Littérature latine et banques de données : *ne quid nimis*

Joseph Denooz*

Summary : The number of available digital versions of ancient Latin texts has increased in almost excessive proportions. These versions are recorded either on cd-roms (PHI3, BTL, Hyperbas-Lasla-Nice) or (more and more frequently) on websites (e.g. *The Latin Library*, Intratext*, Itinera electronica*,…). It can be observed that most of these digitalised texts are actually not data banks but at the outmost text-bases that do not include any critical, lexicological, morphological, or syntactical information. Moreover these bases are often copied and pasted from each other, thus reproducing the same files with their qualities and flaws. The aim of the present article is to provide a description of what is available and to hoghlight both positive and negative aspects.

Keywords : littérature latine – banques de données textuelles – qualités – défauts.

Les œuvres littéraires latines antérieures à la chute de l'Empire romain copiées et recopiées au fil des siècles par les moines, les clercs et les lettrés font l'objet, depuis l'invention de l'ordinateur et son utilisation en linguistique, de transcriptions numérisées dont on établirait difficilement un inventaire exhaustif.

Les premières applications aux textes latins antiques ont été réalisées au Laboratoire d'Analyse statistiques des Langues anciennes de l'Université de Liège[1], fondé en 1961.

À partir de cette date de nombreux projets vont voir le jour dans d'autres centres de recherche mais sans aboutir au développement de bases de données ; il s'agit le plus souvent de programmes dont le but est de produire des instruments de travail consacrés à un auteur, voire à une seule œuvre[2].

Cet article ne prend pas en compte ces travaux ponctuels ; il s'attache aux seules banques de données qui ont une certaine amplitude et une large diffusion.

Pour procéder à une analyse critique de ces banques textuelles et en évaluer l'utilité pour le chercheur ainsi que la qualité, nous avons sélectionné les principales d'entre elles en citant d'abord celles qui contiennent le plus d'œuvres latines. Nous avons préféré laisser de côté des bases dédiées à un auteur particulier ; leur examen n'aurait rien apporté à cette analyse.

Les bases de données textuelles diffusées sur CD-Rom ou par le WEB ne sont en général que des copies des reproductions anastatiques d'éditions existantes, parfois assez anciennes.

La première est celle qui a été réalisée sur CD-Rom par la *Packard Humanities Institute* (PHI)[3], elle est diffusée dans une version qui date de 1991 et peut, depuis quelque temps, être obtenue

* Joseph.Denooz@ulg.ac.be

1 En abrégé L.A.S.L.A.

2 Il s'agit de travaux tels que *Aurelius Victor* : concordance réalisée par le G.I.T.A. [Groupe d'informatique et de traitement automatique – Université Libre de Bruxelles], U.L.B. : Ghislaine Viré et René Patesson, *De Casesaribus*, s.d. ou encore de Bengt M. Löfstedt, David W. Packard, *A Concordance of the Sermons of Bishop Zeno of Verona*, American Philological Association, 1975.

3 Le *Packard Humanities Institute* (PHI) a été fondé en 1987 pour développer des instruments de travail dans les domaines de l'histoire, de la littérature, de la musique, etc. Le PHI s'est substitué à la fondation David and Lucile Packard dont les activités touchaient ces domaines.

gratuitement. On examinera ensuite la *Bibliotheca Teubneriana Latina* (BTL), diffusée sur CD-Rom, elle contient la majorité des textes publiés aux éditions Teubner depuis les textes latins les plus anciens jusqu'à une époque récente[4].

À côté des supports CD-Rom ou DVD, on trouve sur la toile de nombreux sites Web consacrés aux principales œuvres de la littérature latine ou à un ou plusieurs auteurs.

Le premier site à voir est *The Latin Library*[5] ; on examinera ensuite *Itinera electronica*[6] organisé par l'Université catholique de Louvain (Belgique) et en troisième lieu, le site de « L'antiquité grecque et latine »[7] coordonné par plusieurs professeurs belges de langues anciennes. D'autres sites, moins complets seront rapidement passés en revue ; on retiendra entre autres les bases intitulées *Nimis pauci*[8] et *Intratext*[9] qui comprennent quelques auteurs latins, des auteurs grecs, etc.

Enfin, on citera, pour mémoire, l'un ou l'autre site consacré à un auteur particulier, tel Cicéron pour lequel on se réfère notamment soit à un site de l'Université du Texas[10], soit à des éditions plus ou moins anciennes ; ainsi pour les discours, on trouve une édition du XVI^e^ siècle[11] sur le site Gallica et ailleurs une du XIX^e^ siècle[12]. Il s'agit fréquemment de textes numérisés en format PDF. On trouve encore des sites consacrés à des genres littéraires particuliers. C'est le cas sur le serveur de l'Université de Nice pour la philosophie[13].

Toutes ces bases textuelles présentent un inconvénient majeur : elles contiennent uniquement le texte et des éléments de référence. N'y figurent jamais de données lexicales, morphologiques ou syntaxiques. Dès lors, elles n'ont pas de contenu linguistique ou philologique, ce qui dans une langue comme le latin génère de sérieuses difficultés – il en va de même pour le grec – : la recherche d'un mot conduira dans de nombreux cas à repérer des formes auxquelles on ne s'intéresse pas (quelques exemples entre mille : *ducis* venant de *dux* ou de *duco*, *legis* de *lex* ou de *lego*, *tempus* signifiant « le temps » ou « la tempe », *occidit* composé de *cado* ou de *caedo*, …).

A.- Banques de données sur CD-Rom

A.1 – PHI CD #5.3

La version actuelle du PHI contient l'ensemble de la littérature latine depuis les origines jusqu'à la fin du II^e^ siècle de notre ère, ainsi que quelques auteurs postérieurs, tels Servius, Justinien, Zénon de Vérone *et alii* ; en outre les concepteurs du PHI ont fait figurer sur le CD le texte de la *Bible*. Les éditions choisies, souvent de bonne qualité, ne sont pas toujours très récentes. Ainsi, pour Plaute, l'édition de référence est celle de Friedrich Leo qui date de 1895[14].

D'une manière générale, la transcription des textes est correcte et fidèle à l'édition de référence.

