Collatinus, un outil polymorphe pour l'étude du latin

Collatinus est à la fois un lemmatiseur et un analyseur morphologique de textes latins. Il est donc capable, si on lui donne une forme déclinée ou conjuguée, de remonter au(x) lemme(s) dont elle peut venir. Il donne en même temps les différentes analyses morphologiques possibles. Sur le programme d'origine, destiné d'abord à des enseignants, se sont greffées de nouvelles fonctionnalités susceptibles d'intéresser aussi bien les utilisateurs non latinistes, désireux de comprendre un texte latin, que les chercheurs plus spécialisés. Collatinus permet de consulter des dictionnaires, de scander un texte ou de fléchir un lemme.

Collatinus est un programme libre, gratuit et ouvert[1]. Il est disponible en version résidente pour les principaux systèmes[2] dans la boîte à outils de l'équipement d'excellence Biblissima[3]. Pour qu'il soit également utilisable sur les appareils mobiles, nous en avons fait une version orientée vers le web[4]. Cette version web reprend les principales fonctions de la version résidente. Plus récemment, nous avons développé Eulexis, qui est le pendant grec de Collatinus. Ce dernier n'existe, pour l'instant, qu'en version web[5]. Comme Collatinus, il permet de lemmatiser une forme ou un texte et de consulter simultanément plusieurs dictionnaires. En revanche, la scansion est absente et la flexion se limite à énumérer les formes connues du lemme.

Lemmatisation et analyse morphologique

Deux approches très différentes permettent de construire un lemmatiseur-analyseur morphologique. La première, qui est aussi la plus courante, consiste à établir la liste de

---

[1] Collatinus fait partie des logiciels dits libres. Sa dernière version, numérotée 10.2, est placée sous licence GPL (General Public License). Cela signifie, en simplifiant, que son code source est obligatoirement accessible à celui qui l'acquiert, qu'il est librement copiable et modifiable, à condition que la copie ou le logiciel dérivé obtenu soit placé sous une licence compatible.

[2] Le programme a été développé entièrement en langage C++, et il utilise la bibliothèque graphique Qt. Il est développé sous GNU-Linux, mais il a été compilé avec succès sous Windows et Macintosh.

[3] L'équipex Biblissima, un observatoire du patrimoine écrit du Moyen Âge et de la Renaissance, a développé une boîte à outils dans laquelle ce programme s'est inséré très naturellement. Collatinus est disponible à l'adresse : http://outils.biblissima.fr/collatinus/

[4] La version web a été développée avec l'aide précieuse de Régis Robineau de l'équipe Biblissima. Elle est accessible sur le site : http://outils.biblissima.fr/collatinus-web/

[5] http://outils.biblissima.fr/eulexis/ développé avec Régis Robineau. Ce projet a bénéficié du soutien de Philipp Roelli, d'Eduard Frunzeanu et d'André Charbonnet (alias Chaeréphon).

toutes les formes utilisées dans un vaste corpus. Cette liste peut être établie par le traitement automatique de tous les textes que nous possédons maintenant au format numérique[6]. En face de chaque forme, il faut alors saisir ses différents lemmes et analyses. L'ensemble est placé dans une base de données. La lemmatisation et l'analyse seront alors une simple recherche de la forme, et l'affichage du lemme et de l'analyse qui lui correspondent. L'avantage de cette méthode est sa simplicité et sa robustesse[7]. Ses inconvénients sont d'une part le gros volume de données à explorer, et d'autre part son aspect très rudimentaire.

L'auteur de Collatinus a choisi une seconde voie, plus proche du fonctionnement d'un cerveau humain, qui consiste à décomposer une forme en deux éléments : le radical et la désinence, entendus, pour les besoins du programme, comme les parties constante et variable des formes fléchies[8]. Toute forme est coupée en deux de toutes les façons possibles. Le programme cherche alors les deux morceaux parmi les radicaux et les désinences et vérifie qu'ils peuvent aller ensemble. Au lieu de la liste de toutes les formes existantes, on ne doit saisir, par conséquent, que la liste des lemmes du lexique latin, chacun accompagné de son modèle de flexion, de ses radicaux (génitif pour les formes nominales, infectum, perfectum, supin pour les formes verbales), et de ses traductions en langue moderne. On enregistre ensuite, dans une seconde liste, toutes les désinences avec leurs caractéristiques (modèle, cas, genre, degré, etc.). La lemmatisation et l'analyse seront le résultat de la recherche d'une addition radical + désinence qui corresponde à la forme à analyser, et qui satisfasse à deux exigences :

– que le modèle du lemme soit le même que celui de la désinence ;
– que le type de radical soit celui demandé par la désinence.

