“ Lemmes ”: un groupe de travail sur les outils de lemmatisation et les corpus de textes médiévaux lemmatisés

Eliana Magnani

« Lemmes » : un groupe de travail sur les outils de lemmatisation et les corpus de textes médiévaux lemmatisés

Eliana MAGNANI
CNRS - LaMOP

Paru dans *Archivum Latinitatis Medii Aevi - ALMA*, 76, 2018, p. 340-344 (impr. 2019)

Alors que depuis les années 2000 le nombre de sources textuelles médiévales numérisées et les bases de données pour les consulter n'ont pas cessé de s'accroître, quelles sont les opérations indispensables pour que la recherche sur des très larges corpus dépasse la simple requête indiciaire et s'engage résolument dans l'utilisation des méthodes numériques, dans les analyses statistiques et de sémantique structurée ? Comment peut-on mettre en pratique le postulat de la relation existant entre le sens des mots et les changements historiques, à l'instar de la théorie des champs sémantiques de Jost Trier (1894-1970), et reconstituer ainsi la logique d'un système social de représentations manifesté par la sélection des vocables utilisés et mis en relation entre eux ?[1] Ce sont ces questions qui nous orientaient lors de la création du groupe « Lemmes », l'un des cinq groupes de travail actuels du Consortium Sources Médiévales 2 (COSME2 - 2017-2020)[2].

Dans le cas des langues flexionnelles et à forte variation graphique, comme celles employées dans l'Occident médiéval – le latin et les langues vernaculaires –, toute ambition de développement de procédures de recherche formalisées et assistées informatiquement implique la lemmatisation des corpus utilisés, c'est-à-dire le regroupement des différentes formes d'un mot sous leur lemme. Ce procédé – manuel, semi-automatisé ou automatisé –, n'est pas chose nouvelle[3], mais au cours des dernières années, plusieurs équipes ce sont attelées à la création de lemmatiseurs ou de paramètres pour la lemmatisation des langues médiévales, tandis que d'autres ont mis ou sont en train de mettre à disposition des chercheurs des corpus textuels lemmatisés. Dans ce contexte de multiplication d'initiatives, il a semblé nécessaire de réunir et faire dialoguer les acteurs de ces différentes expériences, de favoriser le développement et l'évolution des applications existantes ainsi que leur utilisation sur les nouveaux corpus constitués. Le groupe « Lemmes » a donc été conçu comme un lieu d'échanges en construction permanente, ouvert à tous les intéressés : historiens, philologues, linguistes.

Au moment où nous écrivons cette chronique (juin 2019), quatre ateliers réunissant entre 15 et 30 personnes ont été réalisés en vue d'organiser et déployer les activités du groupe. Une première étape consiste à connaître les équipes et les applications concernées, afin de dresser, dans la mesure du possible et sans prétention à l'exhaustivité, un panorama de l'existant. Une constatation s'impose d'emblée : la lemmatisation de larges corpus n'est pas l'intérêt principal de tous les concepteurs d'outils, chaque projet ayant ses propres objectifs. L'aide à la traduction, l'accentuation et la scansion automatiques, l'assistance à l'édition, la constitution de bibliothèques textuelles sont, parmi d'autres, autant de pratiques qui font appel à la lemmatisation. Cela a une incidence sur les outils élaborés, les interfaces d'utilisation proposées et les développements continuels réalisés. Sans que l'on puisse entrer ici dans le détail de chaque application, les projets et outils présentés au cours des différents ateliers sont : Collatinus[4], CompHistSem[5], Hydra[6], LASLA[7], OMNIA[8], PALM[9], Pandora (devenu Pie) et l'interface de correction Pyrrha[10].

D'un point de vue technique, la lemmatisation automatique regroupe plusieurs opérations : l'acquisition du ou des textes dans un format numérisé élémentaire ; la tokenisation, c'est-à-dire le découpage du texte par lexème, au cours de laquelle le texte peut être pré-formaté selon les choix et besoins des concepteurs (par exemple, nettoyage des caractères spéciaux,

séparation d'enclitiques) ; l'étiquetage morphosyntaxique des formes (*POS tagging = part-of-speech tagging*) avec un jeux d'étiquettes très variable d'un outil à autre ; et le regroupement des formes sous le lemme correspondant. Plusieurs difficultés doivent être surmontées dans ce processus, notamment la désambiguïsation des homographes et les variations graphiques des formes.

