Logiciels d'exploitation de la banque de données de textes latins du L.A.S.L.A.

Sylvie MELLET

Abstract. Firstly will be repeated the typical features of the L.A.S.L.A. textual database and underlined the great usefulness of the lemmatization for the Latin texts. Then will be presented two programmes, aimed at doing automatic research not only on lemmas or forms, but also on grammatical categories such as nominal casus, tense and mood, or subordinate clause and cooccurrence of two of these criteria. Furthermore, thanks to *Estela* programme it is possible as well to do indices of frequency, to create numerical tables which are then available for summing up the data, and last to list the lemmas which include some prefix or other, some suffix or other.

Keywords: Textual database, lemmatization, textual data processing analysis, latin, cooccurrences, index of frequency, quantification.

Mots-clés : Base de données textuelles, lemmatisation, traitement informatique des textes, latin, cooccurrences, index fréquentiel, quantification.

On voudrait présenter ici deux logiciels d'exploitation automatique du texte, applicables aux textes latins analysés et codés par le L.A.S.L.A.[1]. L'un d'eux a été mis au point, depuis quelque temps déjà, par J. Denooz; l'autre, plus récent, a été élaboré par Y. Deschamps[2] et S. Mellet, et peut être fourni à tout lecteur intéressé[3]. Créés pour répondre à la demande de divers chercheurs travaillant sur le lexique, la morpho-syntaxe ou la stylistique latine à partir

[1] La liste de ces textes peut être obtenue sur simple demande auprès du L.A.S.L.A.; elle s'allonge très régulièrement.

[2] Ancien maître-assistant à l'université d'Orsay.

[3] Ce logiciel a été baptisé *Estela* (Exploitation Statistique des TExtes LAtins); s'adresser à l'auteur.

✉ INaLF (UPR 6861); Faculté des Lettres; 98, boulevard Édouard Herriot; BP 209; F–06204 Nice Cedex 3 (France).
Fax : + 33 93 37 54 45 E-mail : mellet@naxos.unice.fr

des textes lemmatisés par le L.A.S.L.A., ces deux logiciels offrent certains points communs (ils répondent souvent aux mêmes questions), mais sont de conceptions assez différentes; chacun, sans doute, a ses avantages propres et ses inconvénients ou ses faiblesses : nous laisserons le lecteur en juger à travers les présentations qui en sont faites ici. Cependant, l'un et l'autre s'appuient fondamentalement sur l'analyse que les programmes du L.A.S.L.A.[4] ont préalablement associée au texte et qui est enregistrée avec lui dans un fichier que nous appellerons désormais le fichier «TEXTE». Les avantages de cette analyse — et notamment de la lemmatisation qui l'accompagne — sont tels et ses caractéristiques sont à ce point déterminantes dans la conception des deux logiciels que nous commencerons par un bref rappel à ce sujet avant d'aborder la présentation proprement dite du logiciel de J. Denooz et d'*Estela*.

1. Les caractéristiques de la banque de données : rappels

1.1. Le codage des formes et les différents fichiers

La principale caractéristique de la banque de donnée du L.A.S.L.A. est, on le sait, d'être constituée de textes lemmatisés, c'est-à-dire dans lesquels chaque forme est dotée d'une analyse grammaticale et se trouve rapportée à son entrée de dictionnaire; l'analyse grammaticale est conservée au moyen d'un code alpha-numérique associé à la forme au sein du fichier «TEXTE». Ainsi, à la forme *ingenio* (Cic., *Lael.* 1) est associé, en plus de sa référence exacte, le code 12F00 :

1 pour «substantif»,
2 pour «deuxième déclinaison»,
F pour «ablatif singulier»;
les deux 0 correspondent à des colonnes inutilisées dans ce cas.

Un programme simple balayant les diverses colonnes de ce code permet donc de différencier les verbes (5 dans la première colonne) des substantifs par exemple (code 1), les ablatifs singuliers (F dans la troisième colonne) des ablatifs pluriels (code 0), les déponents de toutes les conjugaisons (codes J, K, L, M, N ou O dans la deuxième colonne) des formes passives (codes A, B, C, D, E ou F), etc.

---

[4] Cf. DENOOZ (J.) : 1978, «L'ordinateur et le latin, techniques et méthodes», *Revue de l'organisation internationale pour l'étude des langues anciennes par ordinateur*, 4, pp. 1–36.

À la forme *machinaris* (Cic., *Cat.* I, 2) est associé le code `5JB11LN` :

`5` pour « verbe »,
`J` pour « déponent de la première conjugaison »,
`B` pour « 2e personne du singulier »,

`1` pour « indicatif »,
`1` pour « présent »,
`LN` pour « verbe en subordonnée relative introduite par une forme du pronom *qui* »;

en effet, les formes verbales sont dotées, en plus de leur analyse morphologique, d'une analyse syntaxique qui permet de déterminer de quelle proposition elles sont le noyau prédicatif. Chaque type de proposition est caractérisé par un code de deux lettres et l'ensemble des codes attestés dans un texte est regroupé par ordre alphabétique au sein d'un fichier qui constitue ainsi un index des propositions subordonnées; chaque code y est accompagné du nombre total de ses occurrences et des références exactes de celles-ci.

Au total, la lemmatisation a donc permis :

1) la constitution d'un fichier « TEXTE » reprenant le texte latin, mais où chaque forme est en outre affectée de son analyse codée et de sa référence complète;
2) la constitution d'un fichier des lemmes classés par ordre alphabétique (fichier caractérisé par l'extension `.FRQ`); à côté de chaque lemme se trouvent sa fréquence et les numéros d'ordre dans le texte des formes qui constituent les occurrences de ce lemme; le va-et-vient du texte à l'index est donc aisé;
3) la constitution d'un fichier de même type que le précédent, regroupant tous les verbes subordonnés classés selon l'ordre alphabétique du code de subordination qui leur est associé (fichier caractérisé par l'extension `.SUB`); de la même façon, les numéros d'ordre des diverses occurrences de ce code permettent un rapide retour au texte[5].