---

4 Publiée chez Brepols.
5 http://www.thelatinlibrary.com/
6 Le site web d'*Itinera electronica* est accessible à partir de plusieurs adresses ; celle qui donne accès le plus directement aux textes latins est : http://agoraclass.fltr.ucl.ac.be/concordances/intro.htm.
7 http://remacle.org/
8 http://ugo.bratelli.free.fr/
9 http://www.intratext.com/LAT/
10 http://www.utexas.edu/depts/classics/documents/Cic.html#Texts.
11 *M. T. Ciceronis Orationum* , éd. de : Parisiis : apud Simonem Colinaeum, 1532.
12 *Œuvres complètes de Cicéron*, traduites en français, texte en regard, Paris, Fournier, 1817.
13 http://www.ac-nice.fr/philo/textes/biblio.htm
14 *T. Maccius Plautus. Plauti Comoediae.* Friedrich Leo, Berlin, Weidmann, 1895 (2 t.)

Parmi les avantages que présente le Cd-Rom du PHI, on retiendra que le progiciel d'exploitation et de recherche est *Silver Mountain Software*, logiciel utilisé aussi pour les recherches dans le *Thesaurus Linguae Graecae* (TLG), ce qui implique que l'apprentissage d'un seul système permet de travailler indifféremment sur les œuvres grecques et latines.

Bien accueilli par les latinistes, le CD-Rom PHI#5.3, très répandu, n'exploite pas suffisamment les possibilités actuelles des ordinateurs et ne répond qu'imparfaitement aux exigences et aux besoins des chercheurs. On relève des problèmes qui sont liés à la conception déjà ancienne de cette base de textes. Quoi qu'il en soit, le PHI permet de faire des recherches simples ou complexes (sur base de l'algèbre booléenne) soit sur l'ensemble du corpus, soit sur un auteur, soit encore sur une seule œuvre. En outre, l'utilisateur peut effectuer des recherches ponctuelles (un ou plusieurs mots – en succession immédiate ou séparés par un nombre maximum de mots que l'on indique), il peut constituer des concordances, des index, des listes inverses ou aussi des listes avec indications statistiques.

Malgré les possibilités qu'offre le PHI, il faut, pour l'utiliser efficacement, tenir compte de plusieurs défauts qui peuvent conduire à des résultats erronés.

En premier lieu, les textes sont enregistrés dans une forme que l'on a appelée le « Bêta code » qui s'inspire des normes adoptées pour le TLG. Voici, à titre d'exemple, le début du *Pro Quinctio*.

> ï€°´·´ÿï °°±ÿï,Ñõéîãôÿï *f* ÃéãÿŸôÿ@@@@{1PRO P. QVINCTIO ORATIO}1 '@@Quae res in civitate duae plurimum possunt, eae contra €nos ambae faciunt in hoc tempore, summa gratia et elo-€quentia; quarum alteram, C. Aquili, vereor, alteram metuo. €Eloquentia Q. Hortensi ne me in dicendo impediat, non €nihil commoveor, gratia Sex. Naevi ne P. Quinctio noceat, id vero non mediocriter pertimesco[15].

Parfaitement fidèle aux éditions utilisées, le PHI marque les coupures de mots en fin de ligne et présente les résultats dans une typographie en tous points conforme à celle de l'édition. Si cette solution semble à première vue correcte, elle présente néanmoins de très graves inconvénients qui faussent les recherches.

Le fait de couper des mots en fin de ligne a pour conséquence qu'on ne les trouve pas avec la procédure de recherche normale. Ainsi, si l'on souhaite obtenir les occurrences de la suite de caractères *eloquentia* dans le *Pro Quinctio*, on n'obtiendra pas pour le texte ci-dessus, la forme « elo-€quentia » qui se trouve à la 3e ligne et qui est en fait répartie sur deux lignes dans l'édition.

Un autre type de problèmes est lié aux graphies : pour respecter les éditions de référence, la banque de données du PHI n'a normalisé ni les lettres « u » et « v » ni les lettres « i » et « j » et le progiciel ne tient pas compte de cette distinction. Le résultat est que si l'on tente de repérer dans les œuvres de Cicéron les occurrences de *Iuppiter*, on obtient 50 références pour la suite de caractères *Iuppiter*, il faut alors procéder à une nouvelle recherche pour obtenir les 3 emplois de ce mot qui commence par la lettre « J ». De même, si l'on recherche *arma*, on obtient 10 références mais on constate que Virgile n'est pas repris ; la référence à l'*Énéide* n'est repérée que si l'on cherche « arma *uirumque cano* en raison de la graphie « u » à l'initiale de *uirumque*.

D'autres types de recherches se révèlent lacunaires sans que l'on trouve avec certitude une explication. Ainsi, pour la requête *arma uirum*, on obtient pour l'ensemble du PHI 14 emplois et

---

15 *M. Tullius Cicero. Pro Quinctio* (*M. Tulli Ciceronis Orationes*. Vol. 4, ed. A. C. Clark, 1909).

pour *arma virum* 21, soit au total 35 occurrences. La recherche des mêmes mots formulée en recourant à l'algèbre de Boole (*arma uirum* | *virum*) donnera 34 références. Le progiciel ne repère pas une référence à l'œuvre de Suétone pour des raisons difficiles à comprendre. En « Bêta code », le texte se présente comme suit :

> &`Prat 176.14 &`principium semel, initium saepius: principium ut&1 'arma&`[16]
> &`Prat 176.15 &1&`uirumque&` cano', initium 'musa mihi causas memora'. in-

Il est difficile d'expliquer cette erreur : on peut supposer qu'elle est due au logiciel qui traite de manière incomplète, approximative, certaines expressions booléennes en fonction de la mise en page du texte.

Les variantes graphiques constituent souvent un obstacle à des recherches précises. Pour reprendre l'exemple de *arma uirum* (ou *virum*) , on s'aperçoit que, selon les éditions, deux ou quatre références à Servius sont ignorées car l'éditeur a choisi une typographie en lettres capitales de sorte que les deux mots qui se présentent comme suit « ARMA VIRVM » ne sont pas repérés par PHI[17] à cause du « V » à la suite de la lettre « R ».

