Gérald PURNELLE
Université de Liège
LASLA

UTILISATION D'UNE BANQUE DE DONNÉES DE TEXTES LATINS LEMMATISÉS ET ANALYSÉS.
PROBLÈMES SPÉCIFIQUES AUX DONNÉES LINGUISTIQUES

Mon propos, au cours de cette communication, sera de vous présenter conjointement deux réalisations du LASLA, le Laboratoire d'Analyse Statistique des Langues Anciennes de l'Université de Liège. Il s'agit d'une part de sa banque de données de textes latins classiques lemmatisés et analysés, et ensuite du CD-Rom sur lequel sera très prochaine diffusé cette banque de données.

Je présenterai d'abord brièvement notre Laboratoire. Le LASLA a été fondé à l'Université de Liège en 1961 par Louis Delatte et son directeur actuel est M. Joseph Denooz. Comme son nom l'indique, ce laboratoire s'était fixé pour but, dès ses débuts, l'étude des textes grecs et latins au moyen des méthodes statistiques. Conjointement, ses membres ont très tôt entrepris d'appliquer à ces textes les techniques informatiques, qui commençaient tout juste, à l'époque, à entrer dans l'usage du domaine philologique. C'est ainsi que des programmes de lemmatisation et d'analyse morphologique ont été développés, initiale-

ment par M. Etienne Evrard, et appliqués à divers textes des deux langues. C'est au moyen de ces procédures que s'est progressivement constituée notre banque de données.

Je dirai quelques mots de cette procédure appelée lemmatisation, importante pour la compréhension de la suite de cette présentation. "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 *fero* sera, selon les cas, le verbe *fero*, "porter" ou l'adjectif *ferus*.

Je n'insiste pas maintenant sur l'extrême utilité de la lemmatisation dans le cas du latin : Mlle Sylvie Mellet en parle ici même. Je vais toutefois décrire brièvement les procédés de lemmatisation employés au LASLA.

Tout d'abord, il faut savoir que la procédure de lemmatisation que nous utilisons pour le latin permet également de produire pour chaque forme son analyse morphologique complète. L'ajout, par des procédés automatiques, de ces informations, accroît spectaculairement la richesse de nos données, déjà assurée par la lemmatisation.

Pour lemmatiser les formes d'un texte, nous utilisons au LASLA deux procédures différentes, qu'il n'est pas inutile de décrire brièvement. La méthode utilisée est plus simple et plus sommaire pour le grec ancien, plus riche et complexe pour le latin et le français.

Pour le grec, tout texte est lemmatisé au moyen d'un lexique comprenant les formes déjà connues du système. Chaque forme est accompagnée de son lemme, enregistré une fois pour toutes, et d'un seul code de catégorie grammaticale. Etant donné le faible taux d'amphibologie du grec (comparé au latin), les formes donnant lieu à plusieurs lemmatisations sont relativement rares; la tâche du philologue qui examine les résultats de la procédure consiste donc essentiellement en une vérification. Par ailleurs, les formes qui n'ont jamais été rencontrées et enregistrées précédemment ne sont pas lemmatisées, et le philologue doit le faire manuellement. Le lexique est progressivement enrichi au moyen des formes nouvelles. Il permet actuellement de lemmatiser automatiquement plus de 90 % des formes d'un texte grec en prose.

La méthode utilisée pour le latin est toute différente : la richesse de l'information produite automatiquement est plus grande, mais le