1.2. L'utilité de la lemmatisation

On a quelques scrupules à revenir une fois encore sur les atouts offerts à tout chercheur par la lemmatisation des textes; pourtant, force est de constater que quelques collègues restent sceptiques, mettant en balance le coût du traitement en temps et en argent. Paradoxalement, ce sont les linguistes francisants

[5] Pour plus de précisions, cf. DENOOZ (J.) : 1989, « Latin et linguistique informatique », *Revue. Informatique et statistique dans les sciences humaines*, 25, pp. 194–204.

travaillant sur le *T.L.F.* qui nous ont parfois envié notre outil de travail; car eux savent bien qu'il y a peu encore il leur était impossible de trier automatiquement les homonymes (par ex. une recherche sur les noms de saison associait au substantif masculin *été* toutes les occurrences du participe passé du verbe *être*); et que, d'autre part, ils doivent toujours multiplier les entrées (donc le temps de recherche) pour dénombrer toutes les occurrences de certains verbes (*all-*, *ir-*, *aill-* ou *sav-*, *su-*, *sach-*, *saur-*, etc.).

De telles recherches, déjà fastidieuses en français, le sont encore davantage en latin, langue flexionnelle. Le système de parenthésage de la finale — qu'on pourrait envisager comme moyen de résoudre la difficulté — s'avère peu performant; dans le meilleur des cas il laisse subsister trop de bruit dans l'information : ainsi une recherche sur *gener(.)* destinée à relever tous les cas obliques de *genus* donnerait également toutes les formes de *gener*, *generi*, celles du verbe *genero*, *-are* et quelques adjectifs ou adverbes tels que *generalis*, *generosus*, *generatim*, etc. Dans le pire des cas, il ne répond pas à la demande et laisse de côté une large part des éléments recherchés : ainsi en va-t-il pour tous les verbes dont le paradigme est constitué à partir d'un phénomène de supplétisme, tels *sum* ou *fero*; cas finalement assez rares, dira-t-on, qui pourraient faire l'objet d'un relevé exhaustif et bénéficier d'un traitement spécifique. Mais une autre caractéristique de la langue latine rend le parenthésage inadéquat : l'orthographe du latin n'est pas entièrement stabilisée; dès lors, on observe de nombreuses fluctuations, par exemple dans la transcription des assimilations consonantiques (*inlicio/illicio, quamquam/quanquam*), des haplologies (*exsilium/exilium, exspecto/expecto*) et du *h* initial (*olus/holus*), voire semi-initial (*exibeo/exhibeo, nil/nihil*). Tous ces exemples montrent que la recherche sur formes graphiques n'est pas, en latin, une méthode satisfaisante.

En outre, en l'absence de lemmatisation, il est bien évident qu'aucun travail de morpho-syntaxe ne saurait s'appuyer sur une recherche automatique : comment, en effet, à partir de la seule reconnaissance formelle des séquences graphiques, la machine pourrait-elle relever tous les imparfaits du texte, tous les comparatifs du texte, tous les ablatifs du texte?

Mais les syntacticiens ne sont pas les seuls intéressés à ce traitement préalable du texte proposé par le L.A.S.L.A.; les philologues et les lexicologues devraient l'être également. Outre que la recherche d'un lemme est, après lemmatisation, l'affaire d'une fraction de seconde (résultat garanti non bruité), la recherche informatisée d'une catégorie grammaticale — comme celle des déponents par exemple — ou celle d'un type de suffixés devient possible; les études diachroniques sur l'évolution de telle ou telle classe de mots trouvent donc là un auxiliaire appréciable.

L'organisation des fichiers offre en effet de nombreuses possibilités d'exploitation : dans le fichier «TEXTE», recherche d'une forme ou de tel code grammatical; dans le fichier .FRQ, recherche d'un lemme particulier ou recherche de lemmes caractérisés par telle ou telle séquence initiale ou finale; dans le fichier .SUB, recherche d'un type de subordination et du mode régi; et, bien sûr, recherche de cooccurrences entre l'un et l'autre de ces critères. C'est ce que réalisent les deux logiciels décrits ci-dessous.

2. Le logiciel L.A.S.L.A.

La base de données de textes latins lemmatisés et analysés du L.A.S.L.A. comporte actuellement quelque deux millions de mots. Elle se constitue depuis 1961, date de création du laboratoire. Les données sont stockées sur ordinateur central et ne sont pas accessibles *on-line*, mais le L.A.S.L.A. dispose de programmes d'interrogation et peut répondre à toute demande.

Dans les années quatre-vingts, des parties de la base de données ont été distribuées à des chercheurs qui en avaient fait la demande, sur des disquettes DOS pour micro-ordinateurs. Le logiciel décrit ici a été réalisé par Joseph Denooz pour permettre une interrogation simple sur PC, à des fins d'utilisation interne au laboratoire et non de diffusion. Il a toutefois été mis à la disposition de quelques chercheurs.

Pour répondre à la demande toujours croissante d'interrogation de sa base de données, le L.A.S.L.A. en diffusera la totalité sur CD-ROM début 1995, avec un logiciel d'interrogation qui devrait permettre d'obtenir facilement une information exacte et précise, notamment de type lexicologique, morphologique et syntaxique, sur la langue latine.

Écrit en BASIC en 1988, le logiciel de J. Denooz fut le premier à permettre une exploitation rapide et facile des textes lemmatisés sur micro-ordinateur.

Le déroulement du menu prévoit d'abord deux questions préliminaires; il demande à l'utilisateur :

1) sur quel(s) texte(s) il souhaite travailler; il lui est loisible d'en sélectionner plusieurs en même temps;
2) sur quel(s) support(s) seront envoyés les résultats : écran, disque, imprimante ou association de plusieurs de ces supports.

Vient ensuite l'étape de la recherche proprement dite. Le menu propose une recherche de lemme, une recherche d'analyse morphologique ou une recherche de type de proposition. Pour chacune d'elles une sous-catégorisation est possible.