Les outils actuels pour associer aux formes une étiquette (POS) correcte, se partagent en deux groupes principaux. Les premiers utilisent des tagueurs probabilistes basés sur un lexique et des règles prédéfinies associés ou pas à des entraînements successifs qui améliorent la reconnaissance des formes et des lemmes. Les seconds sont basés sur les technologies plus récentes dites des « réseaux de neurones » (ou *deep learning*), des algorithmes qui, à partir d'un corpus pré-annoté (ou corpus d'entraînement) visent à apprendre l'application à créer les lemmes qui ne figurent pas dans le corpus d'entraînement, en fonction de leur représentation sémantique (les mots coocurrents). Dans les deux cas, un travail manuel en amont est nécessaire, l'établissement d'un lexique et des règles, ou l'annotation d'un corpus. En aval aussi l'intervention manuelle est nécessaire dans les deux cas pour corriger les corpus annotés par les lemmatiseurs.

D'une manière générale, qu'ils appartiennent à l'un ou à l'autre de ces deux types et quels que soient leurs objectifs, tous les outils présentés lors des ateliers du groupe « Lemmes », sont estimés performants à environ 90 % (± 5 %), ce qui est déjà un seuil satisfaisant pour le traitement des très larges corpus. Les 5-15 % de fautes ou de non-reconnaissance (*unknow*) sont les aspects qui demandent réflexion, d'autant plus qu'ils recèlent souvent des problèmes historiques. L'expérience montre, par exemple, que la reconnaissance des noms propres, de personne et de lieu, figure parmi les erreurs récurrentes d'étiquetage. Le groupe entend ainsi poursuivre des travaux sur ce point. Une première piste, du côté technologique, se situe dans les recherches actuelles de détection automatique d'entités nommées. C'est dans cette perspective que deux travaux doctoraux empruntant cette voie ont été exposés et discutés lors de l'un des ateliers du groupe[11].

Une autre question soulevée lors des discussions concerne l'absence d'évaluation systématique et comparative des tagueurs. Une analyse raisonnée permettrait de connaître les avantages des solutions proposées par chaque outil, et peut-être de les combiner un jour dans une sorte de « méta-tagueur ». La mise en œuvre technique d'une telle opération a donné lieu à des vifs échanges. La tâche s'avère complexe en raison de la diversité des choix de chaque application (les jeux d'étiquettes, les formats des données, entre autres), des langues vernaculaires dont certaines ne disposent pas encore d'un corpus suffisamment important, mais peut-être aussi de la surcharge de travail par ailleurs des membres du groupe et des inerties toujours à surmonter dans ce type d'entreprise collective. Mais le groupe compte sur beaucoup de bonnes volontés et prépare la mise en ligne, sur la plateforme Ménestrel[12], des fiches descriptives des outils et paramètres, de manuels d'utilisation et d'un guide d'initiation à la lemmatisation. Le groupe œuvre ainsi à la diffusion de l'information scientifique, après avoir consacré son quatrième atelier à une journée de formation réunissant des étudiants, doctorants, ingénieurs, chercheurs et enseignants-chercheurs intéressés par la prise en main effective des outils de lemmatisation. On peut espérer que la synergie créée entre les utilisateurs potentiels et les concepteurs d'outils enrichissent, voire infléchissent, les développements à venir.

L'incitation à la constitution de corpus lemmatisés, ou la lemmatisation des corpus déjà existants, est l'autre volet du groupe de travail « Lemmes », et il est indissociable des actions relatives aux outils. Deux équipes participant au groupe, le CIFM (Corpus des inscriptions de la France Médiévale)[13] et le CBMA (Corpus Burgundiae Medii Aevi)[14] se sont lancées dans ce cadre dans un projet commun, la réalisation d'un corpus épigraphique plurilinguistique, relatif à la Bourgogne médiévale (VIII^{e}-XV^{e} siècle)[15]. Ce corpus de plus de 1400 inscriptions en latin et

en ancien français, souvent les deux mélangées, est le premier corpus multi-langues lemmatisé. Sa lemmatisation a été l'occasion de tester et comparer différents outils et paramètres, et d'expérimenter des procédures pour détecter automatiquement la langue, ou le degré de mélange linguistique, d'un texte[16].

Cette expérience, comme d'autres déjà menées ou en cours, confirment que la façon la plus efficace de réfléchir sur les outils de lemmatisation et les corpus de textes médiévaux lemmatisés est de prendre acte que la lemmatisation n'est pas seulement une opération fondamentale, mais qu'elle est aussi désormais un préalable indispensable à l'étude des textes et à l'exploitation des corpus si l'on veut que la révolution numérique se traduise par de réelles transformations dans la recherche. Le groupe « Lemmes » entend contribuer à cette évolution, à son échelle certes, mais volontairement.

