UNION ACADÉMIQUE INTERNATIONALE
1919-2019

Centenaire de l'Union Académique Internationale

Mercredi 27 novembre 2019 – Institut de France
9h30-12h30 – Salon Bonnefous

Le *Dictionnaire du Latin Médiéval*
Réflexions méthodologiques à propos du tournant numérique

Élaboré sous l'égide de l'UAI à partir de 1920, le *Dictionnaire du Latin Médiéval* visait à fournir aux médiévistes un instrument lexicographique répondant aux critères scientifiques contemporains, comparable au *Thesaurus Linguae Latinae* dédié au latin classique . Quelle qu'ait été la qualité des méthodes mises en œuvre dans ces ouvrages, elles introduisaient un certain nombre de biais, aggravés, dans le cas du latin médiéval, par la masse des textes conservés.
En effet, les dépouillements manuels ne peuvent retenir qu'un nombre limité des attestations d'un vocable donné, ce qui a pour effet une surreprésentation des termes ou des valeurs sémantiques rares et laisse une place excessive à la subjectivité des rédacteurs de fiches. Par ailleurs, l'organisation des articles apparaît assez souvent dépendante des conceptions sémantiques modernes et, procédant par découpage du matériau lexical, elle fait apparaître des distinctions tranchées plus que des continuums et des liens sémantiques.
Ces faiblesses sont inhérentes aux conditions dans lesquelles s'opère le travail et elles ne peuvent être corrigées qu'à la marge. Le développement des technologies numériques permet au contraire d'envisager un renouvellement profond des méthodes, qui s'appuie sur la constitution de vastes corpus numérisés et l'élaboration d'outils d'interrogation linguistique adaptés aux spécificités d'une langue « morte ».
On sait depuis longtemps que le vocabulaire traduit toujours les conceptions que telle ou telle société se fait d'une réalité d'ordre pratique, social ou idéel. Se dégager de nos propres conceptions pour reconstituer celles des hommes du Moyen Age représente une difficulté presque insurmontable en lexicographie traditionnelle. Par leur puissance et leur « neutralité », les outils numériques permettent de surmonter cette difficulté, en faisant apparaître les connexions lexicales qui définissent le système de sens dans lequel s'inscrit un mot donné. Ils permettent donc aussi de développer une véritable sémantique historique et une approche renouvelée des questionnements historiens à partir des textes. Les exposés présentés illustreront ces perspectives nouvelles, à partir de programmes en cours de réalisation en Allemagne, Catalogne, France et Pologne.

***

9h45 – Quatre siècles européens dans un seul dictionnaire : les défis du *Novum Glossarium Mediæ Latinitatis*

Renaud ALEXANDRE, IRHT-CNRS, Comité Du Cange (Paris)

À la fin du XIX^{e} et au début du XX^{e} siècle, différents projets lexicographiques naissent. Il s'agit alors de pourvoir à un besoin d'outils de référence qui aideraient à comprendre les langues et les cultures anciennes de l'Europe. L'un de ces projets, le « Dictionnaire du latin médiéval », a pour objet la description du vocabulaire latin sur l'ensemble de l'Europe ; publié en fascicules à partir de

1957, le *Novum Glossarium Mediæ Latinitatis* couvre une période qui va de la fin du VIII^e^ siècle au début du XIII^e^ siècle.
Pour mener à bien cette lourde tâche, un grand nombre de textes médiévaux font l'objet de dépouillements. Des articles de dictionnaire sont rédigés à partir de ces dépouillements ; dans ces articles, les rédacteurs tentent de rendre compte des sens des vocables en rendant sensible leur évolution dans l'espace et le temps. Des biais sont susceptibles d'intervenir à différents moments de ce processus, la matière elle-même rendant impossible un dépouillement – et donc un traitement – exhaustif.
La généralisation de l'emploi des outils informatiques en sciences humaines et sociales, au début du XXI^e^ siècle, modifie profondément ce cadre de recherche. Le champ des possibles s'ouvre considérablement, à condition toutefois de prendre la mesure de cette transformation et des difficultés qui lui sont inhérentes.
Dans cette communication, je souhaiterais décrire ce tournant tel qu'il a été mis en œuvre par l'équipe du *Novum Glossarium*. Dans un premier temps, il s'agira de faire le bilan des outils, méthodes et difficultés propres à la lexicographie traditionnelle du latin médiéval, et, dans un second temps, de retracer les chemins dans lesquels l'équipe de rédaction de ce dictionnaire s'est engagée afin de tirer pleinement profit du numérique.

