

***Opera Latina :***  
**la banque de données de textes latins**  
**du L.A.S.L.A. sur CD-ROM**

Gérald PURNELLE

**Abstract.** University of Liège's L.A.S.L.A. is about to publish on CD-ROM its own database of lemmatized and analyzed Latin texts. The CD-ROM includes 77 works by 18 different authors. Its contents are described in this paper, as well as the computer program allowing its consultation. The program features such functions as corpus definition, simple or combined search (based on cooccurrence), result exploitation, access to a work outside search and on-screen display of statistics. Possible search objects are form, lemma, the syntax of subordinate verbs and the whole range of morphological criteria.

**Keywords:** Lemmatized Latin Texts, CD-ROM, **Mots-clés :** Textes latins lemmatisés, CD-ROM, latin literature, littérature latine.

Dans le précédent volume de cette *Revue*, Sylvie Mellet a présenté deux logiciels d'exploitation de la base de textes latins du L.A.S.L.A.<sup>1</sup> Dans cet article, où sont décrits les deux systèmes qui, jusqu'alors, avaient été développés pour cette exploitation, elle annonçait la publication prochaine de ladite banque de données sur support CD-ROM<sup>2</sup>.

L'objet du présent article sera de décrire le contenu de ce CD-ROM — baptisé *Opera Latina*, et dont la parution est imminente — et surtout l'interface d'interrogation qui l'accompagne. Ce nouvel instrument de recherche,

---

<sup>1</sup> Cf. MELLET (S.) : 1994, « Logiciels d'exploitation de la banque de données de textes latins du L.A.S.L.A. », *Revue. Informatique et statistique dans les sciences humaines*, 30, pp. 91–108.

<sup>2</sup> *Ibid.*, p. 95.

mis à la disposition de la communauté scientifique, sera amené à lui rendre de grands services dans différents domaines tels que la linguistique et la philologie.

### La lemmatisation et l'analyse morphologique

Avant de décrire en détail le contenu du CD-ROM, la nature des données qui y sont rassemblées et les possibilités de recherche qu'il offre, il n'est pas besoin de rappeler longuement les principes qui président à la production de notre banque de données. Le lecteur se souviendra facilement que depuis plus de 30 ans le L.A.S.L.A. applique aux textes latins des procédures semi-automatiques de lemmatisation et d'analyse morphologique, et que celles-ci permettent la constitution de fichiers-textes dans lesquels sont conservées les formes du texte, à raison d'une forme par enregistrement, accompagnée de toutes les informations qui lui sont attachées : son lemme, son analyse morphologique complète et le codage de son éventuelle subordination (dans le cas des verbes).

Le concept de lemmatisation ne nécessite plus, lui non plus, d'abondantes précisions : rappelons que « lemmatiser » un texte consiste à rapporter chaque mot de ce texte à son lemme, c'est-à-dire à la forme qui le représente dans un dictionnaire de référence. Ainsi, le lemme de *patrem* sera *pater*, celui de *feci et facta* sera *facio*, celui de *imperatorum* sera, selon les cas, *imperare* ou *imperator*. Dans son article, Sylvie Mellet a détaillé les atouts de la lemmatisation, particulièrement importants dans le cas de la langue latine<sup>3</sup>. Je me bornerai ici à rappeler que la lemmatisation permet de :

- regrouper (et donc rechercher) toutes les formes d'un mot, sans devoir les énumérer, malgré une éventuelle variation graphique ou formelle parfois très importante (p. ex. *fer-*, *tul-*, *lat-* pour *FERO*);
- éliminer les formes homographes non pertinentes (la distinction des lemmes *LEX* et *LEGO* permet de distinguer parmi les formes *legis*, *legi* et *lege* celles qui appartiennent à l'un ou à l'autre);
- résoudre les problèmes de variations graphiques (sous le lemme *AF-FIRMO* sont rangées toutes les formes en *adfirm-* et en *affirm-*);
- distinguer les lemmes homographes, qu'il s'agisse de mots étymologiquement différents (*LABOR* substantif ou verbe) ou d'emplois différents

<sup>3</sup> *Ibid.*, pp. 93-94.

(*QVANDO* adverbe relatif, adverbe interrogatif, conjonction de subordination ou adverbe indéfini).