Pour les lemmes, il est possible de relever toutes les occurrences dans le texte sélectionné ou de ne retenir que les contextes dans lesquels le lemme intéressant est employé en cooccurrence avec une forme à définir; l'écran correspondant à cette étape est représenté à la figure 1.

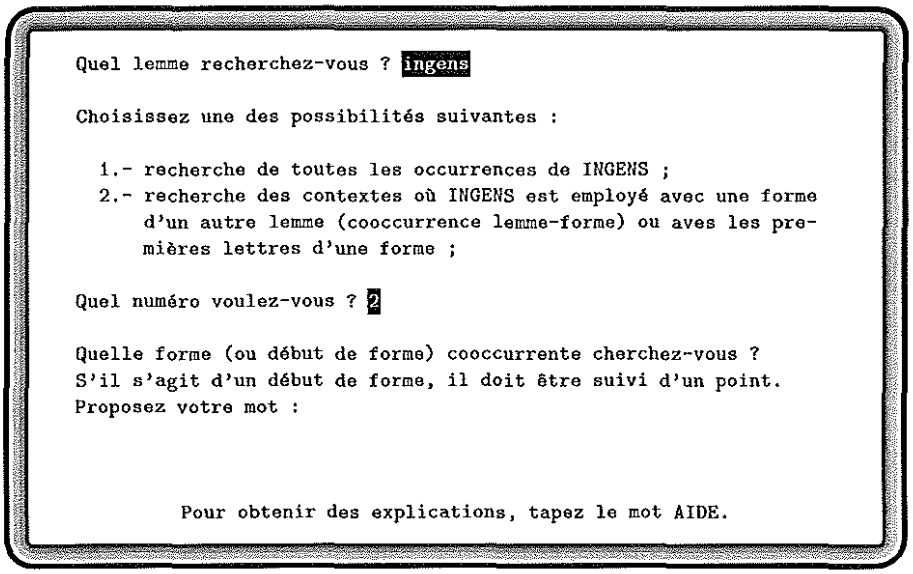


Fig. 1.

La cooccurrence peut avoir pour cadre la phrase toute entière ou un contexte plus réduit, défini par le nombre de mots avant et après le lemme recherché.

Pour l'analyse morphologique, les sous-catégorisations utilisent bien sûr les différentes colonnes du code associé à chaque forme : on pourra donc rechercher tous les verbes d'un texte (figure 2), mais réduire aussi la recherche à tel mode et/ou tel temps verbal (figure 3).

Enfin, pour les types de propositions, on peut sélectionner les principales d'un côté, les subordonnées de l'autre; et, pour ces dernières, faire un nouveau tri en fonction du temps et du mode de son verbe et/ou de la nature de son subordonnant.

Notons que pour toutes ces recherches l'affichage et l'enregistrement dans un fichier peuvent fournir les phrases complètes dans lesquelles apparaissent les éléments sélectionnés ou seulement leurs références dans le texte.

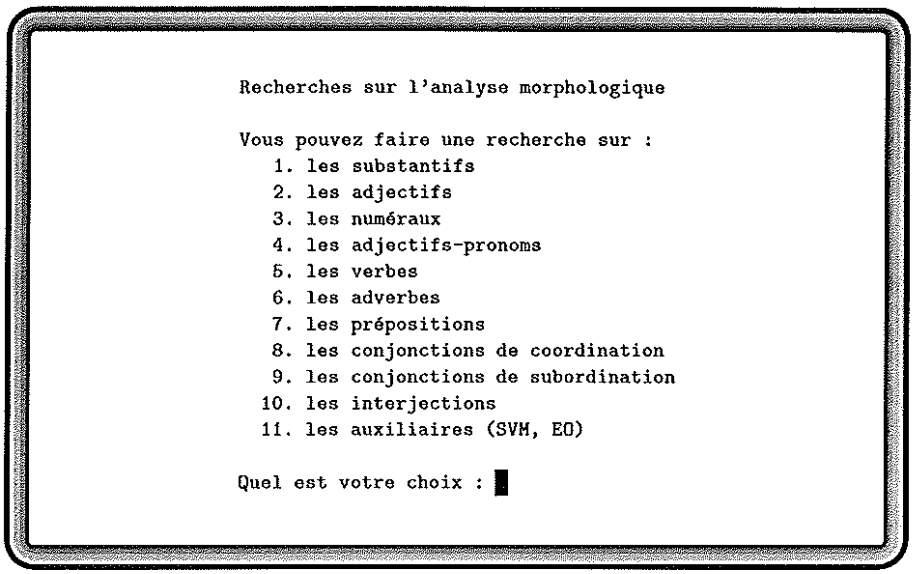


Fig. 2.

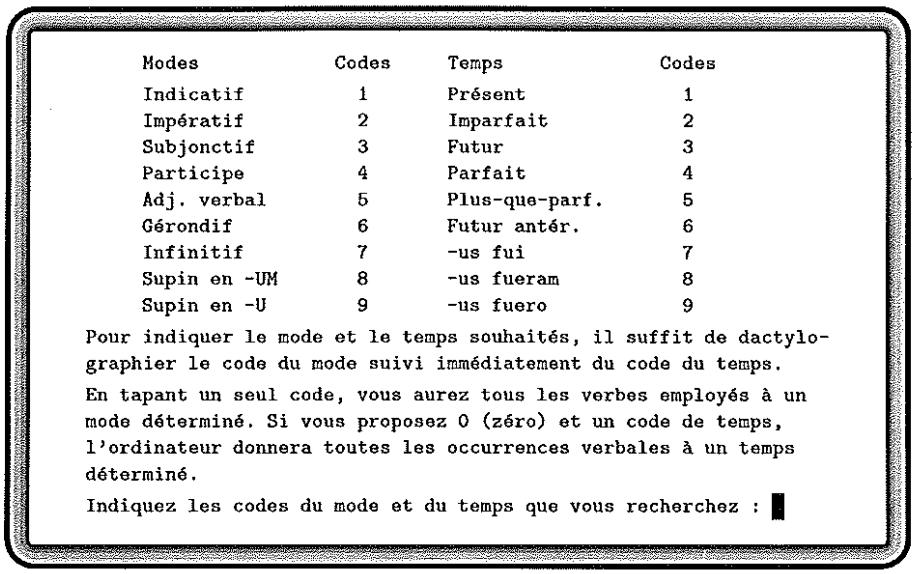


Fig. 3.