***

10h15 – L'édition numérique du *Glossarium Mediae Latinitatis Cataloniae* : après la réflexion, quelle conceptualisation et quelle pratique pour quels résultats ?

Ana GOMEZ RABAL, Institution Milá y Fontanals, CSIC (Barcelone)

Au sein du groupe du *Glossarium Mediae Latinitatis Cataloniae* (GMLC), la création d'un corpus numérique des textes utilisés comme matière première pour la rédaction du GMLC et aussi pour les recherches parallèles des membres de l'équipe vise, depuis 2012, au développement et à l'expansion réguliers du *Corpus Documentale Latinum Cataloniae* (CODOLCAT). Celui-ci doit être défini comme une base de données lexicale de publication périodique qui permet l'accès, de façon libre et gratuite, à une partie de plus en plus large du corpus textuel utilisé pour la rédaction des articles lexicographiques du GMLC. Depuis 2012, une nouvelle version du CODOLCAT est préparée et publiée chaque année et, en tant que publication périodique, le CODOLCAT a un numéro d'ISSN et un taux d'utilisation remarquablement soutenu comme instrument de consultation parmi les philologues latinistes ou romanistes, les historiens, les historiens du droit, etc.
Pendant ces années de mise en service et extension du CODOLCAT, l'équipe du GMLC a envisagé la publication numérique de son dictionnaire : avec cet objectif, l'équipe est en train de développer le processus de numérisation et d'encodage des articles déjà composés, ainsi que la préparation des outils nécessaires pour permettre la rédaction de nouveaux articles avec une issue numérique. La préparation de la publication numérique du GMLC constitue maintenant un enjeu concret pour le groupe en raison de la compatibilité des deux outils qui veulent être offerts, c'est à dire, le glossaire même en version numérique et la base de données lexicale CODOLCAT.
L'objectif ultime des intégrants du GMLC est la publication complète de tout le glossaire d'une façon agile, fiable, commode pour un lecteur qui pourra passer de la consultation de l'œuvre rédigée, explicative et forcement sélective, à la consultation de la base de données lexicale, ordonnée mais massive, apportant tous les exemples, tous les contextes, tous les résultats. L'intention des membres de l'équipe est, cependant, plus large : pourquoi cette expérience d'intégration entre les deux outils d'un même projet ne servirait-elle pas de piste d'essai afin de tracer des lignes inspiratrices d'un parcours plus ambitieux ? Au sein du projet européen commun

de lexicographie latine médiévale qui a commencé à prendre corps lors de la publication en 1957 de la lettre L du *Novum Glossarium Mediae Latinitatis*, le développement, dans un avenir pas trop lointain, d'une plateforme de consultation commune des dictionnaires européens de latin médiéval et – si possible – de leurs bases de données lexicales serait un moyen de connaissance du latin du Moyen Âge dans toute sa diversité, géographique, stylistique et chronologique, moyen très utile pour la communauté scientifique et le public intéressé. La technologie actuelle permettrait de créer cette plateforme commune (non un serveur unique) où chaque groupe déposerait les données par lui-même décidées. Malgré les diversité des buts déjà atteints par les différentes équipes, malgré celle des méthodes et des moyens employés, la collaboration entre les équipes avec un tel horizon trouverait son lieu de discussion le plus convenable dans le contexte de la célébration du centenaire de l'Union Académique Internationale.

***

11h15 – Le « Latin Text Archive » à l'Académie des sciences de Berlin-Brandebourg. Des nouvelles possibilités pour la recherche historique et les dictionnaires numériques.

Tim GEELHAAR, Goethe-Universität Frankfurt am Main (Francfort)