Une fois qu'un texte a été traité au moyen de ces méthodes, le fichier qu'il constitue peut être utilisé de diverses façons. On peut, notamment, produire de manière automatique et publier un *index verborum* de l'œuvre. C'est ce que fait le L.A.S.L.A. depuis ses débuts, pour la plupart des textes auxquels il s'est intéressé. On peut aussi produire une concordance lemmatisée. Ce sont là les usages les plus visibles d'une banque de données comme la nôtre, ceux auxquels un utilisateur songera en premier lieu. Mais, on le sait, les fichiers du L.A.S.L.A. peuvent encore être utilisés à d'autres fins. Les logiciels décrits par S. Mellet ont pour fonction d'« interroger » la banque de données selon un nombre important de points de vue. Il est ainsi possible de la consulter pour y chercher toutes les occurrences d'un lemme (un lexème), d'une catégorie grammaticale, d'un cas, d'un mode, d'une combinaison de critères grammaticaux, d'un type précis de subordination.

## Le CD-ROM

À la faveur des récents développements de la micro-informatique et considérant l'ampleur de la collection de textes ainsi constituée, le L.A.S.L.A. a résolu d'améliorer significativement la façon dont il met cette information à la disposition de la communauté scientifique, en la diffusant sur un support informatique moderne.

Le corpus enregistré sur le disque comprend 1 276 023 mots; il couvre 77 œuvres de 18 auteurs différents (on se reportera à la liste placée en annexe). Le plus ancien est Caton, dont les œuvres sont datées du début du II<sup>e</sup> siècle av. J.-C. Les plus récents sont Tacite, Juvénal et Pline le Jeune, dont les œuvres furent écrites à la fin du I<sup>er</sup> et au début du II<sup>e</sup> s. de notre ère<sup>4</sup>. Certains auteurs sont intégralement repris dans le corpus : ainsi César, Caton, Catulle, Quinte-Curce, Horace, Juvénal, Lucrèce, Pétrone, Tibulle, Virgile. D'autres y figurent presque entièrement : de Sénèque ne nous manquent que les *Questions naturelles*, de Salluste les fragments, de Tacite les *Histoires*, de Properce le dernier des quatre livres. Certains sont certes amplement présents, mais il reste encore une bonne part de leur œuvre à traiter : ainsi Cicéron ou Ovide. Deux auteurs, Ausone et Pline le Jeune, ne sont représentés que par

---

<sup>4</sup> En réalité, un auteur plus récent est présent : il s'agit d'Ausone, qui date du IV<sup>e</sup> s., mais qui n'est représenté dans notre banque que par ses seules *Épigrammes*.

une faible partie de leur œuvre. Enfin on note la présence de quatre textes anonymes, les trois guerres césariennes et la tragédie d'*Octavie*.

La banque de données est constituée par l'ensemble des formes de chacun de ces textes. Le CD-ROM contient essentiellement le corpus complet des textes et un ensemble d'index fondés sur les différentes informations liées à chaque mot du texte (lemme, forme, catégorie grammaticale, sous-catégorie (déclinaison, conjugaison, etc.), cas, nombre, mode, temps, voix, personne et type de subordination); il contient également un logiciel d'interrogation assez performant. Avant d'en exposer toutes les fonctionnalités, il est intéressant de signaler, dès à présent, que ce logiciel sera indifféremment utilisable sous système *Windows* (y compris *Windows 95*) ou *Macintosh*. Cette qualité est propre à satisfaire tous les utilisateurs.

### **Le programme de consultation**

L'écran proposé présente sept feuillets différents, dans lesquels l'utilisateur formule ses choix. Ils s'intitulent « Corpus », « Statistiques », « Recherche 1 », « Recherche 2 », « Recherche 3 », « Résultats » et « Œuvre ».

Dans le premier, l'utilisateur définit son corpus de recherche, c'est-à-dire l'ensemble des œuvres dans lequel il entend effectuer une recherche. Toutes les formules de choix sont possibles : l'ensemble du corpus, un seul auteur, plusieurs œuvres d'un auteur, une seule œuvre, plusieurs auteurs, en totalité ou en partie. Un corpus ainsi défini peut être sauvegardé sous la forme d'un fichier.

Les trois feuillets « Recherche » permettent de définir un ou plusieurs objets de recherche (un par feuillet). Dans chacun, l'écran propose plusieurs champs, correspondant aux différents types d'objets que contient la banque de données : lemme, forme textuelle, syntaxe des propositions, critères morphologiques.

La zone de saisie « Lemme » est couplée avec un menu déroulant qui contient la liste alphabétique des lemmes attestés dans la banque de données. Au fur et à mesure qu'un mot est dactylographié dans la zone, l'affichage de cette liste se positionne progressivement sur le premier lemme commençant par les caractères frappés, jusqu'à ce qu'un lemme complet soit atteint.

