Sylvie Mellet
C.N.R.S., « Bases, Corpus et Langage »
(UMR 6039, Nice)

Lemmatisation et encodage grammatical : un luxe inutile ?

C'est à partir d'une base de données de textes latins que la question posée en titre de cet article sera examinée ; on verra cependant que les points abordés, les réponses fournies, les enjeux détectés ont une portée qui dépasse largement le cadre des études classiques et que la contingence de nos compétences spécifiques ne doit pas occulter l'importance d'un problème méthodologique très général.

La première base de données textuelles latines ayant enrichi les textes bruts a été constituée par le L.A.S.L.A. (Laboratoire d'Analyse Statistique des Langues Anciennes) de l'Université de Liège, à partir de 1961. Cette base de données rassemble aujourd'hui une vingtaine d'auteurs latins différents (pour la plupart d'époque classique ou du haut-empire), soit encore 160 textes et environ 1,7 million de mots. L'objectif était, au départ, d'offrir à l'utilisateur les moyens de retrouver automatiquement toutes les occurrences d'un même vocable, quelles que fussent ses variantes orthographiques et ses formes flexionnelles ; ce qui, pour une langue comme le latin dans laquelle non seulement les verbes se conjuguent, mais aussi les noms et les adjectifs se déclinent, impose d'emblée le choix de la lemmatisation, c'est-à-dire l'établissement d'un lien entre chaque forme graphique (ou mot) du texte et l'entrée correspondante dans un dictionnaire de référence (ou lemme).

Ce regroupement des formes sous leur lemme implique une analyse morphologique préalable : ainsi pour savoir si la forme *legis* doit être rattachée au verbe *lego* ou au substantif *lex*, il convient d'avoir analysé la proposition dans laquelle apparaît cette forme et d'avoir déterminé si l'on a affaire à un verbe à la deuxième personne du singulier (*tu lis*) ou à un substantif au génitif singulier (*de la loi*). L'idée s'est alors imposée de conserver aussi ces analyses sous la forme d'un code alpha-numérique qui suit chaque mot dans la linéarité du texte et de prévoir des outils de lecture et d'exploitation automatique de ces codes.

On le voit, lemmatisation et encodage grammatical sont deux opérations distinctes qui, pour être complémentaires, ne sont pas nécessairement associées et, surtout, pas toujours conservées sous un format récupérable dans les bases de données. Le L.A.S.L.A. a eu le mérite de comprendre les avantages des deux opérations dès le début des années 60. La méthode de traitement est semi-automatique : à partir d'un dictionnaire des radicaux (ou bases) d'une part et d'une liste de toutes les finales envisageables en latin d'autre part, le programme effectue une lecture régressive de chaque forme et en propose toutes les analyses morphologiques possibles. Par exemple, la forme *naturae* est ainsi lue et analysée :

natura-e	rien	
natur-ae	génitif singulier	
	datif singulier	du substantif *natura*
	nominatif pluriel	
	vocatif pluriel	
natu-rae	rien	
nat-urae	participe futur du verbe *nascor*	
nat-urae	participe futur du verbe *no*	

C'est alors le philologue qui intervient pour sélectionner manuellement l'analyse juste en contexte. Cette dernière étape est évidemment longue ; mais elle permet de disposer *in fine* d'un texte lemmatisé, totalement désambiguïsé et enrichi d'étiquettes morphologiques fiables.

Chaque base textuelle est donc composée de trois fichiers (cf. Mellet : 1996) : le fichier comprenant le texte étiqueté, l'index des lemmes où, sous chaque entrée, sont rassemblées toutes les formes afférentes accompagnées de leur référence exacte dans le texte (et c'est là que s'enregistrent aussi les levées d'homonymie) et enfin l'index des propositions subordonnées. Une telle structuration

permet d'effectuer automatiquement et de manière très efficace (*i.e*. rapide, exhaustive et non bruitée) des recherches très variées portant soit sur une forme graphique, soit sur une chaîne de caractères (suffixes, préfixes par ex.), soit sur les occurrences d'un lemme quelles que soient ses formes fléchies (*sum*, *es*, *est*, *sunt*, *fuit*, *ero*, *eram*, etc.), soit sur un type de subordination (toutes les relatives au subjonctif), soit sur une catégorie grammaticale (tous les noms de la 3ème déclinaison à l'ablatif singulier). Nous n'insisterons pas sur la richesse de telles possibilités, que nous avons présentées ailleurs ; on sait bien en effet que la lemmatisation offre une solution élégante et efficace à certains problèmes majeurs du traitement automatique des textes tels que la levée d'ambiguïté et le traitement préalable des homonymies ; on sait aussi que l'encodage grammatical ouvre les corpus étiquetés à tous les types de recherches linguistiques là où les textes bruts n'intéressaient guère que les lexicographes et lexicologues. En revanche, nous voudrions revenir ici sur un autre des atouts majeurs de la lemmatisation, trop souvent négligé et sous-estimé, à savoir le traitement des variantes orthographiques et morphologiques.

