre d'éléments supérieur à 1 et inférieur à leurs longueurs respectives (flèche gauche et droite <->).

Le renvoi vers un groupe supérieur permet de retrouver les occurrences d'un groupe inclus; seules figurent sous un groupe les occurrences qu'aucun développement à gauche ou à droite dans le texte ne fait passer à un groupe plus long. Pour l'exemple ci-dessus, les autres occurrences de DOLIVM IN INFVNDO~2 se trouvent sous DOLIVM EODEM IN INFVNDO~2.

Les indications numériques à la suite de chaque relation sont : le numéro du groupe relié; le nombre d'éléments impliqués sur le total du groupe entrant (2/2, 2/3); le nombre d'éléments du groupe-hôte.

Enfin, il est facile de produire un index des vocables impliqués dans les groupes récurrents d'un texte, avec renvoi aux groupes concernés.

3. Analyse des résultats.

Après avoir décrit les procédures permettant de produire la liste des groupes récurrents d'un texte, j'examinerai les résultats obtenus, sous le seul angle de leur contenu grammatical; l'analyse du contenu sémantique des formules apparaissant dans le texte du *Agricultura* de Caton fera l'objet d'une étude future.

3.1. Le contenu grammatical et la validité des groupes.

Une réflexion intéressante des membres du Laboratoire de Saint-Cloud, bien que ne portant que sur les cooccurrences, peut nous servir d'introduction[11] :

> "La proximité de deux formes est due à différentes causes : elle peut avoir trait à l'ordre syntaxique, ou à des faits sémantiques, ou même à quelque effet du hasard. Les deux formes apparaissent dans la même phrase et c'est la seule chose que l'ordinateur puisse détecter sans analyse linguistique préalable. C'est au moment de l'interprétation que nous devons clarifier les raisons qui ont produit les combinaisons statistiquement remarquables."

On en retiendra, pour la recherche de groupes récurrents, qu'ici aussi il faut compter avec la part de hasard et que l'action de l'ordinateur, si précises et sévères que soient les règles dont on lui impose le respect, ne peut, en aucun cas, éviter la production de cas non pertinents, ni dispenser le chercheur d'une épuration des résultats.

Parmi les groupes tirés des trois textes étudiés, les exemples suivants sont à rejeter. On voit qu'ils ont échappé au filtre des critères de sélection basés sur la nature grammaticale des lemmes pris en compte; les uns présentent un substantif et une préposition, les autres, un subordonnant et un verbe, mais sans relation de dépendance entre ces éléments :

uinum ad *aluum mouendam concinnare* (CAT., 115,2) / uinum ad *isciacos sic facito* (123,1)

ne plus quadriennium in *sole siueris* (113,2) / *post* quadriennium in *cuneum componito* (113,2)

ad deorum *immortalium numen* (CIC., 5) / ad deorum *religionem* (30)

pro re publica (36) / pro rei publicae *salute* (38)

operito ne *odor exeat* (CAT., 113,1) / *amphoras* operito ne *aqua accedat* (113,2)

Il doit être possible, en affinant la méthode, de repérer ce type de cas parasites, par exemple en se basant sur l'ordre des éléments ou les codes de cas

---

[11] "... proximity of two forms is due to different causes : it may be related to syntactic order, or to semantics facts, or even to some chance effect. The two forms appear in the same sentence and this is the only thing the computer can detect without preliminary linguistic analysis. It is when interpreting that we have to clarify the reasons that have produced the statistically remarkable combinations." (GEFFROY Annie, LAFON Pierre, SEIDEL Gill, TOURNIER Maurice, *Lexicometric Analysis of Co-occurrences*, dans AITKEN A.J., BAILEY R.W., HAMILTON-SMITH N. ed., *The Computer and Literary Studies*, Edinburgh, University Press, 1973, p. 131.).

dans le cas d'une préposition et d'un substantif, ou en contrôlant les codes de subordination attribués aux verbes, dans le second cas.

On trouve un second type de groupes, un peu plus intéressants déjà, dans les paires de termes contigus sans lien syntaxique; chacun exerce dans la phrase une fonction indépendante de l'autre :

> *si uoles scire in* uinum aqua *sit necne* / *si habebit* aquam, uinum *effluet* (CAT., 111,1)

Cette classe, assez rare, constitue en fait un cas particulier de la cooccurrence, celui de la récurrence en séquence d'un paire de vocables.

