Les nouveaux corpus CBMA : hagiographie, épigraphie, alia. Bilan et perspectives (2017-2020)

Eliana Magnani

Bulletin du centre d'études médiévales d'Auxerre | BUCEMA
Collection CBMA

Les nouveaux corpus CBMA : hagiographie, épigraphie, *alia*. Bilan et perspectives (2017-2020)

Communication dans le cadre de la journée d'études CBMA-LaMOP - Des applications aux manuscrits. Expériences de transcriptions automatiques de manuscrits et développements du Corpus Burgundiae Medii Aevi – 19 mars 2020

Eliana Magnani

Édition électronique
URL : http://journals.openedition.org/cem/17087
ISSN : 1954-3093

Éditeur
Centre d'études médiévales Saint-Germain d'Auxerre

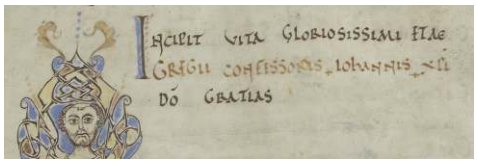
Référence électronique
Eliana Magnani, « Les nouveaux corpus CBMA : hagiographie, épigraphie, *alia*. Bilan et perspectives (2017-2020) », *Bulletin du centre d'études médiévales d'Auxerre | BUCEMA* [En ligne], Collection CBMA, Les journées d'études, mis en ligne le 20 mai 2020, consulté le 20 mai 2020. URL : http://journals.openedition.org/cem/17087

Ce document a été généré automatiquement le 20 mai 2020.

Les nouveaux corpus CBMA : hagiographie, épigraphie, *alia*. Bilan et perspectives (2017-2020)

Communication dans le cadre de la journée d'études CBMA-LaMOP - Des applications aux manuscrits. Expériences de transcriptions automatiques de manuscrits et développements du Corpus Burgundiae Medii Aevi – 19 mars 2020

Eliana Magnani



1 Le propos de cette présentation est de revenir sur les principales réalisations du projet CBMA entre 2017 et 2019. Il s'agit aussi de traiter des développements lancés ou en cours en cette année 2020 et des perspectives à venir.

2 Le projet CBMA, né en 2004, a connu une évolution importante en 2017. La plateforme de documents diplomatiques devient alors une plateforme de documents pluri-typologiques. On passe des Chartes au Corpus[1]. Dans le contexte de numérisation toujours croissante et inédite d'une grande masse de documents conduisant à des nouvelles méthodes de recherche, l'objectif de l'équipe est de mettre à la disposition des chercheurs, en accès libre, un corpus raisonné, géographiquement circonscrit, facilement accessible et adaptable : le corpus peut ainsi être utilisé pour lui-même, peut servir de point de comparaison avec d'autres corpus et être utilisé à des approches à différentes échelles : locales, régionales, européennes.

L'équipe et les partenaires institutionnels

3 Toutes les réalisations présentées ci-dessous sont le résultat d'un travail d'équipe : Mathieu Beaud (Univ. Paris 1 – LaMOP), Pierre Brochard (CNRS – LaMOP), Hélène Caillaud (Univ. Limoges), Davide Gherdevich (UVSQ - DYPAC), Estelle Ingrand-Varenne

(CNRS – CESCM), Eliana Magnani (CNRS – LaMOP), Aurore Menudier (CESCM), Nicolas Perreaux (Univ. Paris 1 - LaMOP), Coraline Rey (Univ. Bourgogne - Univ. Angers).

4 Plusieurs des membres de l'équipe, hautement qualifiés, n'ont que des emplois précaires. Nous tenons à rappeler avec véhémence que sans une politique forte d'emplois statutaires pérennes, nos recherches ne pourront pas se développer.

5 Ces projets sont portés par le LaMOP, en collaboration avec le CESCM de Poitiers pour le volet épigraphique, et ont été soutenus par le Labex Hastec (en collaboration avec l'IRHT, l'EnC – Centre Jean Mabillon, le CERCOR-LEM) et par le consortium Cosme2.