À côté de l'omission de certaines occurrences, on doit aussi avoir à l'esprit que la requête *arma uirum* | *virum* signifie pour le système qu'il y a lieu de rechercher toutes les suites de caractères « arma » suivies d'un mot où figurent les suites de caractères *uirum* ou *virum*, ce qui a pour résultat de repérer des références qui ne correspondent pas à ce que l'on recherche. Ainsi, dans le cas de la formule de recherche reprise ci-dessus, on obtiendra parmi les références correctes le vers suivant de Stace :

> fixa tremunt **arma**ntque **uirum**; saepe aspera passus[18]

On trouve encore d'autres équations de recherche qui conduisent soit à du silence[19], soit à du bruit[20] dans les résultats. Cela signifie que l'on ne peut se fier aux données fréquentielles fournies par le système en procédant à une seule requête – il faut penser à tout – et sans vérifier tous les contextes proposés.

Enfin, le logiciel du PHI accepte aussi des signes critiques qui permettent de rechercher soit des débuts de mots, soit des fins de mots.

Ainsi, la référence inadéquate *armantque uirum* de Stace peut être éliminée si on fait appel aux signes > et < qui permettent respectivement de préciser ce que l'on recherche. Ainsi, en dactylographiant « >arma », ne seront repérées que les occurrences des formes qui commencent par les quatre lettres « a r m a ». La formulation « >arma< » signifie que seules les formes « arma » doivent être repérées. De même, « >*iter* » repérera toutes les formes qui commencent par les 4 lettres « iter », tandis que « iter< » fournira tous les mots qui se terminent par « iter ».

Pour terminer cette partie consacrée au PHI, on retiendra, d'une part, que les résultats fournis par les requêtes simples et surtout complexes ne sont jamais entièrement fiables, que subsistent des erreurs ou des lacunes dues soit à des questions de typographie (le bêta code), soit à des variantes

---

[16] Suétone, *Prata*, 176, 14.
[17] Servius, *Commentaires à l'Énéide*, *ARMA VIRVM addidit …*
[18] Stace, *Theb* 2, 605.
[19] Omission de références.
[20] Texte inadéquat par rapport à la requête.

graphiques (*aff-*, *adf-*, …), soit surtout aux diverses formes d'un même mot résultant de la morphologie.

En regard de ces critiques, on doit souligner que le progiciel du PHI comporte des fonctions intéressantes : la constitution d'un index des formes, l'établissement de concordances ou des relevés statistiques, ou des listes classées à partir de la finale des mots (listes inverses).

Enfin, lorsque l'on travaille sur la totalité de la base, on regrettera la relative lenteur du logiciel de consultation.

A.2 – Le CD-Rom BTL

Diffusée depuis 2004 ans dans sa version 3, le CD-ROM BTL est la version numérisée de la *Bibliotheca Scriptorum Romanorum Teubneriana* ; l'éditeur responsable de ce support distribué par Brepols est le Centre *Traditio Litterarum Occidentalium* (*CTLO*), qui a collaboré avec le Cetedoc (UCL) pour la version électronique de la *Bibliotheca Scriptorum Romanorum Teubneriana*, ainsi qu'avec le L.A.S.L.A. pour des œuvres antérieures au IIe siècle de notre ère.

Selon les responsables de la BTL, la banque de données contient les œuvres latines datées entre le IIIe siècle avant notre ère et le IIe siècle PCN ainsi que les principaux auteurs non chrétiens jusqu'au IXe siècle (+/- 280 auteurs et 600 textes) ; elle inclut en partie les *Grammatici Latini* et *Servius Grammaticus* plus quelques textes ; enfin elle incorpore des textes latins du Moyen Âge et des auteurs modernes. Au total, on y trouve 454 auteurs et, selon la publicité, environ 1000 textes – en réalité 908.

Les textes de la BTL sont répartis comme suit selon la chronologie :

Antiquitas : origines jusqu'à la fin du IIe siècle
Infima Antiquitas/Aetas Patrum : débute à la fin du IIe siècle et se termine en 735
Medium Aeuum : 736 à 1500
Recentior Latinitas : après 1500

La BTL offre plusieurs méthodes de recherches et d'accès à l'information ; elle autorise des requêtes sur les points suivants *auctor*, *titulus, clauis*, *aetas*, *formae* lesquels concernent respectivement, la sélection de l'auteur, du titre d'une œuvre, du numéro (*clauis*) de l'auteur dans le *Handbuch der Lateinischen Literatur der Antike*, de l'époque et enfin d'un mot-forme. Ces différents critères de sélection peuvent être combinés.

Après avoir choisi un texte, on peut faire appel à la rubrique *Memento* qui donne des indications sur l'édition, des données statistiques et des précisions sur l'origine de la version numérisée. Voici à titre d'exemple les données fournies pour les *Commentaires de la guerre civile*.

Caesar (Caius Iulius Caesar)
100 a. Chr. - 44 a. Chr.
Commentarii belli ciuilis - s. 1 a.c. - prosa
LLA 260 - TLL CAES. civ.
Teubner (A. Klotz, 1950)

Summa formarum : 33268
Summa formarum dissimilium : 8165
Summa notarum (bytes) : 235610

- L'œuvre date de 45 av. J.-Chr. environ.

- Nous tenons à remercier vivement le 'Laboratoire d'Analyse Statistique des Langues Anciennes' (LASLA) de l'Université de Liège qui nous a transmis une version magnétique de cette œuvre selon l'édition Teubner retenue.
Cette copie nous a été de la plus grande utilité pour l'élaboration du fichier intégré dans cette base de données.

À partir de l'entrée *formae* sont repérés tous les contextes (*sententiae*) qui correspondent au mot-forme recherché. La fréquence d'emploi de ce mot est indiquée au bas de l'écran. Le logiciel offre la possibilité d'afficher les différentes *sententiae* précédées de leur référence (généralement conforme à l'*ars citandi*). Le logiciel trouve en une seule requête *virum* ou *uirum*, il associe indifféremment « u » et « v ».

L'affichage d'une *sententia* conduit aussi à la lecture de l'intégralité de l'œuvre dans laquelle elle apparaît, avec possibilité de sauvegarder des résultats.

Pour ce qui relève de la formulation des requêtes, il faut souligner que la BTL offre des solutions intéressantes tant au point de vue linguistique qu'au point de vue technique. En voici un exemple qui fait usage de l'astérisque.