Ce logiciel se caractérise à la fois par sa simplicité d'utilisation et par sa très grande fiabilité.

3. Le logiciel *Estela*

Il s'agit d'un logiciel d'exploitation des textes latins lemmatisés dont les programmes sont écrits en langage Turbo C++ et qui est utilisable sur les micros PC ou compatibles travaillant sous un système DOS de niveau 3 minimum.

Les recherches effectuées sont automatiquement enregistrées dans un fichier d'extension `.DOC`; en revanche, et contrairement au logiciel précédent, celui-ci ne prévoit aucune sortie directe imprimante.

3.1. Les différents menus

Après avoir ouvert le programme général et donné le nom du fichier dans lequel seront enregistrés les résultats de la recherche, le logiciel propose différents menus :

```
A. Terminer
B. Analyse
C. Stat. conv.
D. Exploite stat.
E. Fréq. conv.
F. Les fréquences
G. Pref. Suff.
```

parmi lesquels on choisira soit en tapant la lettre correspondante (A, B, C, ...), soit en déplaçant la surbrillance au moyen des touches de direction ↓, ↑ et en tapant ensuite sur la touche Enter [6].

3.1.1. L'analyse

Ce menu offre le choix entre l'analyse morphologique simple ou la recherche de cooccurrences. La première option permet de relever toutes les occurrences d'un lemme, ou celles d'un type de subordonnée, celles d'une forme, celles d'un cas nominal, celles d'une classe de pronoms, ou celles enfin d'une analyse verbale. Ainsi, en sélectionnant la catégorie «NOM» dans la liste, on

[6] Toutes les autres sélections, sur les écrans suivants, se feront par les mêmes moyens.

obtient sur l'écran suivant le choix entre NOMINATIF, VOCATIF, ACCUSATIF, GÉNITIF, DATIF, ABLATIF ou INDIFFÉRENT (cette dernière ligne permettant de sélectionner toutes les formes nominales du texte sans tenir compte du critère casuel). Lorsque le choix est fait, le logiciel offre ensuite la possibilité d'afficher soit la phrase complète dans laquelle se trouve la forme retenue (elle y apparaîtra alors en surbrillance), soit la seule référence de la forme en question. Vient enfin la sélection du texte sur lequel doit porter la recherche.

Le principe est donc de définir d'abord la recherche et de choisir après le texte sur lequel elle s'appliquera. À la fin de la première recherche apparaît sur l'écran la question : « Voulez-vous continuer sur un autre texte ? »; en cas de réponse positive la liste de tous les textes disponibles sur disque dur est à nouveau proposée et l'on peut recommencer sur l'un d'eux la même recherche sans avoir à en définir de nouveau les critères. C'est, pour cette partie de l'analyse, la seule différence fondamentale avec le logiciel L.A.S.L.A. dans lequel le choix du texte servant de base à la recherche se faisait avant la définition des autres critères; pour le reste, les performances de l'un et l'autre programmes sont ici presque identiques. Notons simplement que, lors d'une recherche sur les verbes, la sélection des critères est un peu plus détaillée avec *Estela* qui peut prendre en compte non seulement le temps et le mode, mais encore la voix. Les écrans successifs sont représentés aux figures 4 à 8.

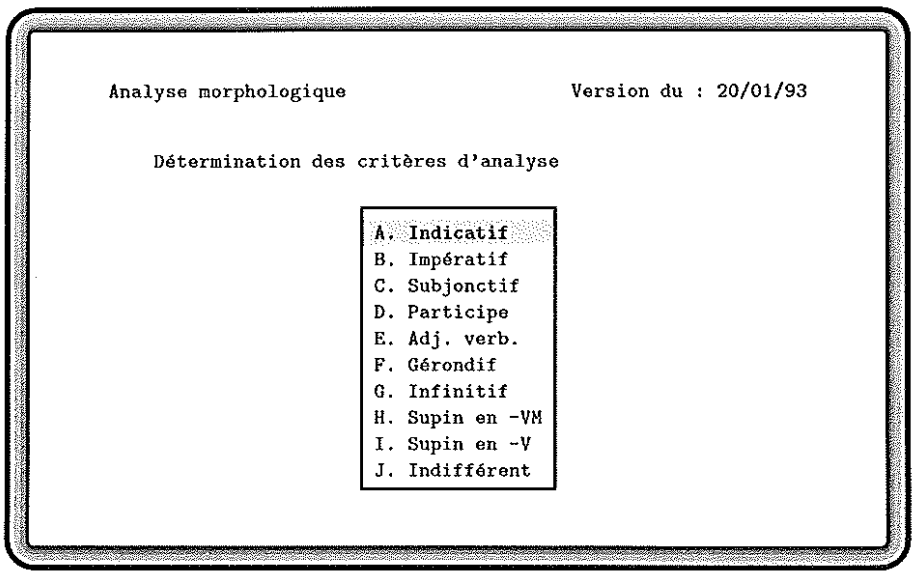


Fig. 4.