Un corpus structuré et hétérogène de textes latins médiévaux

6 À partir de 2017, l'équipe CBMA a œuvré à la constitution d'un corpus structuré et hétérogène de textes latins médiévaux relatifs à la Bourgogne, du Ve au XV^{e} siècle. Il est parti de l'acquis l'existant : un corpus électronique et une base de données mono-typologiques de 29 000 documents diplomatiques (chartes), transformé en corpus comprenant différents types documentaires (narratifs, théologiques, normatifs, épigraphiques, etc.)[2]. Entre 2017 et 2019, deux sous-corpus principaux ont été réunis.

7 Le premier est constitué de 328 textes hagiographiques (vies, passions, translations, recueils de miracles relatifs aux saints)[3]. Le deuxième sous-corpus est constitué d'inscriptions épigraphiques. Il contient 1418 textes, dont 471 avaient déjà été édités auxquels ont été ajoutés 947 textes supplémentaires, surtout de la fin du Moyen Âge. Ce corpus est constitué de 60% d'inscriptions funéraires. Avec les inscriptions sont entrés dans le corpus plusieurs textes en ancien français[4]. Ce fait amène à des nouvelles propositions pour l'évolution à venir des CBMA, celle d'un corpus plurilinguistique. En plus des inscriptions concernant la Bourgogne, 850 inscriptions relatives à la « Provence »[5] ont été réunies sous les mêmes principes en 2019[6].

8 Actuellement l'inventaire des textes « autres » à incorporer est en cours. Du point typologique ils concernent des productions narratives, liturgiques et normatives. Ont été acquis ou saisis récemment (fin 2019) : la *Vie de Garnier de Saint-Etienne de Dijon*, éditée par Jean-Luc Chassel (1993) ; *La Chronique de Falcon de Tournus* éditée par René Poupardin (1905), révisée par François Bougard et Dominique Poirel (2019) que nous remercions de nous avoir transmis leur travail ; le *Martyrologe de Marcigny*, édité par Regina Hausmann (1984) ; le coutumier clunisien de Bernard, édité par D. Marquard Herrgott (1726 – *Vetus disciplina monastica*).

Acquisition et formalisation des corpus

9 Le processus d'acquisition et de formalisation des corpus se compose d'un certain nombre d'étapes principales. Nous réalisons d'abord un inventaire des textes édités concernés devant être récupérés ensuite. Pour l'hagiographie, nous sommes partis de la *BHL* (*Bibliotheca Hagiographica Latina*) et des listes réalisées par l'équipe de Guy Philippart pour le projet « Hagiographies ». La récupération des textes hagiographiques édités s'est appuyée sur les éditions anciennes disponibles en ligne - les *Acta Sanctorum,* la *Patrologie latine*, les *MGH* (*Monumenta Germaniae Historica*). Pour l'épigraphie, le projet a été mené en étroite collaboration avec les épigraphistes du CESCM de Poitiers. Les

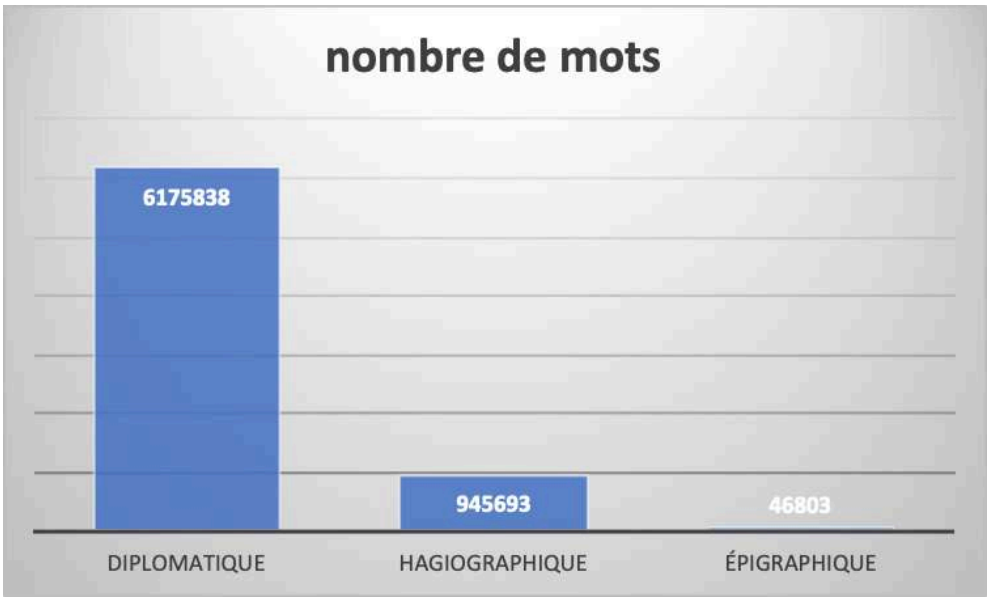
textes ont été récupérés à partir des volumes océrisés du *Corpus des Inscriptions de la France Médiévale* (jusque 1300) auxquels ont été ajoutés les inscriptions de la fin du Moyen Âge répertoriées dans le fichier papier du CESCM de Poitiers. Il y a eu ainsi un travail, pour les inscriptions épigraphiques, d'édition d'inédits.