La requête dans la zone *formae* de la suite de lettres *moment** relève toutes les formes qui commencent par *moment-* quel que soit le nombre de caractères qui suit le « t-» . On obtient 611 formes venant de *momentum*, *momentana*, *momentaneus*, *momentarius*, *momentosus* et *momentaliter*.

À l'inverse, la recherche de **iter* fournit tous les mots qui se terminent par –*iter*, à savoir 9 103 occurrences qui sont en majorité des adverbes (*aliter*, *miserabiliter*, *pariter*, … mais aussi des substantifs tels que *iter*, *arbiter* ou *Iuppiter*).

Enfin, si on pose comme interrogation **iter**, le système fournit 9 431 références qui reprennent tous les mots où figure la suite de caractères *iter* quelle que soit sa position dans un mot, on obtient ainsi, par exemples, *aliter*, *miserabiliter*, *liternus* et des formes telles que *confitere*, *constiterunt*, …

Le logiciel de la BTL propose d'autres formules de requête, ainsi le point d'interrogation remplace un caractère quelconque – un seul – ou aucun caractère à l'intérieur d'un mot ; en outre, on peut utiliser plus de deux points d'interrogation par mot. Si on dactylographie « m?t?e? » à côté de *formae* le système repère *matre*, *matrem* et *matres* et aussi *mittet, mittes, metuet,* …

L'emploi du point d'interrogation permet de trouver en une seule requête des formes pour lesquelles il y a assimilation, par exemple, la requête « a ?fici ? » va isoler 76 occurrences quelle que soit la lettre qui suit le « a » initial ou celle qui remplace le « ? » final[21] : *Aeficio*, *adficio*, *afficio, adficis, afficis, adficit, afficit.*

La BTL apparaît comme une base de données de qualité supérieure au CD-ROM du PHI. Elle offre davantage de possibilités d'interrogations – que nous n'avons pas reprises ici pour ne pas allonger cette partie –. Au rang des avantages que l'on reconnaît à la BTL, on mentionnera la rapidité du système qui fournit immédiatement les résultats, le fait que les lettres « u » et « v » sont repérées en une seule recherche même si elles sont distinguées dans l'édition, la pertinence

[21] La différence de résultats produits par « * » ou par « ? » est que l'astérisque remplace un nombre variable de caractères. Si « a ?fici ? » identifie 76 occurrences, « a ?fici* » en repère 396 comme *adficiatur*, *afficiuntur*, *adficitur*, *afficias*, *adficiatis*, …

des résultats (*pater* mais pas *paternus*) et leur exhaustivité. En outre, on dispose d'un bon aperçu du contexte et de statistiques précises. Enfin, il est aisé de conserver les résultats pour les utiliser dans un traitement de texte.

La BTL présente néanmoins l'inconvénient de ne contenir aucune information lexicologique ou morphologique. Pour reprendre les exemples cités précédemment, on constate que les 666 occurrences de *ducis* doivent être analysées par le chercheur pour savoir de quel mot chaque emploi est dérivé. En dernier lieu on soulignera le fait que l'on ne peut obtenir que des impressions partielles des contextes en raison du fait que Teubner tient à préserver ses droits.

A.3 – Le CD-Rom « Hyperbase »

Le système « Hyperbase » est construit sur la base de données du L.A.S.L.A.[22] Chaque mot d'un texte est donc accompagné d'informations lexicales, morphologiques et syntaxiques.

Ce système a été développé et mis au point par Étienne Brunet et Sylvie Mellet de l'Université de Nice[23] qui en ont donné une description.[24]

« Hyperbase » possède des fonctions multiples qualitatives et quantitatives. Sans entrer dans le détail, on retiendra qu'il permet notamment de rechercher toutes les occurrences d'un vocable, d'une forme, de caractères morphologiques ou de structures syntaxiques et d'en fournir la concordance.

Par ailleurs, « Hyberbas » possède de nombreuses possibilités d'analyse statistique des données : écart type, graphique, analyse factorielle[25] ou analyse arborée[26].

Les possibilités qu'offre « Hyperbase » sont trop nombreuses pour être décrites ici. On signalera simplement que son emploi demande une formation de base en informatique et une connaissance certaine des méthodes statistiques.

---

[22] Le lecteur trouvera une description de la base de données du L.A.S.L.A. dans différents articles de Joseph Denooz; il consultera, par exemple, « Littératures classiques et banques de données » dans *Informatica e scienze umane, Mezzo secolo di studi e ricerche*, Lessico Intelletuale europeo Leo S. Olschki, 2003, p. 107-128; « *Opera latina* : une base de données sur internet », dans *EVPRHOSYNE*, 32 (2004), p.79-88. Deux articles plus anciens contiennent des descriptions plus détaillées de la méthodologie du L.A.S.L.A. : « Le traitement des textes latins, grecs et français au Laboratoire d'Analyse statistique des Langues anciennes », dans *Revista de la Universidad Complutense*, 25 (1976), p. 143-167 et tout particulièrement « L'ordinateur et le latin, Techniques et méthodes », dans *Revue de l'organisation internationale pour l'étude des langues anciennes par ordinateur*, 1978, 4, p. 1-36.

[23] Dominique Longrée et Gérald Purnelle de l'Université de Liège y ont collaboré pour l'adaptation de la base de données et pour les tests du logiciel.

[24] Mellet, Sylvie, Brunet, Étienne, *Hyperbase : Logiciel hypertexte pour le traitement documentaire et statistique des corpus textuels conçu et développé par Étienne Brunet. Manuel de référence pour la base de Littérature latine*, Nice, s.d.,46 p.

[25] Françoise et Jean-Paul Benzécri, *Analyse des correspondances [et classification]*, Paris, Dunod, 1984.

[26] Michel Juillard et Xuan Luong, « Des feuilles aux racines : du discours à la langue », dans *Revue informatique et statistique dans les sciences humaines*, 24(1988), p. 221-240. Xuan Luong et Sylvie Mellet, « Mesures de distance grammaticale entre les textes », dans *Corpus* 2 (2003). Sur l'aspect théorique de la mesure des distances entre textes et la comparaison de différentes méthodes, on se reportera à Étienne Brunet, « Peut-on mesurer la distance entre deux textes ? », dans *Corpus* 2 (2003). Ces deux derniers titres sont en texte intégral et accessibles à partir de l'adresse : http://corpus.revues.org/sommaire52.html (consulté le 18 mai 2009).