travail demandé au philologue également. Le programme de lemmatisation et d'analyse consulte, pour chaque forme, deux fichiers : un lexique de radicaux et une table de désinences. Chaque radical est codé selon sa morphologie, de même que chaque désinence. Sommairement, la procédure peut se décrire comme suit. Chaque forme est progressivement découpée en deux parties depuis sa fin; pour chaque segment final ainsi isolé la table des désinences est consultée; si le segment est identifiable comme désinence, le reste du mot (le premier segment) est recherché dans le lexique des radicaux; s'il s'y trouve et porte des codes morphologiques compatibles avec la désinence, une lemmatisation et une analyse morphologique complète sont produites. Cette méthode permet d'atteindre une information maximale, tout en correspondant bien au haut degré d'amphibologie de la langue latine. Cependant, ces deux facteurs (l'ambiguïté de la langue et la richesse de l'information grammaticale produite) font que pour la plupart des formes d'un texte le programme propose plusieurs lemmes et analyses. C'est au philologue qu'il revient, pour chaque mot du texte, de choisir le bon lemme et la bonne analyse parmi les propositions de l'ordinateur. Etant donné les caractéristiques lexicologiques, morphologiques, syntaxiques et stylistiques du latin, il est actuellement impossible de développer une méthode de levée automatique des ambiguïtés pour cette langue. La tâche humaine est donc plus lourde, mais c'est à ce prix que l'on peut s'assurer tous les avantages de la lemmatisation et tout le gain d'une analyse morphologique complète.

On voit tout l'intérêt qu'il y a à disposer de fichiers de textes lemmatisés - je n'y reviens pas.

J'en termine avec cette description des procédures en signalant que, dans notre système, chaque verbe subordonné est marqué par un code indiquant son type de subordination ou la nature du subordonnant qui l'introduit. Cette opération de marquage est strictement manuelle.

En soumettant un texte latin à cette méthode, on obtient finalement un fichier séquentiel, dont chaque enregistrement correspond à un mot du texte et présente sa forme, son lemme, son analyse morphologique complète et son éventuel code de subordination. Les réalités morphologiques codées sont (selon les formes) la catégorie grammaticale, la sous-catégorie, le cas, le nombre, le degré, la voix, le mode, le temps, la personne.

J'ajouterai un dernier point, important. Une des caractéristiques du latin est de présenter des vocables dont les lemmes sont homo-

graphes. Deux exemples parlants : il y a dans le dictionnaire deux mots *LABOR*; l'un est un substantif et signifie "travail", l'autre est un verbe et signifie "glisser"; il y a deux mots *POPVLVS*, tous deux substantifs; l'un signifie "le peuple", l'autre "le peuplier". Pour distinguer ces lemmes homographes nous utilisons des indices chiffrés, qui affectent les lemmes homographes dans les enregistrements. Ces indices sont générés automatiquement par la procédure de lemmatisation, au même titre que les lemmes. Ainsi, l'accusatif *laborem* sera lemmatisé *LABOR 1*, mais le verbe *labitur* sera placé sous *LABOR 2*.

Telles sont donc les informations produites et conservées pour chaque mot d'un texte latin. Depuis longtemps déjà, il était possible d'exploiter en de nombreuses directions la banque de données que forment les textes latins lemmatisés au LASLA.

Il suffit, en réalité, de quelques programmes informatiques pour "interroger" notre banque de données selon un nombre important de points de vue. Il est ainsi possible de consulter cette banque 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. Il est également possible de cumuler ces recherches et de produire tous les contextes où deux critères coexistent. Chacune de ces interrogations peut faire l'objet de dénombrements, et ceux-ci peuvent être comparés. Ce ne sont là que quelques exemples des virtualités de nos fichiers. Le nombre de philologues et de linguistes qui y ont recours n'a fait que croître ces dernières années.

Parvenu à ce stade de ces activités, et considérant l'ampleur de la collection de textes latins ainsi constituée, le LASLA a décidé, il y a deux ans, de profiter des récents développements de la micro-informatique, afin d'améliorer significativement la façon dont il met cette information à la disposition de la communauté scientifique. Un accord a été passé avec les Editions Hachette, qui se sont chargées de la réalisation, la fabrication et la diffusion du CD-Rom. Au terme d'une ample entreprise d'harmonisation de la banque de données, le LASLA est aujourd'hui en mesure de publier un volumineux corpus de textes latins lemmatisés et analysés.

Ce corpus comprend 1 276 023 mots; il couvre 75 oeuvres de 18 auteurs différents. Le plus ancien est Caton, dont les oeuvres sont datées du début du IIe siècle av. J.-C. Le plus récent est Ausone, du IVè siècle. Certains auteurs sont intégralement repris dans le corpus,

d'autres n'y figurent que partiellement.