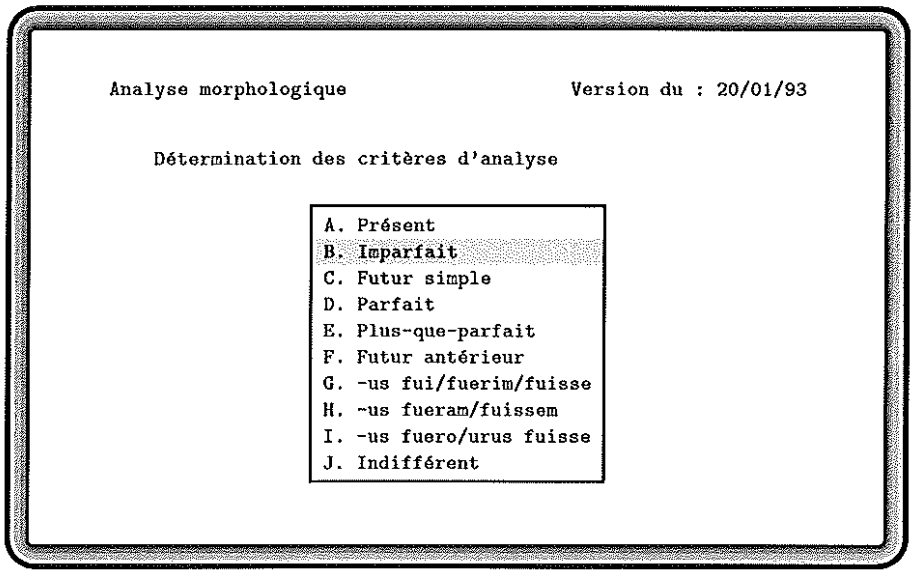


Fig. 5.

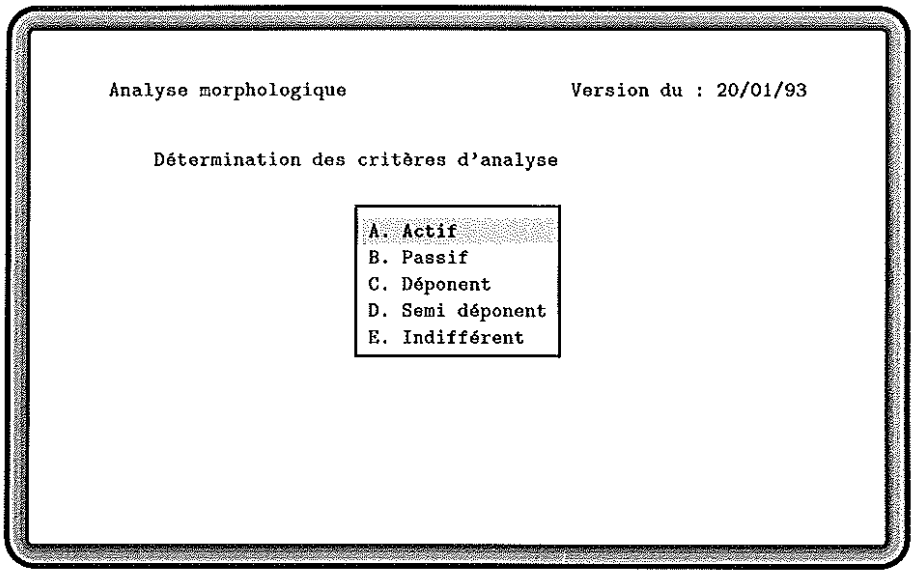


Fig. 6.

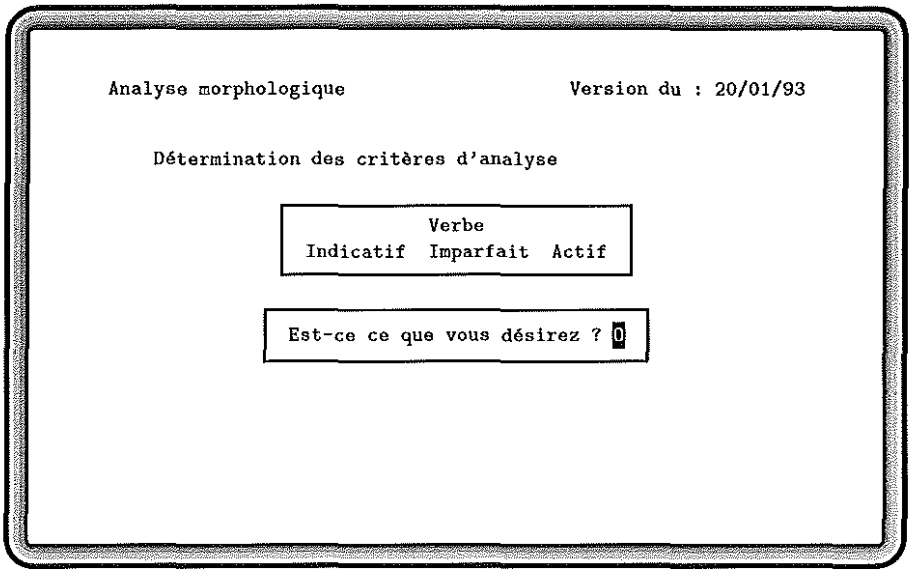


Fig. 7.

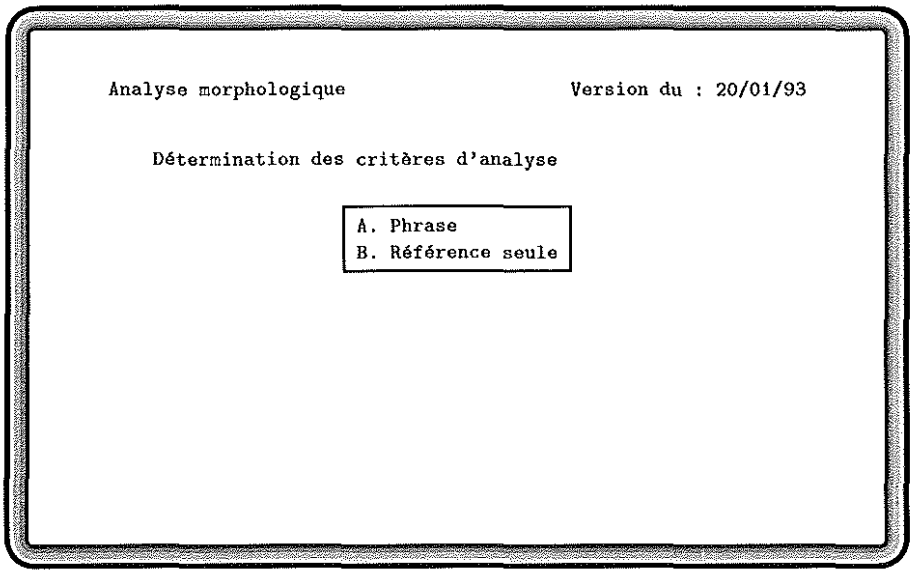


Fig. 8.