À partir de l'année prochaine, une nouvelle plateforme offrira un accès analytique aux sources historiques latines. Le « Latin Text Archive » (LTA), hébergé à l'*Académie des sciences de Berlin-Brandebourg* (BBAW), relie les corpus textuels mis en place à l'Université de Francfort à différents outils d'analyse, relevant du domaine des sciences humaines numériques. Notre objectif est double : il s'agit de mettre à disposition du public le stock de textes actuellement accessible à Francfort, dans un cadre institutionnel fiable et durable, mais aussi de le rendre interrogeable via des méthodes inédites – ceci sans que les utilisateurs aient besoin d'avoir des compétences en informatique. Le LTA fournira un matériel documentaire en libre accès, organisé en sous-corpus, permettant des requêtes diachroniques. Cette possibilité n'est toutefois envisageable que grâce à un travail approfondi sur le lexique latin, qui est à la base du balisage des textes du LTA. Afin d'obtenir la lemmatisation la plus précise possible, le groupe de travail de Bernhard Jussen et de Tim Geelhaar a établi un dictionnaire numérique du latin médiéval, le *Frankfurt Latin Lexicon* (FLL), contenant des informations morphologiques pour chaque entrée lexicale en gardant la variabilité de l'orthographe du latin médiéval, ainsi que des outils permettant de contrôler et d'améliorer les résultats d'un balisage automatisé.
Cette intervention propose de montrer que, tout comme le travail sur les méthodes d'analyse et l'incrémentation du stock de textes ne s'arrête pas en 2019, celui sur le dictionnaire numérique de Francfort doit être poursuivi. Cet effort est en effet nécessaire pour plusieurs raisons. Au fur et à mesure que de nouveaux textes seront ajoutés à la collection, des néologismes et des variantes lexicales rares pourront être identifiés, puis ajoutés au FLL. Contrairement aux dictionnaires traditionnels, un dictionnaire numérique offre en effet des possibilités d'extension simples et efficaces. Par exemple, il est possible d'isoler les noms propres afin de créer des dictionnaires spécialisés, à partir de la banque de données initiale. Au-delà des aspects lexicaux à proprement parler, le LTA permettra aussi d'explorer le sens des lemmes latins, via la production des listes de cooccurrences et la visualisation de champs sémantiques. Un dernier avantage est que toutes les informations d'un dictionnaire numérique peuvent être reliées à volonté à d'autres ressources, en prévision de la construction d'autres outils d'enquête. Cependant, la qualité de tout ce travail dépend des lexicographes, qui seuls peuvent garantir la chose la plus importante : des données philologiques fiables.

***

11h45 – Le corpus électronique du latin médiéval polonais : bilan et perspectives

Krzysztof NOWAK, Institut de la langue polonaise, Académie Polonaise des Sciences (Cracovie)

L'apparition des corpus textuels électroniques a profondément changé la façon dont les chercheurs étudient aujourd'hui les textes du passé. Outre qu'ils ont considérablement facilité la recherche (accès aux textes, vitesse d'enquête, *etc.*), ils en ont aussi souvent modifié la base elle-même. Dans le domaine linguistique, cela entraîne surtout une préférence pour les approches quantitatives, et une concentration sur les phénomènes fréquents et récurrents. Plus spécifiquement, les concordances, les listes de collocations et de fréquences, et les fonctionnalités avancées de tri ont permis aux lexicographes et lexicologues de prêter plus d'attention aux déterminants contextuels de la signification du mot. Cela dit, l'effet de la révolution quantitative sur l'étude du latin médiéval semble avoir été moins sensible, et l'adaptation des méthodes moins enthousiaste. L'une des raisons peut s'en trouver dans la disponibilité limitée de corpus dont la composition et l'architecture seraient fondées sur des critères scientifiques, au contraire de simples collections de textes.
Les origines du projet de corpus du latin médiéval polonais *eFontes* (http://scriptores.pl/efontes) remontent à 2012, quand il a été conçu au sein de la Section du latin médiéval de l'Institut de la langue polonaise (Académie Polonaise des Sciences, Cracovie), cette équipe qui prépare le *Lexicon mediae et infimae Latinitatis Polonorum* depuis les années 1950, sous les auspices de l'Union Académique Internationale. Le corpus *eFontes*, représentatif et balancé, a pour ambition de refléter la variation de la production écrite en latin sur le territoire de la Pologne entre l'an mil et les années 1550. Outre leur dimension chronologique et géographique, les textes sont inclus en tenant compte de leurs propriétés sociolinguistiques telles que la fonction, le genre, le domaine et le milieu. Dans sa version de 2017, le corpus contient environ cinq millions de mots lemmatisés, mais le financement complémentaire obtenu en 2018 pour la continuation des travaux permettra d'élargir considérablement la base textuelle (jusqu'à trois fois dans les cinq ans à venir) et d'en changer l'architecture. Du *corpus maius* principal, d'environ 15 millions de mots, on extraira un *corpus minus*, sous-corpus d'un million de mots, qui sera annoté à la main pour servir ensuite dans l'entraînement du marqueur et lemmatiseur automatique.
Néanmoins, sans attendre l'achèvement de ce grand projet, l'équipe mène déjà de nombreuses recherches lexicales à partir du corpus déjà disponible. Ces travaux visent, entre autres sujets, à étudier les métaphores médiévales, les mécanismes du changement sémantique et la terminologie ecclésiastique.