À ce stade interviennent plusieurs particularités. Si le lemme demandé possède des homographes, ceux-ci sont affichés, et l'utilisateur peut soit sélectionner l'un d'eux, soit choisir de faire porter la recherche sur l'ensemble des lemmes homographes. Ainsi, si l'on a tapé *LABOR*, on choisira plus utilement

entre « substantif » et « verbe » ; à l'inverse, en ayant frappé *CVR*, on pourra sélectionner les deux analyses (adverbe relatif et adverbe interrogatif) ou n'en choisir qu'une.

Il est par ailleurs loisible à l'utilisateur d'affecter d'une troncature la chaîne de caractères dactylographiée dans la zone « Lemme », aussi bien à son début (p. ex. *\*ITIO*) qu'à sa fin (p. ex. *COMMVN\**). On voit que cette option permet de rechercher en une seule opération toutes les occurrences de lemmes s'achevant ou commençant par les mêmes lettres.

La zone intitulée « Forme » s'utilise de la même manière : un menu déroulant donne un accès progressif à la liste des formes attestées dans la banque ; quand deux ou plusieurs formes sont homographes, la liste des différents lemmes auxquels elles appartiennent s'affiche et l'on peut éventuellement sélectionner le lemme pertinent (ou, à l'inverse, assumer la recherche de toutes les occurrences de la forme, quels que soient leurs lemmes) ; les options de troncatures gauche et droite sont disponibles pour la forme également.

Si l'objet à rechercher est une syntaxe des propositions, le résultat de la recherche sera constitué de tous les verbes subordonnés soumis au type de subordination demandé (p. ex. ablatifs absolus, propositions infinitives, verbes introduits par *QVI* relatif, par *DVM*, par *VT*, etc.). En regard de la zone correspondante dans un feuillet se déroule un menu qui contient la liste complète des types de subordinations attestés dans la banque, il suffit d'y choisir celui que l'on souhaite rechercher.

Le quatrième type d'objet est la morphologie. L'usage de cette partie du feuillet est double : les critères morphologiques peuvent être recherchés isolément ou être utilisés comme précision d'une recherche portant sur un des trois autres types (lemme, forme ou syntaxe des propositions).

À l'écran figurent neuf zones, qui correspondent au neuf critères morphologiques enregistrés dans notre banque : catégorie grammaticale, sous-catégorie, cas, nombre, degré, voix, mode, temps, personne. On sélectionnera dans une ou plusieurs de ces zones les critères que l'on veut rechercher (éventuellement en combinaison) dans la banque ; p. ex. : accusatif pluriel (2 critères) ; substantif, première déclinaison, datif pluriel (4) ; adverbe au comparatif (2) ; indicatif présent déponent 2<sup>e</sup> personne singulier (5). Le résultat de telles recherches sera constitué de toutes les formes qui présentent, pour les critères recherchés, les valeurs choisies, quels que soient leurs lemmes et leurs formes.

Indépendamment de ce recours à la morphologie comme objet de recherche, il est possible d'utiliser les mêmes menus pour limiter une recherche

portant sur un objet d'un autre type (lemme, forme ou syntaxe). Un objet ayant été choisi, on pourra restreindre la portée de la recherche en sélectionnant des critères grammaticaux à fonction restrictive. Ainsi d'un lemme *LOCVS*, on ne retiendra que les formes au pluriel; d'un lemme *FACIO*, les formes au subjonctif imparfait; d'une forme *amicorum*, les occurrences de substantif; d'une fin de forme *\*im*, les occurrences comme accusatif singulier; d'une fin de lemme *\*iter*, les occurrences d'adverbe; des verbes introduits par *DVM*, les formes de verbes au subjonctif; etc.

L'utilisateur a donc le loisir de définir ainsi jusqu'à trois recherches (une par feuillet) en même temps. Chaque recherche peut être effectuée indépendamment des autres. Par ailleurs, dès qu'au moins deux recherches sont définies, on peut soit les combiner, soit les cumuler. Cumuler deux ou trois recherches revient à les effectuer en une seule commande, ce qui a pour effet de produire une seule liste de résultats cumulés. L'exécution d'une recherche combinée est plus complexe : il s'agit de rechercher les contextes où les deux ou trois objets définis apparaissent en cooccurrence.