Cette seconde méthode fait appel à des calculs un peu plus complexes, et oblige le programmeur à traiter séparément les formes spéciales[9], invariables ou irrégulières[10]. Cependant, elle manipule un ensemble de données beaucoup plus léger, qui tient très facilement en mémoire vive. La lemmatisation et l'analyse morphologique s'affichent dans une bulle d'aide quand le curseur s'arrête sur un mot du texte. Un simple bouton de la barre d'outils de Collatinus permet de lemmatiser l'ensemble d'un texte, contenu dans le cadre supérieur de la fenêtre (voir figure 1). Le cadre du bas contient le résultat des requêtes qui dépend de l'onglet choisi. Ici, c'est l'onglet « Lexiques » qui est actif et

---

[6] Voir par exemple l'article de Philipp Roelli dans ce volume.

[7] Cette méthode a été adoptée pour Eulexis ; sont utilisées les listes de formes et de lemmes établies par le projet Perseus (http://www.perseus.tufts.edu/) et exploitées également par Diogenes (https://community.dur.ac.uk/p.j.heslin/Software/Diogenes/).

[8] Ici, les notions de radical et de désinence ne concordent pas nécessairement avec celles des grammairiens.

[9] Les formes spéciales sont, par exemple, la forme du nominatif d'un nom de la 3ème déclinaison imparisyllabique. Cette forme est donnée dans le lexique (forme canonique), mais ne peut pas se déduire du radical du génitif.

[10] Il existe deux types de formes irrégulières. La première n'est qu'un double de la forme régulière. Les deux formes, la régulière et l'irrégulière, peuvent se rencontrer toutes deux, quelquefois même dans le même texte. L'ablatif *amni* coexiste avec *amne*. Les formes irrégulières du second type sont exclusives. La forme régulière correspondante est soit très rare, soit non attestée. Par exemple *fert*, troisième personne du singulier de l'indicatif actif de *fero, fers, ferre*. La forme *ferit* ne peut être lemmatisée qu'en *ferio*.

le cadre contient la lemmatisation de la dernière phrase du cadre supérieur. Les perspectives pédagogiques de la méthode choisie sont également très intéressantes, puisqu'il devient possible de mimer en partie le travail intellectuel de l'apprenti latiniste. De plus, la procédure de lemmatisation peut être inversée : à partir d'un lemme, l'ordinateur sait construire toutes les formes fléchies, même si elles ne sont pas attestées. On obtient le tableau de flexion d'un lemme en soudant tous les radicaux qui lui sont associés à toutes les désinences compatibles.