Je vais maintenant décrire en quelques mots les différentes fonctionnalités du logiciel qui accompagnera le CD-Rom et permettra de l'exploiter. Pour chacune, je donnerai un ou plusieurs exemples susceptibles d'illustrer à la fois l'usage que la recherche peut faire de nos données et l'utilité du procédé de lemmatisation.

L'utilisation du logiciel comprend trois étapes principales : la définition d'un corpus de recherche, la définition de l'objet de la recherche, l'examen des résultats. Je décrirai surtout la seconde. Un mot de la première : il est possible de choisir, pour une recherche, tout le corpus, un seul auteur, une seule oeuvre, plusieurs oeuvres de plusieurs auteurs. Chaque corpus défini, c'est-à-dire chaque liste d'oeuvres à exploiter, peut être sauvegardé.

Le logiciel permet deux types de requêtes : la requête simple et la recherche combinée. La première ne porte que sur un seul objet, la seconde sur deux ou trois objets dont on recherche la cooccurrence. Avant de définir les possibilités de la seconde, je m'intéresserai d'abord aux objets de recherche eux-mêmes.

Chaque élément de l'ensemble d'informations associées à une forme textuelle, telles que je les ai décrites, peut faire l'objet d'une recherche : la forme, le lemme, les codes morphologiques, le code de subordination. Ceci permet d'exploiter toutes les virtualités de la base de données.

Il est possible d'exploiter le champ "lemme" des enregistrements (ou plus précisément l'index qui en est extrait) afin d'obtenir toutes les occurrences d'un lexème, quelles que soient les formes qui le représentent dans le corpus. Dans l'interface développé, il s'agit de remplir un champ "lemme", ce qui provoque le déroulement d'une liste alphabétique de tous les lemmes attestés sur le CD-Rom. Dans cette liste, les lemmes homographes sont accompagnés d'une explicitation propre à distinguer leurs sens. L'utilisateur peut choisir un seul lemme ou tous les homographes.

Ce genre de requête intéressera les études lexicologiques ou même historiques, la recherche de passages parallèles, mais aussi la réflexion linguistique, si le lemme est, par exemple, un subordonnant, une préposition, un adverbe, une conjonction. Donnons rapidement quatre exemples qui illustrent les différentes utilités de la lemmatisa-

tion. Chercher le lemme *FERO* permet de pallier la multiplicité des radicaux morphologiques. En cherchant le lemme *LEX* on élimine tout bruit dans les résultats, puisque sont écartées les occurrences de formes *legis*, *legi* ou *lege* appartenant au verbe *LEGO*. En ce qui concerne les lemmes homographes, voici trois exemples : *LABOR* substantif, *POPVLVS* "le peuple", *QVANDO* adverbe indéfini.

On peut, par ailleurs, élargir une requête de type lexicologique en affectant l'objet de recherche d'une troncature : on cherchera ainsi tous les mots commençant par une même chaîne de caractères. Je donnerai deux exemples : tous les mots composés du même préverbe (*inter-*), tous les lemmes tirés d'une même racine (pour autant que celle-ci n'ait pas subi de modifications phonétiques ou graphiques au cours des processus de dérivation) (*COMMVN-*, *ITAL-*, *DVLC-*). On voit l'usage que peuvent faire de cette option les recherches portant sur la sémantique ou la dérivation.

Au moyen d'une troncature ménagée au début de la chaîne de caractères demandée, il est également possible d'effectuer une requête à partir d'une fin de lemme ou de forme. On cherchera, par exemple, tous les noms en *-tudo* ou en *-mentum*, tous les adjectifs en *-ilis*, tous les verbes en *-sco*, tous les diminutifs en *-ulus*. Il n'est pas nécessaire d'insister longuement sur l'utilité de cette option pour les études linguistiques, qu'elles soient lexicologiques ou morphologiques. Je cite un dernier exemple, que j'utiliserai encore plus loin : en cherchant tous les lemmes s'achevant en **iter*, on obtiendra tous les adverbes formés de ce suffixe, mais également un certain bruit, qui contiendra notamment les substantifs *iter*, le chemin, ou *Iuppiter*, Jupiter. Nous verrons plus loin que le logiciel permet d'affiner la recherche.