1. Variantes orthographiques et morphologiques

Le latin est une langue à l'écriture relativement transparente : elle est en effet quasiment phonogrammique, chaque lettre étant toujours et partout prononcée et transcrivant constamment un seul et même phonème. Cependant l'orthographe est mal stabilisée et trahit souvent les évolutions phonétiques en cours ainsi que l'accueil facilement accordé aux formes orales et dialectales. Une telle malléabilité induit dans les textes les variantes orthographiques que nous allons détailler ci-dessous :

1.1. Evolutions phonétiques et absence de norme stabilisée :

On relève à ce titre :

_ des haplologies facultatives : *exspectare* / *expectare* « attendre » ; *exsilium* / *exilium* « exil » ;

_ des assimilations consonantiques transcrites ou non : *inlicio* / *illicio* « je séduis » ; *adtuli* / *attuli* « j'ai apporté » ; *quidquid* / *quicquid* « quoi que ce soit » ;

_ des transcriptions facultatives de phonèmes faibles (le h notamment ou les nasales devant fricatives et sifflantes) : *harena* / *arena* « le sable » ; *exhibeo* / *exibeo* « je montre » ; *consul* / *cosul* « le consul » ;

_ des traces de monophtongaisons en cours : *saeta* / *seta* « la soie » ; *plaudite* / *plodite* « applaudissez » ; *poenicus* / *punicus* « punique, carthaginois ».

Comme on le devine à travers ces quelques exemples, de telles alternances graphiques ne sont pas rares : elles touchent de nombreuses séquences graphiques et des lexèmes très courants ; nous n'en avons pas fait le décompte exhaustif, mais nous pensons que l'impact de ces variantes n'est pas négligeable, même dans des traitements statistiques de quelque ampleur.

1.2. Les élisions, épenthèses, aphérèses et contractions diverses :

Deux phénomènes sont ici à distinguer :

_ la transcription de formes orales (notamment dans les textes de théâtre) qui alternent avec la forme pleine normalement attendue à l'écrit : *certumst* / *certum est* « c'est certain » ; *credin ?* / *credisne ?* « tu crois ? » ;

_ les problèmes liés à l'édition moderne de textes qui, dans leur forme originelle, pratiquaient la *scriptio continua* (écriture continue ne séparant pas les mots par des blancs et ne connaissant aucun signe de ponctuation) : ces choix éditoriaux modernes créent des alternances, sans doute artificielles, du type : *quo modo ?* « de quelle manière ? » / *quomodo ?* « comment ? » ; *et iam* « et déjà » / *etiam* « aussi » ; *animum aduertere* « prêter attention, tourner son esprit vers » / *animaduertere* « remarquer ».

1.3. Les abréviations :

Elles sont nombreuses dans les textes latins ; on peut citer, entre autres :

_ les prénoms : *M.* pour *Marcus*, *M'.* pour *Manius*, *L.* pour *Lucius*, etc. ;

_ les chiffres romains, bien connus, mais aussi les noms de mesure et de monnaie ;

_ les dates : *a.d. VI Kal. Ian. = ante diem sextum Kalendas Ianuarias* « le sixième jour avant les calendes de janvier » ;

_ les formules rituelles : *SPQR = Senatus Populusque Romanus* (marque de l'état romain) ; *V. = Vale* « porte-toi bien » (formule de salutation à la fin d'une lettre).

1.4. Les variantes morphologiques :

Les textes latins portent parfois la trace de l'évolution diachronique de la langue et des incertitudes que celle-ci fait peser sur la morphologie flexionnelle. Cela est particulièrement vrai dans les textes préclassiques tels les comédies de Plaute ; mais la présence de variantes n'est pas exclue non plus ultérieurement.

Ainsi voit-on un même texte hésiter entre la forme classique de génitif singulier *pater familiae* (comme *rosae*) et sa forme archaïque *pater familias* (dont la pérennité ne s'est pas démentie) ; ou entre deux formes concurrentes d'ablatif singulier pour les noms dont le thème se termine par un *–i* : *igni* / *igne* « par le feu » ; ou encore entre deux formes de subjonctif présent du verbe *être* : *sit* / *siet* « qu'il soit » ou deux formes du subjonctif parfait du verbe *faire* : *fecerit* / *faxit* « qu'il ait fait ».