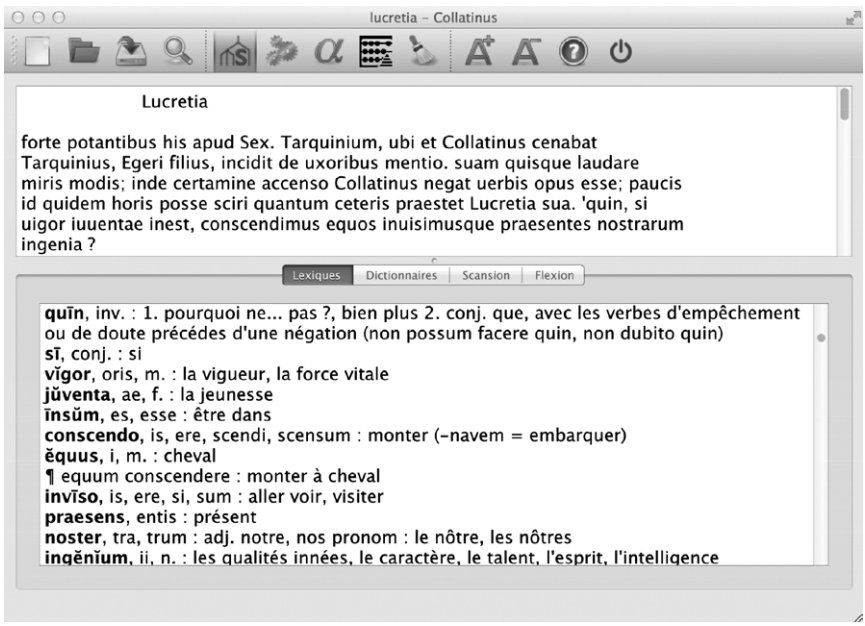

Figure 1 : Copie d'écran d'une lemmatisation avec Collatinus.

Depuis sa version 9, Collatinus tente de faire plus que la lemmatisation et l'analyse morphologique. Il est capable de détecter les expressions, c'est-à-dire l'utilisation conjointe de plusieurs lemmes générant un sens autre que celui de la simple addition des sens de chacun d'entre eux. Si par exemple, dans une phrase, il repère l'épithète *longam*, accordée en genre, nombre et cas avec le nom *nauem*[11], il propose le sens « navire de guerre ». Collatinus dispose d'un fichier contenant ces expressions, qui sont des listes de groupes de lemmes accompagnés des caractères morphologiques qu'ils doivent satisfaire. L'un des lemmes de l'expression est désigné comme celui qui par son occurrence, correspondant à la morphologie définie par le lexique, lancera la recherche des autres lemmes. Dans l'exemple de la figure 1, Collatinus nous signale par un ¶ qu'il a trouvé les mots clefs de l'expression « monter à cheval ».

[11] Le programme utilise des notations ramistes, c'est-à-dire qu'il distingue, comme la plupart des dictionnaires, les paires u/v et i/j. Il sait toutefois reconnaître une forme non-ramiste comme *nauem* et l'associe au lemme *navis*.

Consultation des dictionnaires

En même temps que la lemmatisation, Collatinus donne une brève traduction des lemmes. Six des principales langues européennes sont disponibles, qui ont été vérifiées par un relecteur[12]. La traduction dans d'autres langues a été introduite, mais elle a bénéficié des progrès de la traduction automatique sans avoir été, pour le moment, soigneusement vérifiée. De toute façon, il s'agit d'un résumé très court des sens possibles et souvent le lecteur éprouve le besoin d'en savoir plus. Nous avons donc introduit dans le programme la possibilité d'afficher l'article ou la page du dictionnaire correspondant au(x) lemme(s) de la forme demandée : il suffit pour cela de cliquer sur un mot du texte en ayant choisi l'onglet « Dictionnaires » pour le cadre inférieur. Un mot à chercher peut aussi être introduit dans la fenêtre de saisie sans qu'il apparaisse dans le texte. Une recherche littérale, sans lemmatisation, est également possible. Certains dictionnaires, comme celui de Ch. Lewis et Ch. Short[13], celui de K.E. Georges[14] ou celui de Ch. du Fresne sieur Du Cange[15], sont en texte pur et le programme affiche l'article demandé. D'autres, comme celui de F. Gaffiot ou celui de F. Calonghi, ne sont aujourd'hui disponibles qu'en mode image : c'est donc toute la page qui est affichée. Des boutons très intuitifs permettent de passer à l'article (ou à la page) qui suit ou qui précède.

Scansion

Dernièrement, Collatinus a appris les quantités des lemmes de son lexique ainsi que celles des désinences. Ainsi, il est capable, après avoir analysé une forme, de lui associer ses quantités. Éventuellement, lorsque plusieurs analyses sont possibles, on peut avoir plusieurs solutions scandées différemment. L'une d'elles, la première trouvée et pas nécessairement la bonne, est mise dans le texte et les autres suivent entre parenthèses. Les voyelles, longues ou brèves, sont repérées par les signes habituels « macron » (par exemple, ā) et « breve » (ă) que l'encodage des caractères en utf-8 permet d'afficher facilement[16].