Le deuxième type d'objet est la forme, telle qu'elle apparaît dans le texte. On renonce donc ici à recourir aux avantages fournis par la lemmatisation. Il faut signaler qu'une requête de ce type pourra fréquemment produire des occurrences appartenant à plus d'un lemme. La recherche d'une forme *labor* donnera des occurrences de nominatifs, mais aussi d'indicatifs présents; *obitum* peut être substantif ou participe, etc. Les résultats produits seront donc souvent constitués d'occurrences de nature ou de sens différents. Il y aura des cas où ce phénomène ne se produira pas (p. ex. la forme *dii*), mais d'autres résultats seront plus ambigus.

Signalons, enfin, que les mêmes options de troncature existent pour la forme comme pour le lemme : il est possible de rechercher

toutes les formes présentant un même début ou une même fin.

Le quatrième type d'objet est relatif à la subordination; il permet d'obtenir tous les verbes relevant d'une certaine structure syntaxique. L'utilisateur choisit le type de subordination qui l'intéresse dans une liste exhaustive constituant un menu déroulant. Quelques exemples : les verbes de subordonnées au subjonctif paratactique, de propositions infinitives ou à l'ablatif absolu, ou ceux qui sont introduits par un subordonnant déterminé : *dum*, *qualis* interrogatif, etc.

Comme celui que nous venons d'examiner, le cinquième type d'objet est sans doute de nature à intéresser davantage encore un public de linguistes et de grammairiens. Il s'agit de l'ensemble des critères grammaticaux qui sont enregistrés, pour chaque mot, dans la banque. Non seulement les catégories grammaticales sont accessibles, mais aussi les cas, les modes, les temps et même les voix, les personnes et le degré des adjectifs et des adverbes. On trouvera donc dans notre banque toutes les occurrences de l'ablatif ou de l'indicatif présent dans une texte ! Ces critères sont combinables; on formulera ainsi des requêtes plus ou moins fines, portant sur des objets strictement grammaticaux et dont les résultats seront des formes de lemmes variés. Quelques exemples : rechercher tous les substantifs, ou tous les mots au nominatif est possible, mais les résultats seront cyclopéens ! Par contre, la combinaison de critères produira des résultats plus affinés : les substantifs de la 5e déclinaison (combinaison de catégorie et sous-catégorie), les gérondifs au génitif (mode et cas), les numéraux distributifs (catégorie et sous-catégorie), l'impératif futur (mode et temps), ou même les indicatifs présents déponents 2e personne du singulier (combinaison de mode, temps, voix, personne et nombre).

A ce stade de la formulation d'une requête, le choix de l'objet de recherche, une autre option du logiciel s'avère très utile. Il est loisible à l'utilisateur de *préciser* l'objet initialement déterminé (un lemme, une forme ou une subordination) par un ou plusieurs critères grammaticaux et donc de *restreindre*, ici aussi, l'ampleur des résultats en réduisant, éventuellement, la part d'occurrences non pertinentes. Quelques exemples illustreront mieux cette nouveauté : on peut, après avoir choisi un lemme *locus*, limiter la recherche à toutes ses occurrences au pluriel; on peut ne chercher que les occurrences du verbe *amo* au passif, sélectionner, parmi les verbes introduit par *ut*, ceux qui sont au subjonctif plus-que-parfait.

Cette option permet également d'éliminer, au moins partielle-

ment, le bruit que peut entraîner une requête plus simple : si, comme tout à l'heure, je suis à la recherche des adverbes en *-iter*, il suffit de préciser cet objet de recherche (lemme avec troncature avant) par le critère morphologique catégorie = adverbe.

Elle permet également d'obtenir des résultats identiques en formulant des requêtes différentes. Si je cherche toutes les occurrences du génitif *amicorum* pour autant qu'elles soient substantif, en excluant les adjectifs, je puis soit demander le lemme *AMICVS* substantif et le préciser par les critères génitif pluriel, soit rechercher la forme *amicorum* en la précisant par la catégorie substantif.

La recherche de formes alternatives pour une même analyse est également facilitée par l'utilisation de critères morphologiques. Ainsi, parmi d'autres, le substantif *ignis* présente un ablatif singulier en *-i* ou en *-e*. En précisant la recherche du lemme par le critère ablatif singulier, on obtient les deux formes, sans devoir effectuer deux recherches de formes (*igni* et *igne*).