Dans le deuxième feuillet se trouvent plusieurs zones qui permettent de définir les conditions de cooccurrence des objets recherchés. On peut décider : qu'ils doivent obligatoirement figurer dans la même phrase, ou dans le même vers; qu'ils ne peuvent être distants l'un de l'autre que d'un nombre maximal de mots; que les objets doivent obligatoirement apparaître en contexte dans l'ordre de leur définition (feuillet « Recherche » 1, 2 et 3).

Trois exemples suffiront à illustrer l'utilité des recherches combinées : « lemme *IMPERO* » et « verbes introduits par *VT* dans la même phrase »; « lemme *PHILOSOPHIA* » et « lemme *SAPIENS* » à une distance de 50 mots maximum; « interjection », « vocatif » et « impératif » dans la même phrase.

Une fois qu'une recherche, éventuellement cumulée ou combinée, a été effectuée, ses résultats sont affichés sous la forme d'une liste verticale formant menu; chaque occurrence trouvée occupe une ligne, où sont mentionnés : la forme textuelle, son lemme, la référence complète (auteur, œuvre, livre ou chant, chapitre, paragraphe, ligne ou vers). L'utilisateur parcourt ce menu et, à tout moment, choisit l'occurrence dont il souhaite lire le contexte. Celui-ci apparaît en pleine page; le mot concerné est affiché en gras. L'utilisateur a le loisir, à partir de ce point, de parcourir l'œuvre entière; il peut afficher à la demande, pour n'importe quel mot, toutes les informations qui le concernent : son lemme, une explication éventuelle liée à sa nature d'homonyme, son analyse morphologique complète, sa fonction et son type de subordination s'il s'agit d'un verbe. L'utilisateur dispose donc de tous les moyens nécessaires

pour comprendre en profondeur le sens de la phrase (exception faite, partiellement, de la syntaxe).

L'interface présente en outre les traditionnelles commandes d'exportation (vers traitement de textes) et d'impression; il n'est ni possible ni permis d'exporter une portion de texte d'une longueur supérieure à celle de la phrase où figure le mot trouvé.

Telles sont donc les trois grandes étapes d'une requête appliquée à la banque de données : définition de l'objet, définition du champ de recherche, exploitation des résultats. Il reste à décrire les deux dernières fonctionnalités du logiciel.

Un feuillet intitulé « Statistiques » permet d'afficher, sans effectuer de recherche, la fréquence de certains phénomènes dans la banque. On y trouve deux zones « Lemme » et « Forme », ainsi que la liste des auteurs et de leurs œuvres. Si l'on choisit l'option « Lemme » sans formuler un lemme précis, le programme affiche, en regard de chaque auteur et en regard de chaque œuvre, le nombre de lemmes *différents* qui y figurent. Si l'option « Forme » est choisie, les chiffres affichés représentent le nombre total d'occurrences contenues dans l'auteur ou dans l'œuvre, c'est-à-dire sa longueur en mots.

Les deux zones « Lemme » et « Forme » fonctionnent de la même manière que dans les feuillets « Recherche » : la frappe d'une chaîne de caractères permet de progresser dans une liste alphabétique, d'atteindre un lemme (ou une forme) précis et de le sélectionner en ne retenant, éventuellement, qu'un homographe. Dès qu'un tel choix est fait, la fréquence du phénomène défini (lemme ou forme) est affichée en regard des auteurs et des formes. Un total figure en bas de l'écran.

La dernière fonctionnalité (feuillet « Œuvre ») concerne la consultation d'une œuvre en plein écran, indépendamment de toute recherche. Il suffit de sélectionner l'œuvre désirée; un système de menu arborescent permet d'atteindre progressivement la partie que l'on veut (p. ex. le livre, puis le chapitre, puis le paragraphe). Dans ces menus, chaque partie est illustrée par ses premiers mots. Une fois le choix effectué, le texte est affiché en pleine page; l'utilisateur peut s'y déplacer sans peine; la même possibilité que dans la consultation de résultats est offerte : il suffit de cliquer sur un mot pour faire apparaître toutes les informations qui y sont associées (lemme et analyse complète).

Ainsi se présentent le CD-ROM du L.A.S.L.A. et l'interface d'exploitation qui l'accompagne. On devine combien il sont à même de rendre service au philologue, à l'historien, à l'enseignant, au lexicographe, au grammairien, au linguiste.