Cela simplifie grandement l'approche de la poésie métrique, comme le montre l'exemple de la figure 2. Dans le premier vers, les deux brèves de *vĭrūmquĕ* imposent le choix des mots *Ārmă* et *cănō* pour donner deux dactyles (qui sont suivis de deux spondées puis d'un dactyle et du spondée final). Le programme est aussi attentif à la suite des mots et fera les élisions et les allongements usuels. Dans le deuxième vers de la figure 2, le a final du mot *Ītălĭām* (normalement la voyelle qui précède un m final est brève[17]) est

---

[12] Il s'agit du français, de l'anglais, de l'allemand, de l'espagnol, du catalan et du galicien.

[13] Charlton Lewis & Charles Short, *A Latin dictionary founded on Andrew's edition of Freund's Latin dictionary*, Oxford, 1879, balisé en XML par Perseus http://www.perseus.tufts.edu/

[14] Karl Ernst Georges, *Ausführliches lateinisch-deutsches Handwörterbuch*, Hannover, 1913, balisé en HTML par Philipp Roelli.

[15] Charles du Fresne sieur Du Cange *et al.*, *Glossarium mediæ et infimæ latinitatis*, Paris, 1883-1887, version XML de l'École des chartes, http://ducange.enc.sorbonne.fr/.

[16] Les voyelles communes ne figurent pas dans la panoplie de l'utf-8. Toutefois, elles s'obtiennent simplement en utilisant le « combining breve ». La séquence o-macron suivi d'un combining breve donne o͝, le o commun qui termine beaucoup de formes canoniques des verbes.

[17] Louis Quicherat, *Nouvelle Prosodie Latine*, Paris, 1885 (30e édition). À ne pas confondre avec le *Thesaurus Poeticus Linguae Latinae* du même auteur.

allongé car il est suivi de *fato*. Dans le troisième vers, on a deux élisions qui conduisent à *lītŏră, mūlt[um] īll[e] ēt tērrīs jāctātŭs ĕt āltō* (donc une structure DSSSDS de l'hexamètre). Collatinus fournit ainsi une aide à la scansion de toute poésie métrique, sans se limiter à la versification dactylique.

Collatinus scande aussi la prose, ce qui ouvre plusieurs pistes : scansion des clausules métriques et rythmiques, repérage de vers ou de portions de vers dans la prose. L'étude des clausules métriques ou rythmiques dans la prose latine est souvent jugée difficile car elle nécessite de déterminer les quantités de toutes les voyelles d'un texte, ou au moins la position de l'accent tonique de tous les mots. Pourtant l'usage du cursus ou de certaines clausules métriques peut être caractéristique d'un auteur et pourrait servir dans le cadre d'une critique d'attribution, par exemple. En permettant la scansion d'un texte en prose, Collatinus lève cet obstacle et simplifie la tâche de l'érudit.