J'en viens maintenant à l'autre type de recherche, la recherche combinée. Elle permet de rechercher tous les contextes où apparaissent conjointement deux ou trois objets de recherche. Ceux-ci peuvent prendre n'importe quelle forme, toutes les combinaisons sont permises : deux lemmes, un lemme et un critère grammatical, deux réalités grammaticales.

Une fois les deux objets définis, il est nécessaire de faire certains choix touchant l'extension du contexte autorisé. On peut ne chercher que les cooccurrences dans une même phrase ou autoriser la séparation des deux objets. On peut, et même on doit dans le second cas, déterminer une distance maximale autorisée entre les deux objets; elle est formulée en nombre de mots. Au besoin, le premier objet défini sera obligatoirement situé avant le second. Cinq exemples suffiront à illustrer la recherche combinée. Le premier concerne une recherche liée au contenu sémantique des textes (lemme *philosophia* et lemme *sapiens* dans un contexte de 40 mots maximum). Le deuxième et le troisième relèvent des études syntaxiques (lemme *impero* et syntaxe *ut* + subjonctif dans la même phrase; *in* + accusatif). Dans le quatrième il s'agit d'une recherche de formule (forme *da* et forme *ueniam*, même phrase, écart d'un mot, même ordre). Le cinquième se rapporte à la syntaxe et à la pragmatique (critères morphologiques interjection, vocatif et impératif, même phrase).

Dès que la recherche (simple ou combinée) définie est exécutée,

le logiciel présente à l'utilisateur les résultats obtenus. Il affiche la liste complète des références des contextes repérés, qui précise, pour chaque occurrence, le lemme et la forme exacts, ainsi que la référence complète (auteur, oeuvre, livre, chapitre, vers, paragraphe ou ligne). L'utilisateur peut la parcourir et demander, à tout moment, l'affichage du contexte d'une occurrence.

Cet affichage prend la forme d'une page d'écran où apparaît la portion du texte complet dans laquelle figure le contexte concerné. Afin d'enrichir cette opération, nous avons choisi d'exploiter au maximum les informations contenues dans la banque. Il est possible de parcourir le texte et d'afficher à la demande, en cliquant sur 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).

Il y a évidemment des possibilités d'impression et d'exportation.

Vous aurez vu, je l'espère, que le logiciel au moyen duquel on consultera la base de données permet d'en exploiter toutes les informations. Il me reste à signaler que, dans une future version du CD-Rom, nous ferons figurer d'autres données ; il s'agira de tableaux statistiques présentant la fréquence de divers phénomènes lexicologiques ou grammaticaux. Chaque réalité morphologique y sera dénombrée, par oeuvre, par auteur et dans le corpus total.

Annexes

Opera Latina : Banque de textes latins lemmatisés du LASLA
(Université de Liège) [G. Purnelle]. Listage à sélecter