B.- Banques de données sur internet : quelques exemples

B.1 - *The Latin Library*

La banque de données *The Latin Library*[27] (ci-après *TLB)* contient un grand nombre d'auteurs de diverses époques ; elle comporte cinq subdivisions, à savoir une page où sont repris les principaux auteurs du Ier siècle avant notre ère, du Ier siècle PCN et aussi d'époques plus tardives où figurent, par exemple, Aulu-Gelle et Aurelius Victor.

La 2e rubrique s'intitule *Miscellany* ; elle ne présente aucune cohérence puisque l'on y trouve des auteurs tels que L. Andronicus, Naevius, Germanicus, Ausone, Grégoire de Tours et même une traduction latine du XIXe siècle du Byzantin Jean Zonaras. Les trois subdivisions suivantes sont respectivement consacrées à des auteurs chrétiens, à des auteurs du Moyen Âge et à des textes néo-latins : Descartes, Érasme, Scaliger, Thomas More et même le texte latin de l'examen de maturité de Karl Marx. Au total, *TLB* contient des textes de quelque 323 auteurs.

Les données ne sont rien d'autre que la copie d'une édition avec les indications de référence. Dans les textes de la *TLB*, on ne retrouve pas la disposition en lignes de l'édition et les mots ne semblent pas coupés en fin de ligne. Par ailleurs, aucune mention n'est prévue pour identifier une citation grecque. Ainsi, dans Sénèque, *De ira*, I, 20,8, le texte *C. Caesar* […] *Homericum illum exclamans uersum h[μVαjnαveirV h] ejgw; sev* qui apparaît correctement dans PHI[28] se présente dans *TLB* sous la forme *C. Caesar* […] *Homericum illum exclamans uersum* « g-ê g-m' g-anaeir' g-ê g-egô g-se ».

Cette critique de *TLB* paraît mineure à côté de reproches plus fondamentaux. Le premier et le plus irrémissible est que l'utilisateur ne peut connaître avec précision l'origine des copies numérisées.

Au bas de l'écran d'accueil de la BTL figurent quatre rubriques ; les deux premières sont très importantes et le lecteur devrait y être conduit avant toute utilisation d'un texte ; elles sont intitulées « Credits » et « About These Texts ».

Cette dernière mentionne explicitement le fait que les textes numérisés proviennent de sources diverses ou qu'ils ont été recopiés à partir d'autres sites WEB rarement identifiés et dont certains ont disparu. En outre, les éditions de base sont soit anciennes, soit peu ou pas identifiées. Enfin, les responsables du site reconnaissent qu'en dépit de leurs efforts, subsistent beaucoup d'erreurs dues au scanner ou à la transcription.

La rubrique « Credits » donne des indications peu homogènes sur les auteurs et les œuvres. La première notice dans « Credits » est reproduite ci-dessous.

> **Abbo Floracensis**, *Passio Sancti Edmundi Regis et Martyris* – the story of Edmund King in East England written c. 985, purporting to be a record of the story as told to Abbo of Fleury, who had

[27] http://www.thelatinlibrary.com/

[28] Dans la BTL, les mots grecs sont précédés d'un « g- » : « C. Caesar […] homericum illum exclamans versum : g-ê g-m' g-anaeir' g-ê g-egô g-se ».

> heard it directly from Edmund's sword-bearer. Submitted by Paolo Paoletti (Perugia, Italy) from an unidentified edition.[29]

À la suite du nom de l'auteur et du titre de l'œuvre, quelques mots précisent le sujet du texte. On découvre ensuite le nom et l'origine de la personne qui a numérisé le texte et on constate que l'édition n'est pas identifiée. Cette dernière indication est très critiquable : il est indispensable de connaître l'édition. On ne sait rien ou à peu près rien de la personne qui a soumis le texte.

Les incohérences de la rubrique « Credits » sont diverses. Pour certains textes, il n'y a aucune indication ni sur l'édition ni sur le responsable de la version numérisée. Le cas de Sénèque est particulièrement significatif avec les indications suivantes :

> *Epistulae Morales ad Lucilium*- submitted by Hansulrich Guhl (Frauenfeld, Switzerland) from an unidentified edition and (the later books) by Sally Winchester from the Reynolds edition.
> *de Vita Beata* submitted by Erich Schweizer-Ferrari in Luzern, Switzerland.
> *Medea* from the *Bibliotheca Augustana* with the kind permission of its webmaster Ulrich Harsch.
> *Phaedra* - posted by William Carey from an unknown edition.
> *Apocolocyntosis Divi Claudii* - scanned by William L. Carey from the Loeb edition of 1913.

On cherche en vain des indications sur les œuvres qui sont intégrées à la BTL, à savoir *Quaestiones naturales*, les trois consolations, *De Ira*, les dialogues et la plupart des tragédies.

Dans d'autres cas, se trouve une référence à l'édition sans aucune mention de la personne qui a soumis le texte à *TLB* :

> Alanus de Insulis, *de Planctu Naturae* – from the edition of J. P. Migne.

On pourrait multiplier les exemples de cette absence de rigueur scientifique qui se rencontre même pour des auteurs connus tels que Saint Ambroise et Saint Anselme dont les notices « Credits » respectives sont :

> Ambrose, *de Principio Individuationis* - from an unidentified e-text.
> Anselm, *Epistula ad Urbanum Papam*- from an unknown edition.

TLB ne donne aucune indication précise sur les textes qu'elle rassemble : à première vue, on est porté à croire qu'il s'agit d'une base exhaustive de la littérature latine antérieure au II^e^ siècle de notre ère. Or, des œuvres sont manquantes pour plusieurs auteurs. En ce qui concerne Sénèque le Philosophe, on cherche en vain les tragédies *Hercules Furens*, *Troades* et *Phœnissæ*. Dans les textes d'Ovide ne se trouve pas *De medicamine faciei*.