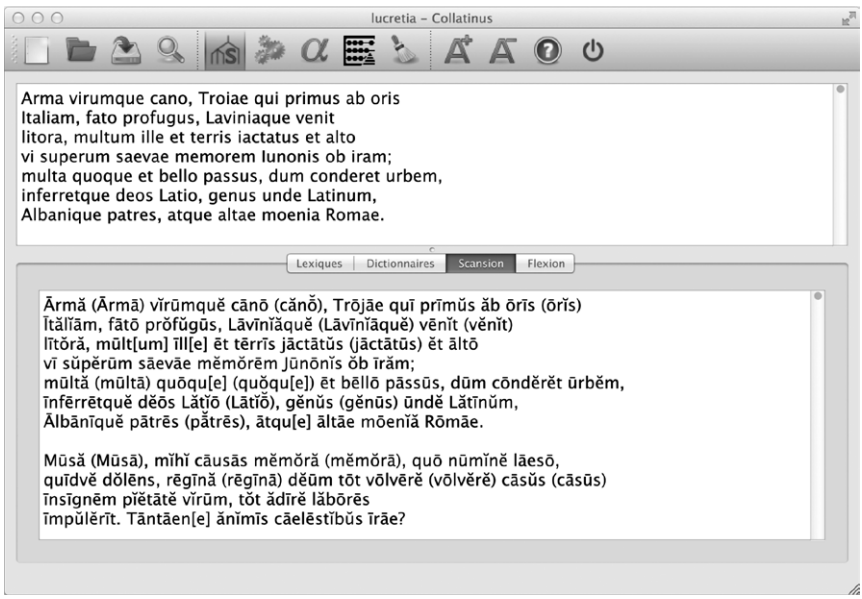

Figure 2 : copie d'écran d'un exemple de scansion avec Collatinus.

Perspectives

Plusieurs voies pour le développement de Collatinus sont envisagées. Chacune apporte son lot de difficultés qu'il faudra surmonter. Inévitablement, elles devront s'étaler dans le temps et l'intérêt qu'elles susciteront dans la communauté nous conduira à établir des priorités. Parmi les choses faciles à faire mais chronophages, nous mentionnerons l'enrichissement du lexique de Collatinus. En effet, ce lexique compte aujourd'hui un peu plus de 10 000 lemmes. Il permet de reconnaître environ 70 % des

formes rencontrées dans la poésie métrique étudiée par PedeCerto[18]. Bien évidemment, Collatinus achoppe surtout sur les formes rares et si l'on tient compte des occurrences de chaque forme, c'est près de 90 % des mots employés dans ce corpus qui sont identifiés. Toutefois, nous disposons aujourd'hui de dictionnaires comme le L&S qui contiennent beaucoup plus de lemmes. De plus, leur format se prête bien à un traitement automatique de l'information. Nous essayons donc d'extraire les lemmes de ces dictionnaires, de comprendre les formes canoniques souvent abrégées[19] et de convertir le tout au format de Collatinus. Une dernière étape de vérification par un latiniste sera nécessaire pour valider le lexique obtenu.

Dans sa version actuelle, Collatinus ne reconnaît que la graphie classique des formes. Or, le latin a évolué au travers des âges et les graphies médiévales s'écartent des formes classiques. On peut citer comme exemple simple la diphtongue ae qui s'est simplifiée en e. Remonter d'une forme médiévale à son parent classique n'est pas chose aisée, les vrais e ne devant pas être remplacés par des ae. En revanche, à partir d'une forme classique, il est facile de générer automatiquement des formes médiévales plausibles. Elles ne seront pas nécessairement attestées, mais si le programme les rencontre, il saura les reconnaître. L'inconvénient d'une telle méthode automatique est que l'on risque d'augmenter inutilement le « bruit » dans le processus de lemmatisation, c'est-à-dire que certaines formes pourraient être rattachées à un lemme sans que cela soit attesté. Cet écueil peut être évité si un philologue complète le lexique de Collatinus avec les graphies médiévales attestées. La lemmatisation y gagnerait en précision, mais cela nécessite une intervention humaine qui pourrait être fastidieuse.

On peut également formuler une critique : toutes les opérations de lemmatisation se font sans tenir compte du sens, élément déterminant chez le lecteur humain. Sans s'en rendre compte, notre lecteur humain élimine très tôt, en anticipant même, un grand nombre de solutions inacceptables pour lui. Des tentatives ont été faites par les spécialistes du TAL[20] afin de trouver un équivalent algorithmique à ces raccourcis. Les fonctions de désambiguïsation permettent, d'après le contexte, de choisir pour un mot le sens qui convient le mieux. Chez les médecins Celse ou Columelle, par exemple, *manus* désigne sans doute la main dans son sens anatomique. Dans un contexte militaire, ce sera un groupe de soldats, et chez un juriste, l'autorité du propriétaire. Ce contexte peut être défini soit automatiquement par l'occurrence de mots caractéristiques, soit manuellement par l'utilisateur. Il deviendrait possible, sinon d'éliminer, du moins de rétrograder dans l'affichage des sens d'un lemme les possibilités les moins cohérentes avec le contexte.