3 1 496 1 ORIS	ORA		11N00		1 6199
	ORA		11O00		2
	OS	1 13D00			3
3 1 496 2 HONOS	HONOR		13A00		1 6200
	HONOR		13B00		2
3 1 496 3 PRIMUM	PRIMVM		3N000		1 6201
	PRIMVS		3KA00 6		2
	PRIMVS		3KC00 5		3
3 1 496 4 ET	ET	1 60000			1 6202
	ET	2 81000			2
3 1 496 5 MULTIS	MVLTA	1 11N00			1 6203
	MVLTA	1 11O00			2
	MVLTI		12N00		3
	MVLTI		12O00		4
	MVLTVS		21N00 1		5
	MVLTVS		21O00 1		6
3 1 496 6 OPTATA	OPTATVM		12J00		1 6204
	OPTATVM		12L00		2
	OPTATVS		21A00 2		3
	OPTATVS		21F00 2		4
	OPTATVS		21J00 6		5
	OPTATVS		21L00 6		6
	OPTO		5AA44 2		7
	OPTO		5AF44 2		8
	OPTO		5AJ44 6		9
	OPTO		5AL44 6	10	
3 1 496 7 LABELLA	LABELLVM	1 12J00		0	1 6205
	LABELLVM	1 12L00		0	2
	LABELLVM	2 12J00		0	3
	LABELLVM	2 12L00		0	4
3 1 497 1 ET	ET	1 60000			1 6206
	ET	2 81000			2
3 1 497 2 PATULAE	PATVLVS		21D00 2		1 6207
	PATVLVS		21E00 2		2
	PATVLVS		21J00 2		3
3 1 497 3 FRONTIS	FRONS	1 13D00			1 6208
3 1 497 4 SPECIES	SPECIES		15A00		1 6209
	SPECIES		15J00		2
	SPECIES		15L00		3
3 1 497 5 CONCRESCERE		CONCRESCO			53071
	1 6210				
3 1 497 6 IN	IN		70300		1 6211
	IN		70600		2
3 1 497 7 UNUM	VNVM		12A00		01 6212
	VNVM		12C00		02
	VNVS		31A00 6	0	3
	VNVS		31C00 5	0	4
3 1 498 1 COEPERE	COEPIO		55L14		1 6213
	COEPIO		55071		2
3 1 498 2 ET	ET	1 60000			1 6214
	ET	1 81000			2
3 1 498 3 GRACILI	GRACILIS		24E00 1	1	6215
	GRACILIS		24F00 1	2	
3 1 498 4 MENTUM	MENS		13M00		1 6216
	MENTVM		12C00		2
3 1 498 5 PRODUCERE	PRODVCO		53071		1 6217
3 1 498 6 ROSTRO	ROSTRVM		12E00	0	1 6218
	ROSTRVM		12F00	0	2

Fichier LASLA

3 1 496 1 ORIS	OS	1 13D00		6199
3 1 496 2 HONOS	HONOR		13A00	6200
3 1 496 3 PRIMUM	PRIMVM		3N000	6201
3 1 496 4 ET	ET	2 81000		6202
3 1 496 5 MULTIS	MVLTI		12O00	6203
3 1 496 6 OPTATA	OPTO		5AJ44 6	6204
3 1 496 7 LABELLA	LABELLVM	1 12J00	0	6205
3 1 497 1 ET	ET	2 81000		6206
3 1 497 2 PATULAE	PATVLVS		21J00 2	6207
3 1 497 3 FRONTIS	FRONS	1 13D00		6208
3 1 497 4 SPECIES	SPECIES		15J00	6209
3 1 497 5 CONCRESCERE	CONCRESCO		53071	6210
3 1 497 6 IN	IN		70300	6211
3 1 497 7 UNUM	VNVM		12C00 0	6212
3 1 498 1 COEPERE	COEPIO		55L14&	6213
3 1 498 2 ET	ET		2 81000	6214
3 1 498 3 GRACILI	GRACILIS		24F00 1	6215
3 1 498 4 MENTUM	MENTVM		12C00	6216
3 1 498 5 PRODUCERE	PRODVCO		53071	6217
3 1 498 6 ROSTRO	ROSTRVM		12F00 0	6218

Codage du CD-Rom

VA 0109OS	1oris	496	A341
VA 0109HONOR	honos	496	A311
VA 0109PRIMVM	primum	496	D5 2
VA 0109ET	2et	496	S
VA 0109MVLTI	multis	496	A262
VA 0109OPTO	optata	496	B112 442
VA 0109LABELLVM	1labella	496	A212
VA 0109ET	2et	497	S
VA 0109PATVLVS	patulae	497	C112
VA 0109FRONS	1frontis	497	A341
VA 0109SPECIES	species	497	A512
VA 0109CONCRESCO	concrescere	497	B3 711
VA 0109IN	in	497	R
VA 0109VNVM	unum	497	A231
VA 0109COEPIO	coepere	498	B5 2 141300
VA 0109ET	2et	498	S
VA 0109GRACILIS	gracili	498	C461
VA 0109MENTVM	mentum	498	A231
VA 0109PRODVCO	producere	498	B3 711
VA 0109ROSTRVM	rostro	498	A261