---

[1] À l'instar d'Alain GUERREAU, *L'avenir d'un passé incertain. Quelle histoire du Moyen Âge au XXIe*

[2] Dirigé par Paul Bertrand et adossé à l'IRHT, COSME puis COSME2 est l'un des consortiums mis en place dans le cadre de la TGIR Huma-Num du CNRS, pour répondre aux besoins d'harmonisation entre les multiples projets numériques en SHS. Les consortiums ont pour tâche la définition et la diffusion de « bonnes pratiques » dans la communauté scientifique concernée (« définition de procédures et standards numériques partagés ») ainsi que la réalisation d'ateliers de formation en vue de « favoriser l'appropriation des outils numériques » (https://www.huma-num.fr/consortiums). Sur COSME2, voir https://cosme.hypotheses.org/. Les autres groupes de travail sont : Alignement des cotes de manuscrits ; Noms de personnes/de lieux ; Dates et formules ; « Valeurs » et mesures.

[3] Une histoire des procédés de lemmatisation reste encore à faire. Nicolas Perreaux a toutefois présenté les principaux jalons lors de l'Atelier 4 du Groupe « Lemmes » le 17 juin 2019. Citons, néanmoins, dès les années 1950, le projet précurseur *Index Thomisticus* (http://www.corpusthomisticum.org/it/index.age). Voir Roberto BUSA, « The Annals of Humanities Computing : the Index thomisticus », *Computers and the Humanities,* 14, 1980, p. 83-90.

[4] Application pour le latin classique avec une extension récente pour le latin médiéval, par Yves Ouvrard, Philippe Verkerk (https://outils.biblissima.fr/fr/collatinus-web/). Voir Yves OUVRARD et Philippe VERKERK, « Collatinus, un outil polymorphe pour l'étude du latin », *Archivum Latinitatis Medii Aevi*, 72, 2014, p. 305-311.

[5] *Computational Historical Semantics,* outils et corpus pour l'analyse sémantique des textes latins médiévaux, présenté par Tim Geelhaar (Goethe-University Frankfurt) (http://www.comphistsem.org/home.html). Voir Steffen EGER, Rüdiger GLEIM, Alexander MEHLER, « Lemmatization and Morphological Tagging in German and Latin : A comparison and a survey of the state-of-the-art ». *Proceedings of the 10th International Conference on Language Resources and Evaluation*, 2016 (http://www.lrec-conf.org/proceedings/lrec2016/pdf/656_Paper.pdf) ; Steffen EGER, Tim VOR DER BRÜCK, Alexander MEHLER, « Lexicon-assisted tagging and lemmatization in Latin : A comparison of six taggers and two lemmatization methods », *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH),* Beijing, 2015, p. 105–113.

[6] Tagueur et lemmatiseur intégrés, développé par Łukasz Gągała (Göttingen University), notamment pour le moyen haut et le moyen bas allemand (https://github.com/Lukasz-G/Hydra). Voir Łukasz GAGALA, « Authorship Attribution with Neural Networks and Multiple Features : Notebook for PAN at CLEF 2018 », In Linda CAPPELLATO, Nicola FERRO, Jian-Yun NIE, Laure SOULIER (éd.) *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*. Avignon, France, September 10-14, 2018, (http://ceur-ws.org/Vol-2125/paper_146.pdf).

[7] Laboratoire d'Analyse Statistique des Langues Anciennes (Université de Liège), fondé en 1961, est la première équipe à avoir appliqué les méthodes automatiques au latin et au grec classiques (http://web.philo.ulg.ac.be/lasla/). Leur base de données Opera Latina a été présentée par Dominique

Longrée et Margherita Fantoli, tandis que les développements récents du LASLA-Tagger ont été présentés par Yves Ouvrard et Philippe Verkerk. Voir Louis DELATTE, Étienne ÉVRARD, « Un laboratoire d'analyse statistique des langues anciennes à l'Université de Liège », *L'Antiquité classique*, 30-2, 1961, p. 429-444.