10 L'opération suivante concerne la documentation de chaque texte, c'est-à-dire, les métadonnées qui sont associées à chaque unité textuelle. Sans être exhaustifs, les principales métadonnées sont relatives à la datation et à la localisation, avec le renseignement des coordonnées géospatiales. Sont aussi renseignés les acteurs en jeu dans la production documentaire : l'auteur, l'institution promotrice, notamment. En amont, nous avons mis en place d'un jeu de métadonnées adapté à chaque type documentaire. Nous avons défini un cahier de charges discuté en atelier avec les spécialistes des différents types documentaires.

11 L'autre opération importante dans la constitution du corpus concerne son enregistrement. Depuis les débuts du projet CBMA, nous avons toujours choisi des formats électroniques d'enregistrement simples et multiples (comme .txt, .csv) permettant une diffusion élémentaire des textes et des données. Parallèlement, des formalisations *ad hoc* du corpus sont réalisées en vue de l'adapter à différents types de recherches assistées informatiquement. Cela concerne la géolocalisation des unités textuelles en vue de l'utilisation du corpus dans des Systèmes d'information géographique (SIG), la lemmatisation des textes en vue de l'utilisation d'outils de lexicométrie et de statistiques sémantique, entre autres.

12 Grâce à ces nouveaux corpus, entre 2017 et 2019, il a été possible acquérir environ 1 million de mots supplémentaires. CBMA contient aujourd'hui 7 168 334 mots[7]. Dans le détail, par type, les CBMA se composent de 6 175 838 mots provenant de textes diplomatiques ; 945 693 mots de textes hagiographiques ; 46 803 mots d'inscriptions épigraphiques (Fig. 1). Le poids de la diplomatique demeure et demeurera toujours prédominante, mais on peut désormais envisager de comparer ces différents types de production écrite.

Fig. 1. Répartition du corpus CBMA par type documentaire en nombre de mots.



13 Pour consulter et télécharger les corpus et les métadonnées, on peut explorer le portail web du projet, hébergé par le TGIR Huma-Num[8]. On trouve aussi sur GitLab toutes les étapes du projet et les dernières mises à jour[9]. Le projet CBMA propose également l'interrogation des corpus à partir de différents outils de fouille de texte (*text mining*). Ces outils permettent de réaliser des tables de concordances, d'effectuer des recherches de cooccurrences, et des analyses statistiques. Les corpus diplomatique et épigraphique ont été lemmatisés et pré-formatés pour TXM. TXM (Textométrie) est un outil pour l'analyse de grands corpus de textes, qui utilise les méthodes la lexicométrie et de la statistique textuelle, avec la possibilité de produire des graphes, des analyses factorielles, entre autres[10]. Les corpus sont aussi interrogeables sous Philologic4[11]. Ce logiciel ne réalise pas de calcul poussé comme TXM ni la prise en compte de la lemmatisation, mais offre une interface d'interrogation aisée du corpus, et pour la recherche des cooccurrences. Trois plateformes ont été mises en place, avec les textes diplomatiques, hagiographiques et épigraphiques[12]. Le corpus diplomatique peut aussi être interrogé avec NoSketchEngine. Cet outil réalise la prise en charge intelligente de corpus lemmatisés, effectuant des calculs avec des différents paramètres, produisant des échantillons aléatoires, entre autres[13].

14 Dans le cadre de la collaboration avec les Archives départementales de la Côte d'Or, parallèlement au corpus textuel, le projet CBMA met à disposition également des documents originaux numérisés en mode image. En 2019 une importante évolution technique a été mise en place avec la création d'une la plateforme compatible avec les standards IIIF et consultable avec le visualiseur Mirador. La bibliothèque numérique CBMA contient actuellement 64 manuscrits, dont des cartulaires, des livres de comptes et différents autres documents comme les tablettes en cire de Cîteaux[14].