Types de requêtes et exemples
lemme simple ¨ FERO
1er intérêt de la lemmatisation : rechercher toutes les formes d'un mot, sans devoir les énumérer, malgré une éventuelle variation graphique ou formelle parfois très importante (fer-, tul-, lat-). lemme simple ¨ LEX

2e intérêt de la lemmatisation : éliminer les formes homographes non pertinentes (demander LEX élimine les formes legis, legi et lege de LEGO qui apparaîtraient lors d'une recherche de formes dans un fichier non lemmatisé). lemme simple ¨ AFFIRMO
3e intérêt de la lemmatisation : supprimer les problèmes de variations graphiques (le lemme AFFIRMO permet d'atteindre tous les formes en adfirm- et en affirm-).
lemme(s) avec indice ¨ LABOR substantif

¨ POPVLVS "le peuple"
¨ QVANDO conjonction de subordination

4e intérêt de la lemmatisation : distinguer les lemmes homographes, qu'il s'agisse de mots étymologiquement différents (LABOR) ou d'emplois différents (catégories grammaticales : QVANDO). lemme avec troncature ¨ COMMVN*
Recherche de tous les lemmes appartenant à une famille étymologique.
lemme avec troncature devant ¨ *ITIO
¨ *ITER
Recherche de tous les lemmes formés d'un même suffixe.
forme simple ¨ dii

Recherche d'une forme précise.
¨ amici
Recherche d'une forme précise, abstraction faite de son analyse (ici gén. sg. ou nom. pl.) et de son appartenance à des lemmes à indices. forme avec troncature ¨ human*
forme avec troncature devant ¨ *iter
¨ *onis
Recherche d'un suffixe, indépendamment du lemme des formes.
syntaxe des propositions ¨ ablatif absolu
¨ DVM
¨ QVALIS interrogatif
Recherche des verbes subordonnés d'une certaine fonction ou introduits par un certain subordonnant.
critère morphologique ¨ vocatif
¨ gérondif au génitif
¨ numéraux distributifs
¨ impératif futur
¨ indicatif présent déponent 2e personne du singulier
Recherche de toutes les formes correspondant à un ou plusieurs critères morphologiques (catégorie, sous-catégorie, cas, nombre, voix, mode, temps, personne, degré). précision morphologique

Objet de recherche	Précision ¨
lemme LOCVS	pluriel
lemme IGNIS	ablatif singulier
lemme AMICVS substantif	génitif pluriel ¨
lemme MAGNIFICVS comparatif	
lemme LEVITER superlatif ¨	
lemme SVM subjonctif imparfait ¨	
lemme tronqué SVPER* substantif ¨	
forme amicorum substantif ¨	

forme tronquée *im accusatif singulier
forme tronquée *onis substantif, génitif singulier
lemme tronqué *iter adverbe
verbes introduits par DVM subjonctif
verbes introduits par VT conjonction indicatif plus-que-parfait

La précision d'un objet par un ou plusieurs critères morphologiques permet de limiter des résultats d'une recherche en ne prenant en compte que les formes répondant à ces critères. Il est ainsi possible de rechercher une partie du paradigme d'un lemme (p. ex. le superlatif) ou de la rection d'un subordonnant (p. ex. DVM suivi de subjonctif); des formes alternatives (p. ex. abl. igni/igne); les emplois d'un verbe sous une certaine catégorie grammaticale (p. ex. forme amicorum substantif).
Recherches combinées

Premier objet	Deuxième objet	Contexte
lemme PHILOSOPHIA	lemmes SAPIENS	40 mots

Recherche thématique : recherche de la cooccurrence de deux termes dans un contexte limité.

lemme IMPERO	syntaxe VT + subj.	même phrase
lemme IN	accusatif	écart de 3 mots, même ordre

Recherche syntaxique : recherche de deux objets grammaticaux dans une même phrase.
forme da forme ueniam même phrase, écart un mot, même ordre
Recherche textuelle : recherche d'une expression précise (deux formes).

morpho : interjection	morpho : impératif	même phrase, écart 12 mots

Recherches cumulées
Premier objet : lemme QVANDO adverbe relatif
Deuxième objet : lemme QVANDO adverbe interrogatif
Troisième objet : lemme QVANDO conjonction de subordination