[8] Outils et Méthodes Numériques pour l'Interrogation et l'Analyse des textes médiolatins, programme ANR (IRHT, EnC, Artehis) (2009-2013) dirigé par Alain Guerreau, ayant élaboré un tokeniseur et les premiers paramètres pour le latin médiéval pour le logiciel TreeTagger (https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/), présentés par Renaud Alexandre (CNRS - IRHT) (http://glossaria.eu/outils/lemmatisation/). Voir Bruno BON, « OMNIA – Outils et Méthodes Numériques pour l'Interrogation et l'Analyse des textes médiolatins », *BUCEMA - Bulletin du centre d'études médiévales d'Auxerre*, 13, 2009, p. 291-292 (http://journals.openedition.org/cem/11086) ; ID., « OMNIA: outils et méthodes numériques pour l'interrogation et l'analyse des textes médiolatins (2) », *BUCEMA - Bulletin du centre d'études médiévales d'Auxerre*, 14, 2010, p. 251-252 (http://journals.openedition.org/cem/11566) ; Id., « OMNIA: outils et méthodes numériques pour l'interrogation et l'analyse des textes médiolatins (3) », *BUCEMA - Bulletin du centre d'études médiévales d'Auxerre*, 15, 2011, p. 333-334 (http://journals.openedition.org/cem/12015).

[9] Plate-forme d'analyse linguistique médiévale, développée dans le cadre du programme ERC *States and Signs* dirigé par Jean-Philippe Genet (Universtié Paris 1 Panthéon-Sorbonne - LaMOP) (2010-2014), dédiée au latin, au moyen français et au moyen anglais, présentée par Mourad Aouini, Chris Fletcher et Aude Mairey (http://palm.huma-num.fr/PALM/). Voir Mourad AOUINI, « A NooJ Module for Named Entity Recognition in Middle French Texts », In Johanna MONTI, Max SILBERZTEIN, Mario MONTELEONE, Maria Pia DI BUONO (éd.), *Formalising Natural Languages with Nooj 2014*, Cambridge, 2015, p. 99-112.

[10] Pandora est un tagueur-lemmatiseur pour le latin et les langues vernaculaires, développé par Mike Kestemont, Jean-Baptiste Camps, Thibault Clérice, Enrique Manjavacas (https://github.com/hipster-philology/pandora), qui se poursuit dans le cadre du tagueur Pie (https://github.com/emanjavacas/pie) développé par Enrique Manjavacas et Mike Kestemont, de son interface web Deucalion, ainsi que de Pyrrha, outil pour la correction en ligne de l'étiquetage automatique développé par Julien Pilla, Thibault Clérice et Jean-Baptiste Camps, présenté par Ariane Pinche, Vincent Jolivet et Thibault Clérice (EnC) (https://github.com/hipster-philology/pyrrha). Voir Mike KESTEMONT, Guy DE PAUW, Renske VAN NIE, Walter DAELEMANS, « Lemmatization for variation-rich languages using deep learning », *Digital Scholarship in the Humanities*, 32-4, 2017, p. 797-815, (https://doi.org/10.1093/llc/fqw034) ; Enrique MANJAVACAS, Akos KADAR, Mike KESTEMONT, « Improving Lemmatization of Non-Standard Languages with Joint Learning », arXiv:1903.06939v1, 2019 (https://arxiv.org/format/1903.06939v1).

[11] Il s'agit des communications de Mourad Aouini (CNRS - CLT), « Approche multi-niveaux pour la reconnaissance des entités nommées en Moyen Français » et de Sergio Torres (UVSQ - DYPAC), « La récupération automatique des entités nommées dans les chartes médiolatines. Modélisation et perspectives d'utilisation ».

[12] http://www.menestrel.fr/

[13] CESCM (Poitiers). Voir http://cescm.labo.univ-poitiers.fr/publi/corpus-des-inscriptions-de-la-france-medievale.

[14] LaMOP (Paris). Voir http://www.cbma-project.eu/.

[15] Eliana MAGNANI, Estelle INGRAND-VARENNE, « Le corpus épigraphique bourguignon (VIII[e]-XV[e] siècle). Des catalogues aux applications numériques », *BUCEMA - Bulletin du centre d'études médiévales d'Auxerre*, Collection CBMA, Les journées d'études (en ligne : http://journals.openedition.org/cem/15591). L'ensemble des données relatives à ce projet sont disponibles en ligne dans http://www.cbma-project.eu/%C3%A9ditions/textes_epigraphiques.html ; https://gitlab.huma-num.fr/lamop/cbma-epigraphie ; http://philologic.lamop.fr/epigraphie/.

[16] Les opérations de lemmatisation ont été réalisées par Nicolas Perreaux (Université de Namur).