Collatinus est aussi tourné vers la formation des futurs latinistes. Pour l'instant, il ne s'attache qu'aux premières étapes de la lecture du latin, la lemmatisation et l'analyse morphologique. Le but qu'il doit évidemment rechercher est de pouvoir accompagner l'élève ou l'amateur le plus loin possible dans la compréhension d'un texte. Le programme

---

[18] La liste des formes avec leurs quantités nous a été gentiment communiquée par Emanuela Colombi et Luigi Tessarolo. Elle a été établie en scandant plus de 240 000 vers de la poésie dactylique dans le cadre du projet PedeCerto (http://www.pedecerto.eu:8080/pedecerto/) et nous a servi en particulier à vérifier la scansion faite par Collatinus.

[19] Par exemple, comment l'ordinateur peut-il construire de radical du génitif (*vīrgĭn*) à partir de l'information *virgo, ĭnis* ?

[20] Traitement Automatique des Langues. NLP en anglais : Natural Language Processing.

devrait être capable de proposer des solutions, ou de vérifier celles que propose l'utilisateur. Il doit pouvoir dire, par exemple, si cet adjectif-ci peut s'accorder avec ce nom-là, si ce datif-ci peut dépendre de ce verbe-là. Selon le but recherché, il s'agirait d'une analyse grammaticale et syntaxique de la phrase ou d'un dialogue homme-machine. Dans ce dernier cas, la traduction se construirait progressivement, à mesure que l'utilisateur identifie les groupes et les dépendances syntaxiques. Un programme qui réussirait à accompagner l'utilisateur, en lui posant les bonnes questions, en validant ou réfutant ses hypothèses, ressemblerait plus à un agent conversationnel. D'un côté, on aurait le texte latin et sa traduction en construction, de l'autre le dialogue avec un assistant, soit muni d'une batterie de cases à cocher et de listes déroulantes, soit tout simplement sous forme de questions et de réponses formulées dans un langage proche du langage naturel. Il y a là un défi à relever.

Collatinus est un outil polyvalent pour l'étude et/ou la compréhension du latin. Nous le mettons à la disposition de tous (élèves, étudiants, enseignants, chercheurs ou simples curieux) dans la boîte à outils de Biblissima. Collatinus permet à l'enseignant de préparer un texte avec le lexique nécessaire. Il offre au latiniste débutant une aide à la lecture en analysant les formes et en proposant les lemmes dont elles peuvent venir, associés à une traduction succincte. Avec ses dictionnaires intégrés, Collatinus permet à l'érudit de recouper les informations, où qu'il soit et sans qu'il ait à déplacer de volumineux ouvrages. La scansion des textes, qu'ils soient en vers ou en prose, ouvre des perspectives intéressantes pour le chercheur. Plusieurs pistes sont explorées pour affiner l'outil, mais tout cela dépendra aussi de l'accueil qu'il recevra et des attentes exprimées par ses utilisateurs.

Yves OUVRARD
Yves.Ouvrard@collatinus.org

Philippe VERKERK
Laboratoire de Physique des Lasers, Atomes et Molécules
UMR8523 CNRS - Université de Lille 1
Philippe.Verkerk@univ-lille1.fr

RÉSUMÉ. – Collatinus est un logiciel de lemmatisation et d'analyse morphologique de textes latins. Il est libre et ouvert. Collatinus peut servir à enseigner ou à apprendre le latin. Mais, il permet aussi, à partir d'une forme fléchie, de consulter l'un des six dictionnaires disponibles en affichant les entrées correspondantes. Collatinus connaît également les quantités de chaque syllabe, ce qui permet à l'utilisateur de saisir la structure métrique d'un vers. Il est encore en cours de développement et reste attentif aux attentes de ses utilisateurs.

ABSTRACT. – Collatinus is a lemmatizer and a morphological analyzer of Latin texts. It is free and open. Collatinus can help in teaching or learning Latin. Moreover, with an inflected form, it is able to open one of the six available dictionaries to display the corresponding entries. Collatinus knows also if the vowels are short or long, allowing the user to catch the metrical structure of a verse. It is expected to evolve in the near future, depending also on the demands formulated by the users.