Recherche exploratoire du corpus hagiographique avec TXM

15 Pour passer à l'utilisation effective des corpus, nous souhaitions montrer quelques exemples de recherches exploratoires en utilisant les corpus CBMA et les outils adaptés comme TXM, utilisé ici dans l'exploitation du corpus hagiographique. Ce sont des analyses et interprétations qui doivent encore être affinées, voire corrigées, et sont proposées ici seulement à titre d'exemple.

16 L'analyse statistique du lexique peut aider à la datation des textes hagiographiques non datés. Dans la Fig. 2 on observe l'analyse factorielle des correspondances par siècle du corpus hagiographique. Il explicite le rapprochement et opposition entre les textes. On distingue deux grands ensembles, à droite, entouré de bleu, le Haut Moyen Âge (V^{e}-IX^{e} siècle) ; à gauche, entouré de vert, le Moyen Âge Central et le Bas Moyen Âge (X^{e}-XIV^{e} s.) ; au milieu, entouré de rouge, les textes non datés (sd = sans date) sont proches du V^{e} et du VI^{e} siècle, ce qui peut nous orienter à pousser les analyses statistiques avec les textes de ces siècles pour essayer de dater les textes non datés. Un autre type d'analyse statistique, le clustering, donne un résultat proche : les textes non datés sont dans la branche des textes du VI^{e} siècle (entouré de rouge) (Fig. 3).

Fig. 2. AFC - Aide à la datation des textes hagiographiques non datés.

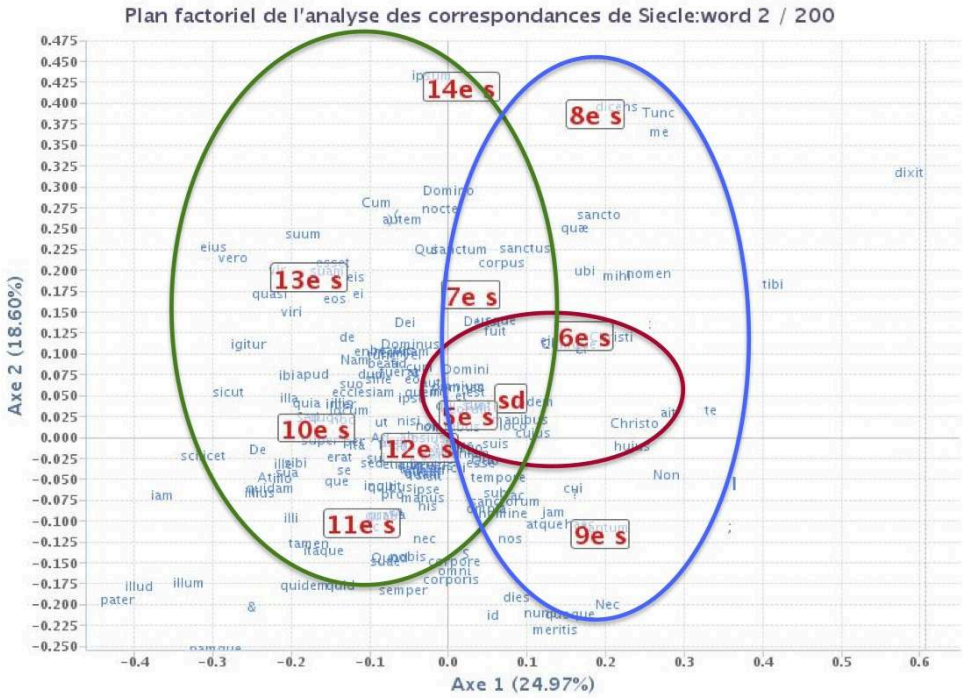
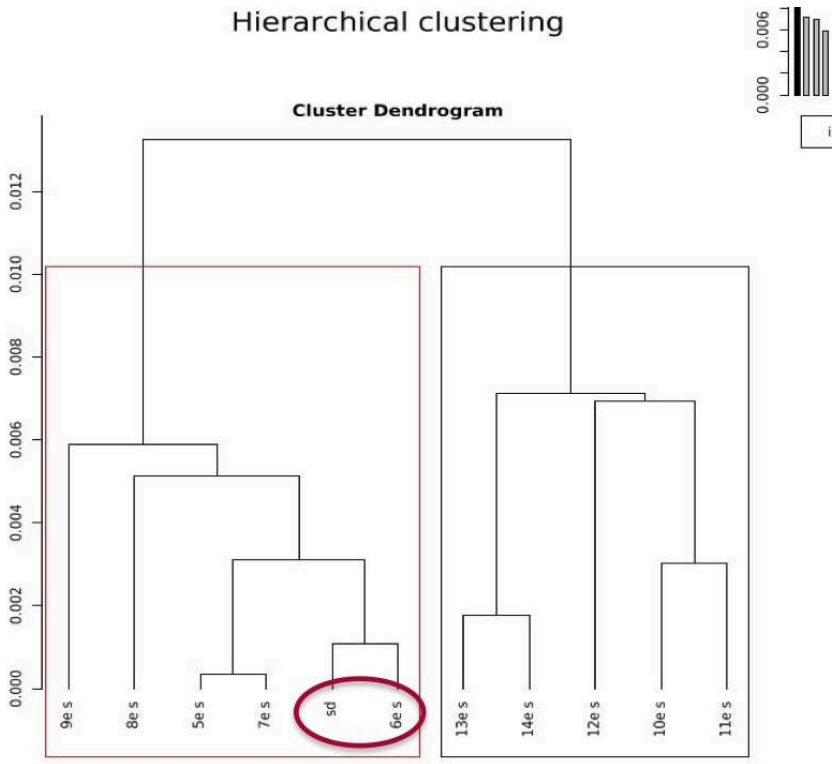