---

[29] La recherche de « Paolo Paoletti » par Google répertorie plusieurs personnes qui portent ce prénom et ce patronyme. Celui qui pourrait être le responsable de la numérisation du texte de la *Passio Sancti Edmundi Regis et Martyris* est peut-être, sans certitude, le Dr. Paolo PAOLETTI (Italie – Voghera, département de Pavie) Diplômé ès Philosophie à l'Université de Pavie qui est directeur de la Bibliothèque Civique « Ricottiana » de Voghera et responsable des services et des institutions culturelles de la Commune; depuis 2002, en tant que directeur du Système Bibliothécaire intégré de l'*Oltrepo* de Pavie, il est chargé de superviser le rôle délicat de ce département. Il a été archiviste des Archives Historiques de la Curie épiscopale de Tortona et a publié différents travaux pour la réorganisation des archives, sur la recherche historique et bibliographique.

Pour terminer ce paragraphe consacré à *TLB*, le lecteur pourra conclure lui-même que l'on ne doit accorder qu'une confiance très limitée à ce site. En outre, il ne contient aucune donnée philologique ou linguistique ; en d'autres termes, il n'apporte rien de plus qu'une édition traditionnelle.

B2. *Intratext*

Le site *Intratext* contient un très grand nombre de textes appartenant à plus de 40 langues. La page d'accueil informe l'utilisateur qu'*Intratext* est une bibliothèque « full-text » réalisée par des experts (*managed by experts*) qui travaillent de manière scientifique.

Ces indications très vagues ne mentionnent pas le fait que beaucoup de textes sont repris à *The Latin Library* ou à des sources difficiles à identifier ou encore à des éditions anciennes. Ainsi, pour la tragédie *Agamemno* de Sénèque, *Intratext* cite comme source numérisée *The Latin Library*, or cette dernière ne fournit d'indication ni sur l'édition utilisée, ni sur l'origine du texte numérisé.

À ces remarques on peut ajouter que l'on peut adresser à la base *Intratext* les mêmes critiques, parfois plus graves encore, qu'à *TLB*. En outre, pour des précisions biographiques sur les auteurs, *Intratext* renvoie presque exclusivement à *Wikipédia*, encyclopédie dont la fiabilité est pour le moins douteuse.

Les défauts de cette base textuelle sont quelque peu atténués par le fait qu'elle offre des possibilités d'utilisation qui ne sont pas dépourvues d'intérêt. Basé sur les techniques hypertextuelles, le logiciel comporte des possibilités diverses d'interrogation des textes, soit uniquement dans la langue source, soit avec traduction. À partir du texte original, tous les mots sont cliquables ce qui permet d'obtenir pour une forme une ligne de concordance à partir de laquelle il est aisé de retourner au texte si on clique sur une partie de la référence.

Ainsi, en choisissant *urbem* premier mot des *Annales*, le système repère les 126 occurrences de la forme et affiche en ordre du texte, une brève concordance de chaque emploi dans tous les textes de Tacite. On trouve ici, à titre d'exemple, le début de cette concordance.

Historiae
Lib. Cap.

1	1, 1 |	erunt. nam post conditam **urbem** octingentos et viginti prioris
2	1, 6 |	perierant. introitus in **urbem** trucidatis tot milibus inermium
3	1, 29|	motus habebamus incruentam **urbem** et res sine discordia translata
4	1, 37|	acceperat. his auspiciis **urbem** ingressus, quam gloriam
5	1, 39|	seditionis et vocibus in **urbem** usque resonantibus, egressum
6	1, 50|	50-~ ~Trepidam **urbem** ac simul atrocitatem recentis

On peut à partir de cette liste cliquer chacun des mots soulignés pour obtenir la liste de ses emplois.

Comme les bases déjà vues, *Intratext* ne contient ni données lexicologiques, ni morphologiques, ni syntaxiques. Les conséquences sont évidentes ; le système n'opère aucune distinction entre les formes qui peuvent appartenir à des mots différents. Ainsi, dans Tacite la forme *ducis* se rencontre 44 fois, le chercheur devra examiner chaque contexte pour distinguer les emplois de *dux* et de *duco*.

La base *Intratext* comporte des lacunes importantes, ainsi les livres 2 à 8 du *Bellum Gallicum* ne s'y trouvent pas ; n'y sont pas non plus les *Verrines*.

Enfin, certains textes sont traduits les uns en anglais, les autres en français ou en italien, sans que l'on sache ce qui a guidé le choix de la langue. Pour certaines œuvres existent deux traductions ; ainsi, pour Salluste, le *De coniuratione Catilinae* est traduit en anglais et en français ; l'*Énéide* existe en anglais et en italien.

B.3 - *Itinera electronica*

Le site *Itinera electronica*[30] est développé à l'Université catholique de Louvain (Belgique). Il contient diverses informations historiques, bibliographiques, etc. pour l'étude des langues classiques ; on ne retiendra ici que la banque textuelle latine dans laquelle sont enregistrés des textes bruts avec, dans de nombreux cas, une traduction française et des textes en présentation hypertextuelle[31].

Selon ses concepteurs, cette base textuelle comprend trois millions de mots[32] ; en examinant de près la présentation qui est faite de cet ensemble, on constate que la majorité des textes ont comme origine soit le site *The Latin Library*, soit d'autres sites tels que celui géré par Philippe Remacle[33]. *Itinera electronica* n'est pas très explicite à ce sujet ; il se contente de mentionner de manière assez discrète[34] :

> en règle générale, les données textuelles latines ont été téléchargées du site américain *The Latin Library* […]. Dans certains cas, dûment signalés au niveau du module LECTURE de l'*environnement hypertexte*, le téléchargement s'est fait à partir d'un autre site, ou, encore, la saisie optique a été réalisée par nos propres soins.

La consultation de plusieurs sites web est souvent nécessaire pour découvrir l'origine des textes numérisés. Il est clair que les critiques formulées à propos de *The Latin Library* et d'*Intratext* s'adressent aussi à *Itinera electronica*. Ceci ressort encore de deux phrases que l'on lit dans la page d'accueil de la base :

> Il n'y a pas eu de relecture *systématique* de ces données latines. Il n'y a pas eu non plus d'investigations quant à l'édition retenue. Les ressources humaines à disposition ne le permettaient pas. Par conséquent, aucune garantie ne peut être donnée ni quant à la valeur scientifique de l'édition de ces textes ni quant à la qualité de leur encodage ou de leur saisie optique. C'est un *matériel* mis à la disposition de tous, gratuitement et en libre accès. Le chercheur veillera à en vérifier la conformité scientifique aux standards actuels préalablement à son utilisation et à son exploitation.