Il reste à donner quelques informations pratiques touchant la parution prochaine du CD-ROM. Il a été développé, sera produit et diffusé par les Éditions Hachette; il doit paraître au mois de décembre 1995. Le CD-ROM et le logiciel seront indifféremment utilisables sur les deux plates-formes *Windows* et *Macintosh*. La banque de données fera l'objet de mises à jour vraisemblablement annuelles. Il est d'ores et déjà possible d'annoncer que dans la version suivante de la banque seront ajoutés sept ou huit pièces de Plaute, de nouveaux discours de Cicéron, la fin de l'œuvre de Properce et de Salluste. En outre, des tableaux statistiques seront rendus accessibles; ils seront relatifs aux différents phénomènes linguistiques contenus dans la banque (fréquence, par œuvre, des différents critères grammaticaux, etc.). En ce qui concerne les développements ultérieurs, nous prévoyons de poursuivre et d'achever la lemmatisation d'Ovide, de Sénèque et des discours de Cicéron, d'intégrer Suétone et de commencer Varron et Térence.



## Annexe

**Liste des auteurs et des œuvres contenus  
dans la banque de données du L.A.S.L.A.**

Decimus Magnus AUSONIUS	<i>Epigrammata</i>
Gaius Iulius CAESAR	<i>Commentarii belli Gallici</i> <i>Commentariorum belli Gallici liber octavus</i>
	[A. HIRTIVS]
	<i>Commentarii belli civilis</i> <i>De bello Africano</i> [apocryphum] <i>De bello Alexandrino</i> [apocryphum] <i>De bello Hispaniensi</i> [apocryphum]
Marcus Porcius CATO	<i>De Agricultura</i> <i>Origines</i> <i>Orationum fragmenta</i> <i>Fragmenta reliqua</i>
Gaius Valerius CATULLUS	<i>Carmina</i>
Marcus Tullius CICERO	<i>Pro A. Caecina oratio</i> <i>Pro A. Cluentio oratio</i> <i>Divinatio in Q. Caecilium</i> <i>Pro M. Fonteio oratio</i> <i>De Imperio Cn. Pompei ad Quirites oratio</i> <i>Pro P. Quinctio oratio</i> <i>Pro S. Roscio Amerino oratio</i> <i>Pro Q. Roscio Comoedo oratio</i> <i>Pro M. Tullio oratio</i> <i>In C. Verrem actiones prima et secunda</i> <i>De Amicitia</i> <i>De Officiis</i> <i>De Senectute</i>
Q. CURTIUS Rufus	<i>Historiae Alexandri Magni</i>
Q. HORATIUS Flaccus	<i>Ars Poetica</i> <i>Carmen saeculare</i> <i>Carmina</i> <i>Epistulae</i> <i>Epodi</i> <i>Sermones</i>
D. Iunius IUVENALIS	<i>Saturae</i>
T. LUCRETIVS Carus	<i>De Rerum natura</i>
P. OVIDIVS Naso	<i>Amores</i> <i>Ars amatoria</i> <i>Epistulae Heroidum</i>

	<i>Medicamina faciei femineae</i>
	<i>Remedia anoris</i>
PETRONIUS Arbiter	<i>Satiricon</i>
C. PLINIUS Caecilius Secundus	<i>Panegyricus</i>
Sextus PROPERTIUS	<i>Elegiae [tres libri]</i>
C. SALLUSTIUS Crispus	<i>Catilina</i>
	<i>Jugurtha</i>
Lucius Annaeus SENECA	<i>Apocolocyntosis</i>
	<i>De Beneficiis</i>
	<i>De Brevitate vitae</i>
	<i>De Clementia</i>
	<i>De Constantia</i>
	<i>Ad Helviam matrem de consolatione</i>
	<i>De Ira</i>
	<i>Ad Lucilium epistulae morales</i>
	<i>Ad Marciam de consolatione</i>
	<i>De Otio</i>
	<i>Ad Polybium de consolatione</i>
	<i>De Providentia</i>
	<i>De Tranquillitate animi</i>
	<i>De Vita beata</i>
	<i>Agamemnon</i>
	<i>Hercules furens</i>
	<i>Hercules Oetaeus</i>
	<i>Medea</i>
	<i>Oedipus</i>
	<i>Phaedra</i>
	<i>Phoenissae</i>
	<i>Thyestes</i>
	<i>Troades</i>
	<i>Octavia [apocryphum]</i>
Cornelius TACITUS	<i>Agricola</i>
	<i>Annales</i>
	<i>Dialogus de oratoribus</i>
	<i>Germania</i>
Albius TIBULLUS	<i>Elegiae</i>
Publius VERGILIUS Maro	<i>Aeneis</i>
	<i>Eclogae</i>
	<i>Georgica</i>