On le voit, le principe est celui d'un menu déroulant, offrant un récapitulatif des critères retenus à la fin de la sélection.

La seconde option à l'intérieur du menu « ANALYSE » est celle de la recherche de cooccurrences. Il suffit alors de sélectionner dans la liste proposée le premier élément de la cooccurrence recherchée : ce peut être un lemme, une forme, mais aussi une catégorie grammaticale (Voir § 3.2). L'écran suivant permet, de la même façon et avec la même variété, la détermination du second élément. Enfin, le programme demande si la cooccurrence a pour cadre l'ensemble de la phrase ou si l'on souhaite en limiter l'étendue en imposant une distance maximale mesurée en nombre de mots entre les deux éléments. Nous reviendrons plus loin sur la spécificité de ce menu qui constitue l'une des principales richesses et originalités du logiciel.

3.1.2. L'exploitation statistique

Le menu « STAT(istique) : CONV(ersion) » permet de créer, pour chaque texte, un fichier associé regroupant toutes les données numériques globales trouvées en scannant les codes contenus dans le fichier « TEXTE » ; sont ainsi enregistrés une fois pour toutes le nombre total de substantifs du texte traité et, à l'intérieur de cette catégorie, le nombre d'occurrences de chaque cas ; de même

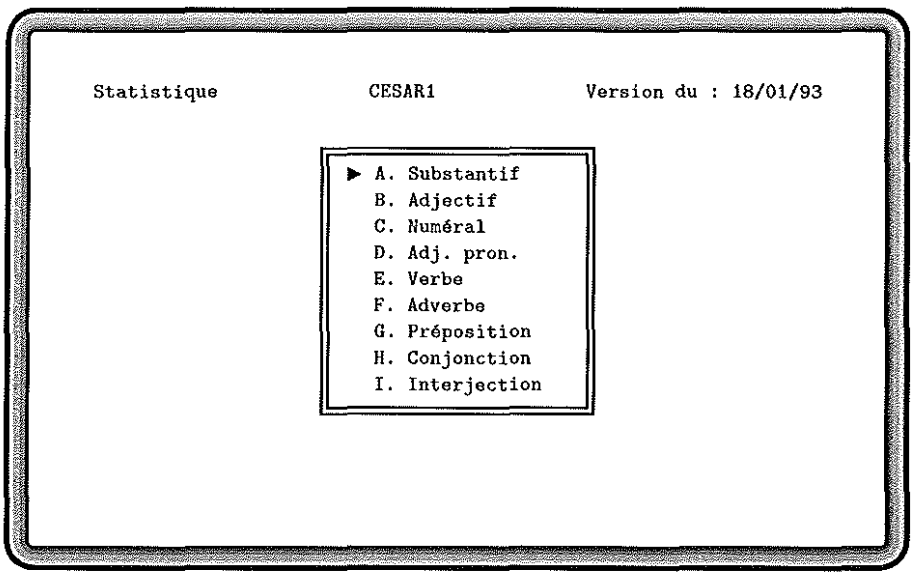


Fig. 9.

pour les verbes et les catégories verbales, etc. Ce fichier numérique récapitulatif peut ensuite être exploité au moyen du menu « EXPLOITE STAT. » ; comme son nom l'indique, ce menu permet de visualiser les données quantitatives sous une forme exploitable, c'est-à-dire sous forme de tableaux sélectifs. Une fois choisi le texte de référence, l'écran (représenté à la figure 9) grâce auquel on sélectionne l'élément de la recherche apparaît.

Par exemple, en sélectionnant le livre I de la *Guerre Civile*, puis la catégorie « ADJECTIF », on obtient le récapitulatif représenté aux figures 10 et 11 où les valeurs numériques sont d'abord données en chiffres absolus, puis en millièmes (simple façon de présenter les pourcentages en évitant les virgules).

Pour le verbe, on peut obtenir un tableau concernant le temps, un autre concernant le mode, un troisième concernant la voix; mais on peut aussi avoir des tableaux à double entrée : temps/mode, mode/voix et temps/voix; les figures 12 et 13 en sont deux illustrations.

3.1.3. La recherche sur préfixe ou suffixe

Il s'agit de retenir dans le fichier des lemmes (.FRQ), les individus qui présentent telle ou telle chaîne de caractères, soit au début soit à la fin du lemme; il n'y a pas de différence de traitement entre les deux recherches et la gestion du menu est extrêmement simple. Le préfixe ou suffixe sélectionné peut comprendre jusqu'à 8 caractères. La liste des lemmes ainsi retenue peut ensuite être présentée dans l'ordre de leur apparition dans le texte ou dans l'ordre alphabétique; dans les deux cas chaque lemme est accompagné des différentes formes qu'il revêt dans le texte et de leurs références.

3.1.4. L'index fréquentiel

Le menu « FREQ(uence) CONV(ersion) » permet de créer et d'associer à chaque texte un index fréquentiel décroissant qui sera ensuite exploité au moyen de la commande « FRÉQUENCES » : celle-ci fait apparaître, pour le texte sélectionné, l'index des lemmes, précédés chacun de leur numéro d'ordre dans ce nouveau classement et de leur fréquence absolue. En déplaçant la surbrillance sur cette liste, on peut en retenir une fraction seulement, par exemple les dix ou les cent premiers lemmes, c'est-à-dire les dix ou cent lemmes les plus fréquents du texte, ou tous les lemmes dont la fréquence est supérieure à une valeur choisie; à l'intérieur de ce sous-ensemble on peut encore opérer une sélection sur la nature grammaticale des mots (première colonne du code) : on conservera ainsi les verbes ou les adjectifs dont la fréquence dans le texte est supérieure à 12 par exemple.