Pour les textes accompagnés de leur traduction française, l'utilisateur est laissé dans la même ignorance en ce qui concerne l'origine des traductions dont on sait qu'elles viennent d'éditions qui ne tombent plus sous les lois des droits d'auteurs (c'est-à-dire des éditeurs) et qu'elles ont été numérisées selon les techniques de reconnaissance optique des caractères (*OCR*). Si dans le menu d'accès à une œuvre on clique sur « texte », on voit apparaître la mention de l'édition latine et l'origine de la traduction. La correspondance de Cicéron est accompagnée des indications qui figurent ci-dessous.

---

30 http://pot-pourri.fltr.ucl.ac.be/itinera/

31 http://agoraclass.fltr.ucl.ac.be/concordances/intro.htm

32 Le site *Itinera electronica* indique : « à la date du 24 mai 2004 ». Aucune indication nouvelle n'a été donnée depuis cette date déjà lointaine.

33 cf. le paragraphe B.4.

34 Cf. le site http://neptune.fltr.ucl.ac.be/corpora/

Traduction française : Œuvres complètes de Cicéron dans : Collection des auteurs latins publiés sous la direction de M. NISARD, t. V, Paris, Dubochet, 1841.

Pour les *Métamorphoses* d'Ovide, on trouve « Traduction (légèrement adaptée) de G.T. Villenave, Paris, 1806 ». Le site de *l'Histoire Auguste* indique quant à lui « Traduction française reprise au site Philippe Remacle. Traduction française : Laass d'Aguen - E. Taillefert, *L'Histoire Auguste*. Tome II. Paris, 1846 ».

La consultation de la liste des auteurs numérisés révèle pour quelques cas que la base est lacunaire : seules quatre comédies de Plaute sont intégrées à *Itinera*, pour Quintilien, les livres VI à IX manquent, pour Térence seules deux pièces sont reprises… Les responsables de la base textuelle annoncent qu'ils combleront progressivement ces lacunes.

Comme *Intratext* – et peut-être inspiré par cette base –, *Itinera* offre des possibilités d'utilisation qui ne sont pas dépourvues d'intérêt. Basé sur les techniques hypertextuelles, le logiciel comporte des possibilités diverses d'interrogation, soit uniquement dans la langue source, soit avec traduction française. À partir du texte original, tous les mots sont cliquables ce qui permet d'obtenir pour une forme une ligne de concordance à partir de laquelle il est aisé de retourner au texte qui dans ce cas est accompagné de sa traduction.

Ainsi, en cliquant sur le premier mot de *Bellum Gallicum*, le système repère toutes les occurrences de la forme *Gallia* – 62 emplois – et affiche soit en ordre du texte soit selon l'ordre alphabétique du mot qui suit immédiatement *Gallia*, une brève concordance de chaque emploi. On trouve ici, à titre d'exemple, pour le livre I de *La Guerre des Gaules* les occurrences de *Gallia* suivies de *citeriore* :

8, 54	nomine quintam decimam, quam	Gallia	citeriore habuerat, ex senatus consulto
8, 23	anno Titus Labienus, Caesare in	Gallia	citeriore ius dicente, cum Commium
1, 24	iugo duas legiones quas in	Gallia	citeriore proxime conscripserat et omnia

On peut à partir de cette liste cliquer sur une des occurrences de *Gallia* et obtenir la totalité du chapitre où ce mot est employé et en regard la traduction française.

Comme les bases vues ci-dessus, *Itinera* ne contient ni données lexicologiques, ni morphologiques ni syntaxiques. Les conséquences sont évidentes ; le système n'opère aucune distinction entre les formes qui peuvent appartenir à des mots différents. Ainsi, dans les *Verrines* la forme *legi* se rencontre trois fois, mais seule une analyse des contextes par le chercheur permettra de distinguer les emplois de *lex* et de *lego* puisque le logiciel donne comme résultat pour la requête :

21, 23 accepti tabulas omnis, quas diligentissime	legi	atque digessi, patris, quoad uixit,
25,69 lex statim promulgata est. Cui	legi	cum uestra dignitas uehementer aduersetur,
22, 31enim et ex iis	legi	et audiui intellego, in qua

La base *Itinera* comporte d'autres possibilités : elle peut fournir une liste des formes classées alphabétiquement à partir de la finale, un relevé du vocabulaire qui n'est en réalité qu'un relevé des formes ; celui-ci comporte des données statistiques partielles. D'autres possibilités de recherche existent mais elles varient d'un auteur à l'autre. Ainsi, pour *l'Histoire Auguste* un renvoi conduit vers le site de Philippe Remacle, pour la *Guerre des Gaules* le renvoi se fait vers la *Bibiotheca classica selecta*[35].

[35] La page d'accueil de ce site bien connu des hellénistes et des latinistes précise le but de ses responsables : *Conçue et maintenue par deux professeurs belges, Jean-Marie Hannick (Université de Louvain, à Louvain-la-Neuve) et Jacques Poucet (Université*

Il est inutile de poursuivre l'examen d'*Itinera*. On peut conclure que la fiabilité des textes latins – et grecs – et des traductions françaises y est aussi sujette à caution que dans *TLB* ou dans *Intratext*.

En outre, on constate un manque d'uniformité et de cohérence dans la méthode de travail puisque les possibilités de consultation diffèrent d'un auteur à l'autre.

Enfin, comme la *TLB*, la base textuelle de l'Université de Louvain ne contient aucune information linguistique. C'est au chercheur qu'incombe tout le travail philologique.

Le fait que certains auteurs ne soient repris qu'en partie se comprend aisément quand on pense à l'ampleur de la tâche. Les responsables de la base assurent que les lacunes seront progressivement comblées.

B.4 – Le site de l'Antiquité grecque et latine

Ce site, composé de sept rubriques, rassemble des textes grecs et latins avec traduction, des ouvrages historiques, bibliographiques, institutionnels, des textes commentés destinés à aider l'enseignant dans la préparation d'une leçon, destinés aussi aux étudiants, etc. Il est l'œuvre de professeurs de latin et de grec de l'enseignement secondaire belge[36].