On le voit, les difficultés rencontrées ne sont pas vraiment spécifiques au latin : d'une part les dialectologues et les historiens du français auront reconnu des phénomènes qui leur sont familiers ; d'autre part, les francisants qui travaillent sur la langue moderne auraient sans doute intérêt à s'interroger davantage qu'ils ne le font généralement sur l'existence de telles variations dans les textes qu'ils étudient. Car, sauf à travailler uniquement sur des textes de facture extrêmement classique, on rencontre nécessairement des graphies oralisantes qui tentent de reproduire la parole d'un personnage et le heurt de ses mots, que celui-ci se traduise par des élisions ou, au contraire, par des liaisons appuyées ; et plus on intègre dans son corpus des romans contemporains, plus le procédé devient systématique (cf. Annie Saumont ou Jean Echenoz, par exemple ; mais Jean Giono ne s'en privait pas non plus). Et si d'aventure on quitte la littérature pour entrer dans le monde journalistique ou, mieux encore, pour télécharger des textes via l'internet, surgissent alors une multitude d'abréviations et de sigles qui sont autant de variantes de lexèmes ou de syntagmes complets, ainsi que des variantes orthographiques et morphologiques qui, pour n'être pas reconnues par l'Académie ni par l'école républicaine, n'en affectent pas moins la qualité des corpus étudiés. Plutôt que de passer outre, il nous semble préférable de se donner des moyens efficaces pour prendre en compte cette réalité.

2. Lemmatisation et traitement de la variation

La lemmatisation permet précisément de dégager les analyses de corpus des aléas de la variation tout en conservant la forme originelle et authentique des textes : elle satisfait donc à la fois aux exigences de la statistique linguistique (en particulier dans le cadre de la comparaison de textes) et à celles de la philologie. Il suffit de garder présent à l'esprit le principe fondamental que tout traitement de lemmatisation doit respecter la graphie d'origine et permettre de la retrouver à tout moment ; il ne s'agit donc pas de transformer le texte, de le lisser au profit d'une norme dont on se demande d'ailleurs ce qu'elle pourrait être. Mais il faut, dans le même temps, pouvoir s'en affranchir.

La solution à cette double exigence réside dans la création d'un ***index alphabétique lemmatisé***. Chaque forme du texte est rapportée à son lemme (déterminé de préférence avec l'aide d'un dictionnaire et d'une grammaire de référence afin que les choix restent stables durant tout le traitement, quelle que soit sa durée et le nombre d'opérateurs qui y participent) et ce sont ces lemmes qui sont classés par ordre alphabétique, constituant ainsi un index (ou dictionnaire du texte) ; sous chaque entrée de cet index sont rassemblées les différentes formes qui s'y rapportent, avec leur graphie d'origine et leur référence précise dans le texte. Tel est le fichier fondamental de la base, le fichier charnière, le médiateur qui permet de circuler rapidement et en toute sécurité du dictionnaire de référence au texte, du lemme à la forme, du standard à ses variantes.

Les quelques exemples suivants, empruntés à l'index de l'œuvre de Caton, vont illustrer la description précédente en reprenant quelques cas de variation évoqués au paragraphe 1 :

Premier exemple : le lemme* adueho *« amener »

Ce lemme a trois occurrences dans l'œuvre de Caton : aux paragraphes 22, 135 et 138 du *Traité de l'Agriculture*.

3 ADVEHO
aduexeris
A. 22, 4 ,22
aruectum erit
A. 135, 7, 29
aruehant
A. 138, 1, 8

Ces trois formes se différencient d'abord par leur flexion puisqu'il s'agit de formes conjuguées ; et l'on peut voir que la conjugaison provoque des modifications formelles non négligeables ; elles se différencient ensuite par une variation dans la forme du préfixe : la forme normale *ad-* pouvait laisser place, surtout en parler rural, à une forme dissimilée *ar-* ; comme on le voit, Caton admet les deux prononciations et les retranscrit fidèlement. Se révèle alors pleinement la vanité des subterfuges qu'ont longtemps représenté, dans les bases non lemmatisées, les caractères jokers tels que l'étoile ou le dollar : le mot connaissant ici des variantes lourdes sur sa finale et une variante qui touche aussi sa deuxième lettre et qui éloigne considérablement ses diverses occurrences l'une de l'autre dans l'ordre alphabétique, il est impensable de pouvoir récupérer de telles formes comme occurrences d'un même vocable en dehors d'une véritable lemmatisation.