CESAR1 Version du : 18/01/93

Cas	Substantif	Adjectif			
		Positif	Compar.	Superl.	Total
Nominatif	345	39	13	18	70
Vocatif	0	0	0	0	0
Accusatif	886	120	17	28	165
Génitif	366	44	0	1	45
Datif	130	16	0	4	20
Ablatif	774	86	11	16	113
Locatif	3	0	0	0	0
Indéclinable	3	2	0	0	2
Total	2488	307	41	67	415

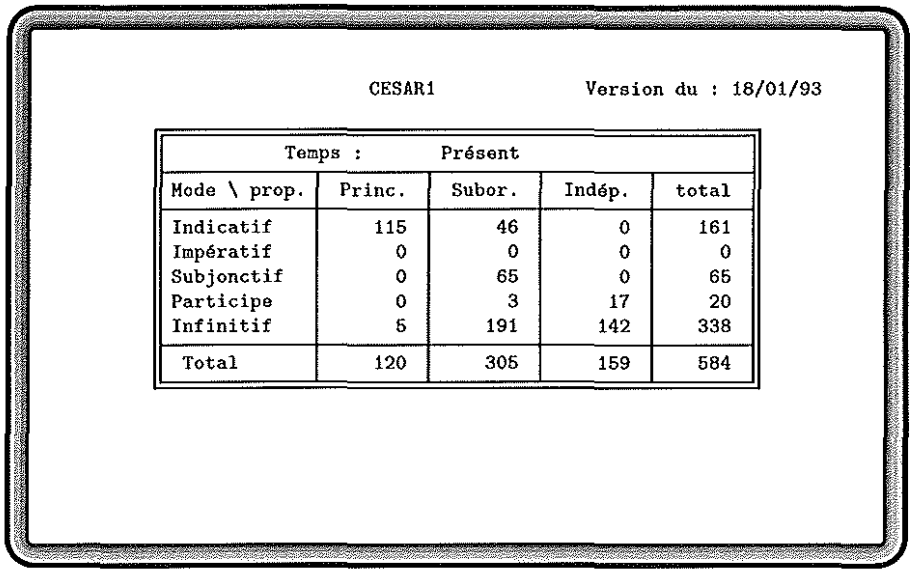
Fig. 10.

CESAR1 Version du : 18/01/93

Cas	Substantif	Adjectif			
		Positif	Compar.	Superl.	Total
Nominatif	139	94	31	43	169
Vocatif	0	0	0	0	0
Accusatif	356	289	41	67	398
Génitif	147	106	0	2	108
Datif	52	39	0	10	48
Ablatif	311	207	27	39	272
Locatif	1	0	0	0	0
Indéclinable	1	5	0	0	5
Total	1007	740	99	161	1000

Valeurs en millièmes

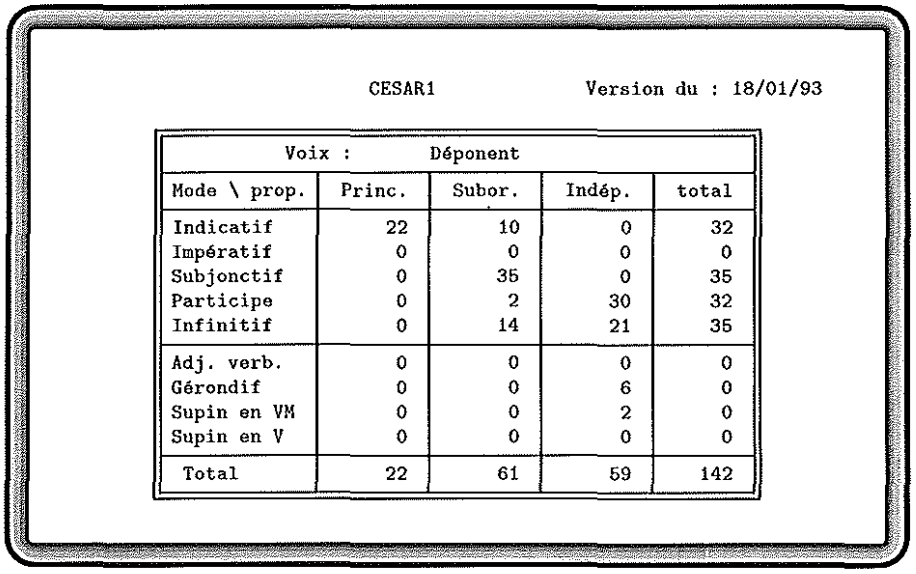
Fig. 11.



CESAR1 Version du : 18/01/93

Temps :	Présent			
Mode \ prop.	Princ.	Subor.	Indép.	total
Indicatif	115	46	0	161
Impératif	0	0	0	0
Subjonctif	0	65	0	65
Participe	0	3	17	20
Infinitif	5	191	142	338
Total	120	305	159	584

Fig. 12.



CESAR1 Version du : 18/01/93

Voix :	Déponent			
Mode \ prop.	Princ.	Subor.	Indép.	total
Indicatif	22	10	0	32
Impératif	0	0	0	0
Subjonctif	0	35	0	35
Participe	0	2	30	32
Infinitif	0	14	21	35
Adj. verb.	0	0	0	0
Gérondif	0	0	6	0
Supin en VM	0	0	2	0
Supin en V	0	0	0	0
Total	22	61	59	142

Fig. 13.