Une troisième catégorie est celle des groupes dont les termes, toujours en liaison syntaxique, présentent, d'une occurrence à l'autre, un lien différent :

> *uites* ablaqueato et signato (CAT., 114,1) / *uites cum* ablaqueuntur signato (115,2)
> *salutem* rei publicae defenderunt (CIC., 27) / *ad* rem publicam defendendam (35)
> *in* magistratibus gerendis (27) / magistratum gerere (28) ius libertatis (11) / iuris et libertatis (12)

Ces groupes présentent souvent une indéniable cohérence, mais on ne peut voir en eux des formules; le premier exemple est plutôt une cooccurrence.

La dernière classe, la plus fréquente, concerne les groupes dont les termes sont unis par le même lien syntaxique dans toutes les occurrences. Parmi ceux-ci, on peut distinguer les cas de coordination de deux termes ou plus :

> oro atque obsecro (CIC., 5) / orat atque obsecrat (37)

Certaines structures sont simples (substantif + adjectif, *aqua marina*; préposition + substantif, *ad uindemiam*), d'autres relativement complexes (*qui rem publicam saluam esse uellent*, CIC. 20 et 34).

3.2. La recherche de groupes récurrents avec intervalles autorisés.

On a vu que la méthode en ordre d'index autorisait la prise en compte d'un intervalle entre les termes supposés constituer un groupe récurrent. Procéder de la sorte, en renonçant au principe de contiguïté, permet effectivement de découvrir de nouveaux groupes récurrents dont les éléments, en certaines occurrences, sont séparés par une ou deux formes, indépendantes ou non. Quelques exemples :

> in alterum dolium *item* transfundito (CAT., 112,2) / transfundito in alterum dolium (112,2)
> commisceto cum *eo* uino (109,1) / *ne* commisceas cum *cetero* uino (114,2)

uinum *Graecum* sic facito (105,1) / uinum *murteum* sic facito (125,1)
operculum *in dolio* imponito (104,2) / operculum imponito (112,1)

Il arrive que deux groupes n'en fassent plus qu'un seul :

uinum Coum *si* uoles facere (112,1) / uinum Coum facere uoles (112,2)

ou qu'apparaisse une nouvelle occurrence d'un groupe déjà identifié :

in dolium *quadragenarium* infundito (105,1)
aquam *ex alto* marinam *sumito* (112,1)

Pour l'échantillon choisi, on passe de 65 groupes différents sans intervalle autorisé à 98, si l'on accepte un intervalle d'un mot, et à 125 pour un intervalle de deux mots.

Cependant, les risques de découvrir des séries non homogènes sont plus grands que dans la méthode normale. Exemples :

eas radices *dato* circum *uitem* (114,11) / *circumponito* circum radices (115,2)
cum eo uino (109,1) / eam *inferuefacito* cum *congio* uini *ueteris* (123)
hoc uinum deterius *non erit quam Coum* (105,2) / *odorem* deteriorem *demere* uino (110,1)

Il sera donc généralement très utile de développer une étude de groupes récurrents par des recherches tenant compte d'intervalles plus ou moins larges; on retiendra toutefois l'inconvénient que cela peut entraîner.

4. Développements.

Pour conclure, je proposerai, pour ces recherches, quelques développements et exploitations, ainsi que le projet d'une importante amélioration.

4.1. Il est possible d'envisager les rapports des groupes récurrents d'un texte entre eux autrement que par le repérage de leurs éléments communs. Un type d'investigation intéressant, inspiré de ce qui se fait pour les cooccurrences, serait d'évaluer pour chaque groupe la distance moyenne séparant ses diverses occurrences. On devrait trouver là une manière supplémentaire de les classer; on serait mieux en mesure, en tout cas, de distinguer les véritables formules de ce qui n'est que simple répétition d'une expression récemment employée dans le texte et dont auteur et lecteur se souviennent.

La recherche des cooccurrences de groupes récurrents, quant à elle, permettra d'identifier, dans l'œuvre étudiée, les chapitres ou paragraphes plus riches en formules, plus typiquement "formulaires" ; si ces cooccurrences de formules sont

répétées, il sera possible de déceler quels chapitres sont plus particulièrement liés les uns aux autres par leur formulaire et, partant de là, d'appréhender par une voie nouvelle la structure du texte.