Deuxième exemple :* quomodo *« comment »

5 QVOMODO adverbe relatif
quo modo
A. 94, 1, 7
A. 142, 1, 21
A. 142, 1, 41
A. 142, 1, 46
quomodo
D. 252, 2, 5
5 QVOMODO adverbe interrogatif
quo modo
A. 2, 1 ,25
A 116, 1, 2
A. 151, 1, 3
quomodo
A. 83, 1, 75
A. 154, 1, 5

La lemmatisation permet ici, non seulement de résoudre le problème de la variante graphique signalé plus haut (un mot ou deux mots), mais encore de régler celui de l'homonymie entre deux adverbes aux fonctions syntaxiques différentes. Ainsi l'index offre au chercheur la possibilité de formuler toutes les requêtes envisageables, depuis la plus spécifique jusqu'à la plus englobante : un logiciel d'interrogation pourra en effet aisément proposer de relever les occurrences de *quomodo* en un seul mot et à fonction d'adverbe interrogatif (deux occurrences) aussi bien que celles du lemme *QVOMODO* adverbe relatif (cinq occurrences) ou encore celles de la forme *quo modo* en deux mots (sept occurrences), etc.

Exactement de la même façon sont enregistrées les abréviations sous leur lemme de référence ainsi que les variantes morphologiques ; on peut même ajouter les variantes éditoriales si besoin est.

Une comparaison entre les textes (calcul de distance intertextuelle) qui souhaite prendre en compte les occurrences des vocables employés sans être soumise aux distorsions des variations de

formes pourra bien sûr construire ses index fréquentiels et ses tableaux de contingence à partir de l'index lemmatisé.

On voudrait, pour terminer, noter un autre avantage de cette lemmatisation et de l'étiquetage grammatical qui l'accompagne : dans la base de données du L.A.S.L.A., toutes les recherches lexicologiques peuvent être croisées avec un paramètre grammatical, en particulier avec une contrainte sur la classe de mots : ainsi non seulement peut-on chercher, par exemple, toutes les occurrences des lemmes terminés par *–ilis* (sans avoir à se soucier de l'impact de la déclinaison qui peut produire des formes en *–ilem, -ile, -ili, ilibus*, etc.), mais encore peut-on restreindre la recherche aux seuls adjectifs en *–ilis*. De même peut-on créer un index fréquentiel totalement fiable (aucune occurrence oubliée, aucun homonyme confondu), mais encore peut-on créer des sous-index selon la classe de mots : on obtiendra ainsi sans difficulté les 30 substantifs les plus fréquents d'une œuvre ou ses 50 premiers verbes – et eux seuls. Les calculs de spécificités et les analyses thématiques s'en trouvent affinées, si on le souhaite.

On aura compris que nous défendons avec ferveur la lemmatisation des corpus informatisés. Pour nous qui travaillons sur la distribution et le signifié des catégories grammaticales, elle n'est pas un luxe inutile, elle est une nécessité. Elle ouvre en effet des champs d'exploration qui, sinon, restent inaccessibles, elle affine et stabilise les traitements quantitatifs. Pour l'ensemble de la communauté linguistique, elle devrait apparaître de plus en plus comme un des critères de réutilisabilité des corpus, à côté de l'annotation documentaire et du choix des formats : car, par les traitements préalables qu'elle impose, elle garantit la qualité du corpus, sa fiabilité et donne accès à des données clairement définies qui seules pourront fournir la base commune à diverses études comparatives. Nous sommes consciente néanmoins qu'elle nécessite un gros investissement dont la rentabilité doit toujours être évaluée en fonction des besoins et des projets de recherche.

Références bibliographiques :

BRUNET, É. (2000) : « Qui lemmatise dilemme attise », dans L. JOSE et A. THEISSEN (éds.), *Scolia*, n°13 (Actes des 11èmes rencontres linguistiques en pays rhénan), Strasbourg, pp. 7-32.

ÉVRARD, É. & MELLET, S. (1998) : « Les méthodes quantitatives en langues anciennes », *Lalies* n°18, pp. 111-155.

LABBE Dominique, (1990), *Normes de saisie et de dépouillement des textes politiques*, Grenoble, Cahier du CERAT.

MELLET, S. (1996) : « Les atouts de la lemmatisation », dans G. MORACCHINI (éd.), *Bases de données linguistiques : conceptions, réalisations, exploitations* (Actes du colloque international de Corte), Corte, pp. 309-316.

PURNELLE, G. (1988) : *Cato, De Agricultura, Fragmenta omnia seruata : Index verborum, liste de fréquence, relevés grammaticaux*, Liège : CIPL, série du L.A.S.L.A. n° 15.

PURNELLE, G. (1996) : « Utilisation d'une banque de données de textes latins lemmatisés et analysés. Problèmes spécifiques aux données linguistiques », dans G. MORACCHINI (éd.), *Bases de données linguistiques : conceptions, réalisations, exploitations* (Actes du colloque international de Corte), Corte, pp. 295-307.