3.2. Principes de la recherche de cooccurrences

C'est à vrai dire cette recherche de cooccurrences qui a motivé la création du logiciel. Souvent en effet le linguiste a une double exigence : il recherchera par exemple volontiers l'association récurrente de *uita* et de *beata* ou celle de *propterea* et de *quod* au sein d'une même phrase. Le logiciel déjà existant de J. Denooz répond en partie à ce besoin : il permet la recherche de cooccurrences entre une forme et un lemme ou entre deux lemmes. Mais le linguiste est souvent personnage capricieux et exigeant : il lui faudra aussi dénombrer les propositions qui contiennent à la fois un impératif et un vocatif; ce dernier n'étant pas trop fréquent, le chercheur peut certes se contenter d'un programme relevant l'ensemble des vocatifs du texte et de trier ensuite lui-même ceux qui accompagnent un impératif. Si, en revanche, il cherche les adjectifs verbaux en *-ndus* noyaux d'un syntagme nominal introduit par la préposition *ad*, sa tâche devient plus rude : à titre d'exemple, on compte 77 occurrences de la préposition *ad* dans le *de Amicitia* de Cicéron et 97 occurrences de l'adjectif verbal, pour seulement 5 syntagmes [*ad* + adj. verbal et nom à l'acc.]. Grâce au logiciel *Estela*, la recherche de la cooccurrence des deux éléments (avec une distance maximale de 4 mots) aboutit à une liste de 6 références : il n'y en a donc qu'une qui ne répond pas à la recherche[7], le tri est rapide ! La même expérience faite sur le *de Ira* de Sénèque donne les résultats suivants : 155 occurrences de la préposition *ad*, 156 occurrences de l'adjectif verbal, mais 11 syntagmes seulement correspondant à la recherche; le menu « COOCCURRENCE » du logiciel fournit 19 références.

On imagine donc aisément le gain de temps que permet un logiciel capable de détecter les cooccurrences, non seulement entre lemmes ou entre forme et lemme, mais aussi entre un lemme (ou une forme) et une catégorie grammaticale (cas, temps verbal, classe de pronoms, type de subordination), voire encore entre deux catégories grammaticales.

Pour utiliser au mieux cette option du menu « ANALYSE » dans *Estela*, il n'est pas inutile de connaître les principes qui ont présidé à l'écriture du programme. Il s'agit de recourir en priorité aux fichiers . FRQ des lemmes et . SUB des codes de subordination, sur lesquels le classement alphabétique rend toute recherche très rapide. Cette priorité impose parfois un ordre de préférence dans la sélection des deux éléments de la cooccurrence.

---

[7] Il s'agit de la séquence *[...] est habendus ad summum [...]* au paragraphe 83.

3.2.1. Cooccurrence de deux lemmes, de deux types de subordination ou d'un lemme et d'un type de subordination

Le programme commence par constituer une pile de références à partir des phrases comprenant le premier élément de la cooccurrence; il fait ensuite une seconde pile avec les références de phrases comprenant le second élément, puis compare les deux piles ainsi constituées et ne retient que l'intersection des deux ensembles. Si l'élément servant de critère de sélection est un lemme, la recherche aboutissant à la création de la pile correspondante se fait dans le fichier .FRQ; si cet élément est un code de subordination, la recherche se fait dans le fichier .SUB. Dans les deux cas il s'agit d'un balayage assez rapide sur un fichier organisé de façon alphabétique et l'ordre d'introduction des deux éléments de la cooccurrence n'a ici guère d'importance.

En dernier lieu, le programme vérifie que la distance imposée entre les deux éléments est respectée; cela est possible sans retour au texte dans la mesure où, dans les deux fichiers .FRQ et .SUB, chaque forme est accompagnée de son numéro d'ordre dans le texte.

3.2.2. Cooccurrence d'un lemme ou d'un code de subordination avec une forme ou avec une analyse morphologique

Cette fois-ci, il est impératif d'observer l'ordre suivant dans la spécification des éléments de la cooccurrence : il convient de choisir comme premier élément le lemme ou le code de subordination. La procédure consiste en effet à constituer d'abord, à partir du fichier .FRQ ou .SUB, une pile des références de phrases répondant au premier critère de sélection retenu; ensuite, c'est en se reportant au fichier « TEXTE » que chacune de ces phrases sera examinée et que le programme vérifiera si elle contient ou non le second élément de la cooccurrence.

Cette procédure a été choisie pour des raisons d'efficacité évidentes : la recherche d'une forme ou d'une analyse morphologique se faisant par lecture au fil du texte, elle prend — on l'a déjà souligné — davantage de temps que la recherche sur les deux autres fichiers qui regroupent sous une seule et même entrée les diverses occurrences du même code; on cherche donc à abréger cette lecture cursive du texte en la réduisant à la lecture des quelques phrases sélectionnées en fonction de l'autre critère.

3.2.3. Cooccurrence de deux formes, d'une forme et d'une analyse morphologique ou de deux analyses

Il est, dans un tel cas de figure, impossible d'avoir recours à l'un des deux fichiers classés par ordre alphabétique : la lecture du fichier « TEXTE »

s'impose. La procédure est alors la suivante : le programme scanne le texte en cherchant la première phrase contenant le premier élément sélectionné pour la cooccurrence; dès qu'il l'a trouvée, et avant de passer à la suivante, il regarde si le second élément de la cooccurrence apparaît aussi dans cette phrase, et à la distance voulue du premier. Si ce second critère de sélection est satisfait, la phrase est conservée; sinon, elle est rejetée; puis la lecture continue selon les mêmes principes, c'est-à-dire en s'arrêtant à la première phrase suivante qui contient le premier élément de la cooccurrence. On voit donc qu'il est avantageux de retenir comme premier élément de la cooccurrence celui qui a des chances d'être le moins fréquent dans le texte : à supposer qu'il soit intéressant de chercher les cooccurrences de la forme *sunt* et d'une interjection dans le *de Rerum Natura* de Lucrèce, le programme s'arrêtera sur 179 phrases si on lui donne la forme *sunt* comme premier critère, sur 23 phrases seulement si c'est la catégorie « interjection » qui est d'abord sélectionnée.

À condition de respecter ces quelques conseils d'utilisation, le menu de recherche de cooccurrences s'avère très efficace; c'est à nos yeux le principal atout de ce logiciel qui, par ailleurs, offre l'intérêt d'associer au sein d'un même menu général plusieurs types de recherches en exploitant au mieux les trois fichiers constitutifs d'un texte enregistré et lemmatisé dans la banque de données du L.A.S.L.A. On espère que cette description des logiciels associés à la banque de données aura convaincu le lecteur de l'importance et de la validité du travail effectué par le Laboratoire de Liège.