4.2. La recherche de groupes récurrents sur base d'un codage syntaxique.

Lors du 5e Congrès international de linguistique latine, tenu récemment à Louvain-la-Neuve, M. Ét. Évrard a proposé un enrichissement des procédures informatiques de lemmatisation et d'analyse morphologique du LASLA qui pourrait s'avérer très utile pour de futurs développements des méthodes de recherche automatique de groupes récurrents[12]. Sa suggestion est d'abord destinée à faciliter les études syntaxiques, mais on va voir qu'il serait aussi possible d'en faire usage dans la perspective de ce travail. Se basant sur la syntaxe dépendancielle de Lucien Tesnière, telle qu'elle est développée dans ses *Éléments de syntaxe structurale*, Ét. Évrard a proposé d'enrichir les fichiers de textes latins lemmatisés et analysés morphologiquement qui sont produits et exploités au LASLA de la manière suivante : dès lors que chaque mot d'une phrase est régi par un autre mot du texte (à l'exception de la base de la phrase qui ne dépend d'aucun mot) et, qu'à l'inverse, la plupart des mots en régissent d'autres, on peut coder la dépendance de chaque forme en reportant manuellement dans son enregistrement le numéro d'ordre dans le texte de son régissant, c'est-à-dire du mot auquel elle est subordonnée. Lorsque cette opération est effectuée pour tous les mots de chaque phrase, il est possible de se livrer à tous les traitements qu'envisage Ét. Évrard, dans une perspective syntaxique : étudier le sens de la rection (dépendance), l'écart entre le régissant et le régi, les associations en dépendance des différentes catégories grammaticales, etc. En ce qui concerne le sujet de ce travail, il me semble qu'un tel codage pourrait grandement faciliter la recherche automatique des groupes récurrents. En effet, il permet de redistribuer chaque phrase selon sa structure, c'est-à-dire de disposer en séquences toutes les formes de la phrase, selon leur ordre de dépendance syntaxique; une telle opération, préalable à des procédures de recherche automatique de groupes récurrents, présenterait l'avantage de pallier pour une grande part les inconvénients que j'ai dégagés précédemment et qui sont dus à la nature même de la langue latine : ordre des mots extrêmement variable, à l'intérieur de la phrase comme à l'intérieur d'une groupe récurrent, même homogène; inclusion d'éléments étrangers à l'intérieur d'un groupe syntaxique; etc. Sans doute,

[12] ÉVRARD Étienne, *Une informatisation de la syntaxe de dépendance en latin*, dans *Actes du V^e Congrès international de linguistique latine* (Louvain-La-Neuve/Borzée, 31 mars – 4 avril 1989), Louvain-La-Neuve, 1989, pp. 115–126.

cette nouvelle voie présenterait d'autres difficultés à surmonter (premièrement, chaque régissant étant susceptible de régir plus d'une forme, les directions dans lesquelles une formule peut se développer à partir d'une forme s'en trouvent multipliées; d'autre part, la dépendance n'est pas le seul système présidant à la formation des groupes récurrents : la juxtaposition et la coordination peuvent aussi y participer), mais elle aurait au moins l'avantage de permettre de distinguer assez rapidement et automatiquement ce qui, parmi les groupes récurrents identifiés, est syntaxiquement cohérent de ce qui est fortuit et agrammatical.