La partie à examiner est, bien entendu, celle qui s'intitule « Traductions d'auteurs latins ». Elle rassemble une partie importante de la littérature latine et, à certains égards, elle est plus complète que *Itinera* puisque l'on y trouve 16 comédies de Plaute. Térence dont il est dit « Œuvres complètes » a cette particularité de présenter, par exemple, pour *Les Adelphes* la seule traduction française, alors que pour les autres comédies on dispose du texte latin et de la version française.

Les critiques que l'on peut formuler ici sont très semblables à celles qui ont été faites aux bases précédentes : il n'y a aucune donnée philologique ou linguistique : ni lexique, ni morphologie, ni syntaxe. Il s'agit donc uniquement de textes.

B.5 – *Perseus*

Initié en 1985 par l'Université de Yale, le projet *Perseus* comprend diverses bases textuelles anglaises, arabes, grecques et latines. On y trouve aussi des banques de données pour l'histoire de l'art, pour l'histoire américaine du 19e siècle, etc.

La plupart des textes anciens sont repris soit au TLG pour le grec soit au PHI pour le latin. Ce site toujours en développement ne contient que 71 textes latins, il présente donc d'importantes lacunes : on n'y trouve, par exemple, ni Sénèque le Rhéteur, ni Sénèque le Philosophe ; ne sont reprises ni *la guerre d'Espagne*, ni *la guerre d'Afrique*, ni *la guerre d'Alexandrie*, de même sont absents du corpus des auteurs tels que Martial et Juvénal, ...

---

de Louvain à Louvain-la-Neuve et Facultés universitaires Saint-Louis à Bruxelles), la BIBLIOTHECA CLASSICA SELECTA (BCS) *se veut une introduction aux études classiques, destinée prioritairement aux étudiants de lettres classiques et d'histoire ancienne, accessoirement à tous ceux qui s'intéressent au monde gréco-romain antique et aux civilisations qui l'entourent.*

36 Le site s'appelle http://remacle.org/. Il est coordonné par Philippe REMACLE, professeur de latin dans l'enseignement secondaire supérieur.

Le site *Perseus* autorise des accès à des dictionnaires et à des données grammaticales pour la plupart des mots d'un texte, mais sans lever les ambigüités. Pour chaque mot une traduction de base est donnée en Anglais.

Pour la plupart des mots figurant dans un contexte, on peut obtenir sous la traduction toutes les analyses possibles : c'est au philologue de choisir. Ainsi, pour la forme *omnis*, le système fournit les analyses qui figurent ci-dessous.

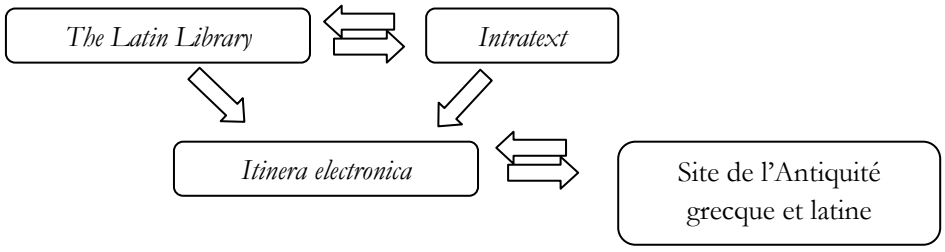
omnis	all, every
omnîs	masc acc pl
omnîs	fem acc pl
omnis	masc gen sg
omnis	fem gen sg
omnis	neut gen sg
omnis	masc nom sg
omnis	masc voc sg
omnis	fem nom sg
omnis	fem voc sg

On constate que *Perseus* ne résout pas plus les ambigüités que les autres systèmes. Bien qu'il comporte des aides à la compréhension et à l'analyse des formes, il n'apporte pas une aide suffisante au chercheur et s'il permet d'établir des relevés, il oblige néanmoins le chercheur à examiner au cas par cas tous les contextes pour en extraire ce à quoi il s'intéresse.

C.- Conclusion

Il serait possible d'étudier encore d'autres bases textuelles latines consacrées soit à un seul auteur, soit à un genre littéraire : leur examen n'apporterait rien à ce qui précède[37].

Pour les bases *The Latin Library*, *Intratext*, *Itinera electronica* et pour le site de l'Antiquité grecque et latine, on voit se dessiner un arbre généalogique qui remet en question la valeur de la transcription des œuvres latines. Ceci conduit à se poser une question : ces bases sont-elles des copies légales ou des plagiats ?



On peut donc dire pour conclure que la multiplication des sites WEB latins n'est guère utile au latiniste, que ce soit sur le plan philologique ou linguistique. Au contraire, il convient de recourir

[37] En voici quelques exemples :
- Horace : http://www.espace-horace.org/
- Salluste : http://www.mediterranees.net/histoire_romaine/salluste/index.html (en français)
- Virgile : plusieurs sites y compris des fichiers PDF contenant la traduction d'une œuvre. C'est le cas de http://misraim3.free.fr/divers/les_bucoliques_de_virgile.pdf où se trouvent les *Bucoliques*.

avec prudence à ces versions digitalisées qui n'ont pas été corrigées avec suffisamment d'attention et qui n'apportent que peu d'aide par rapport aux textes des éditions traditionnelles. En outre, ces sites qui se copient les uns les autres se réfèrent le plus souvent à des éditions anciennes – qui ne sont plus sous les lois des droits d'auteurs – et à des traductions qui sont parfois « de belles infidèles ».

Comme nous l'avons souvent dit et écrit, si l'on numérise la littérature, il est inutile de le faire en se limitant à reproduire les éditions traditionnelles. Seules l'ajout de données philologiques confère valeur scientifique et apporte une aide véritable à la recherche linguistique, qu'elle porte sur le lexique, sur la morphologie ou sur la syntaxe. S'il est vrai que certaines bases textuelles (comme la BTL, par exemple) font appel à des logiciels bien développés et sont utiles aux chercheurs, il n'en reste pas moins vrai qu'ils ne résolvent pas les questions linguistiques les plus pointues.