Annexe : Concordance des groupes récurrents.

```
1    ABLAQVEO 1 SIGNO
1 114,01 satis esse uini tot uites *ablaqueato et signato*
1 115,02 aluum mouendam concinnare uites cum *ablaqueabuntur
           signato* rubrica ne admisceas cum cetero
2    ABLAQVEO 1 VITIS
2 114,01 bonam faciat secundum uindemiam ubi *uites
           ablaqueantur* quantum putabis ei rei satis
2 114,01 rei satis esse uini tot *uites ablaqueato* et signato
3    ACER 2 ACETVM
3 104,01 q(uadrantalia) x in dolium indito *aceti
           acris* q(uadrantalia) ii eodem infundito sapae
3 104,02 si quid superfuerit post solstitium *acetum
           acerrimum* et pulcherrimum erit
4    AD ALVVS
4       -->  AD ALVVS MOVEO  =  5    2/2 -> 3
4       -->  AD ALVVS MOVEO SERVO  =  6    2/2 -> 4
4 125,01 id est *ad aluum* crudam et ad lateris dolorem
5    AD ALVVS MOVEO
5       -->  AD ALVVS MOVEO SERVO  =  6    3/3 -> 4
5       <--  AD ALVVS  =  4    2/2 -> 3
5       <--  ALVVS MOVEO  = 11    2/2 -> 3
5 115,02 uinum *ad aluum mouendam* concinnare uites cum
           ablaqueabuntur signato
6    AD ALVVS MOVEO SERVO
6       <--  AD ALVVS  =  4    2/2 -> 4
6       <--  AD ALVVS MOVEO  =  5    3/3 -> 4
6       <--  ALVVS MOVEO  = 11    2/2 -> 4
6 114,02 si uoles seruare in uetustatem *ad aluum
           mouendam seruato* ne commisceas cum cetero uino
6 115,01 id uinum *seruato ad aluum mouendam*
7    AD VINDEMIA
7 112,02 item transfundito ita relinquito usque *ad uindemiam*
7 125,01 ubi iam passa erit seruato *ad uindemiam* in
           urnam musticontundito murtae
```

```
8    AD VINVM
8 115,02 *uinum ad* aluum mouendam concinnare uites cum
8 123,01 *uinum ad* isciacos sic facito de iunipiro
9    ADDO AQVA
9 104,02 die dies quinque continuos eo *addito aquae*
           marinae ueteris sextarios lxiiii et
9 111,01 si uoles scire in uinum *aqua addita sit* nec
           ne uasculum facito de materia
10    ALTER DOLIVM IN
10        <--  DOLIVM IN  = 35    2/2 -> 3
10        <->  DOLIVM EODEM IN INFVNDO 2  = 34    2/3 -> 4
10        <->  DOLIVM IN INDO  = 36    2/3 -> 3
10        <->  DOLIVM IN INFVNDO 2  = 37    2/3 -> 3
10        <->  DOLIVM IN VINVM  = 38    2/3 -> 3
10 112,02 quod desiderit post dies xx *in alterum
           dolium* item transfundito ita relinquito usque
10 112,02 ubi dies xxx praeterierint transfundito *in
           alterum dolium* puriter et leniter relinquito in imo
11    ALVVS MOVEO
11        -->  AD ALVVS MOVEO  =  5    2/2 -> 3
11        -->  AD ALVVS MOVEO SERVO  =  6    2/2 -> 4
11 114,02 bibito ante cenam sine periculo *aluum mouebit*
11 115,02 cyatum in ceteram potionem indito *aluum mouebit* et
           postridie perpurgabit sine periculo
12    AMPHORA DIFFVNDO 2 IN
12        <--  AMPHORA IN  = 13    2/2 -> 3
12 105,02 xxx dolium oblinito ad uer *diffundito in
           amphoras* biennium in sole sinito positum esse
12 113,02 post dies xl *diffundito in amphoras* et
           addito in singulas amphoras sapae
13    AMPHORA IN
13        -->  AMPHORA DIFFVNDO 2 IN  = 12    2/2 -> 3
13 113,02 implere nimium ansarum infimarum fini et
           *amphoras in* sole ponito ubi herba non
13 115,01 mustum ueratri atri manipulum coicito *in
           amphoram* ubi satis efferuerit de uino
14    AQVA DVLCIS
14        -->  AQVA DVLCIS QVADRANTAL  = 15    2/2 -> 3
14 112,01 dies lxx ante uindemiam quo *aqua dulcis* non perueniet
15    AQVA DVLCIS QVADRANTAL
15        <--  AQVA DVLCIS  = 14    2/2 -> 3
15 104,01 eodem infundito sapae quadrantalia ii *aquae
           dulcis q(uadrantalia)*
15 105,01 quadragenarium infundito seorsum in uas *aquae
           dulcis q(uadrantal)* i infundito salis mo(dium) i
16    AQVA HABEO
16 111,01 aquam habere eo demittito si *habebit aquam*
           uinum effluet aqua manebit nam non
16 111,01 hederacia uinum id quod putabis *aquam habere*
           eo demittito si habebit aquam
```