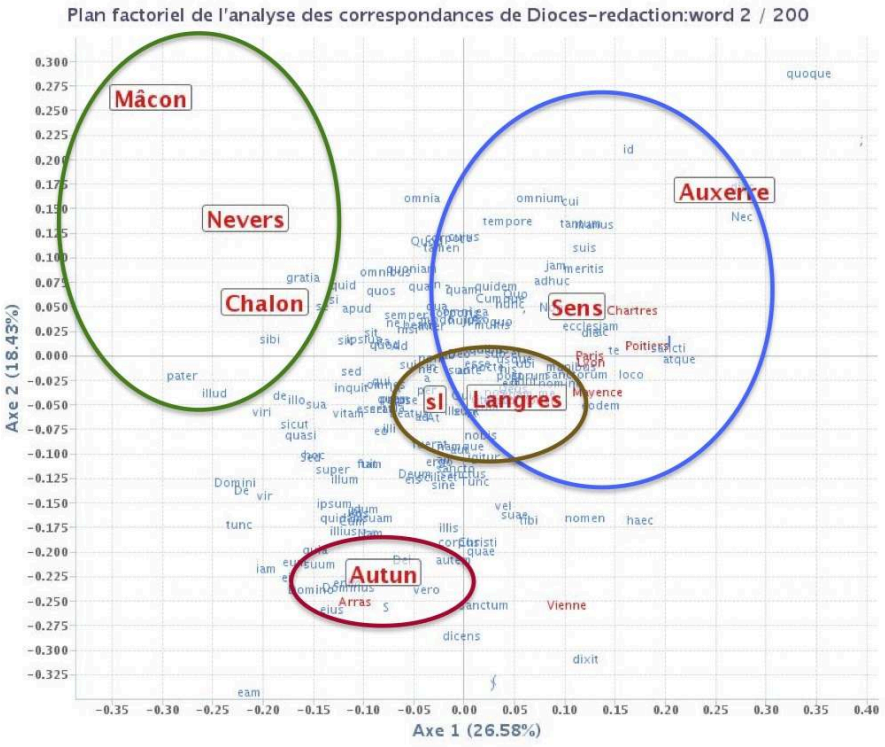
Fig. 3 – Clustering - Aide à la datation des textes hagiographiques non datés.



17 Dans la Fig. 4 on montre analyse factorielle de correspondances des textes hagiographiques par diocèse de rédaction. On observe que les diocèses du Sud (entourés de vert) et du Nord (entourés de bleu) de la Bourgogne sont opposés. En rouge est signalé le diocèse d'Autun, qui dans d'autres analyses (par exemple, du corps

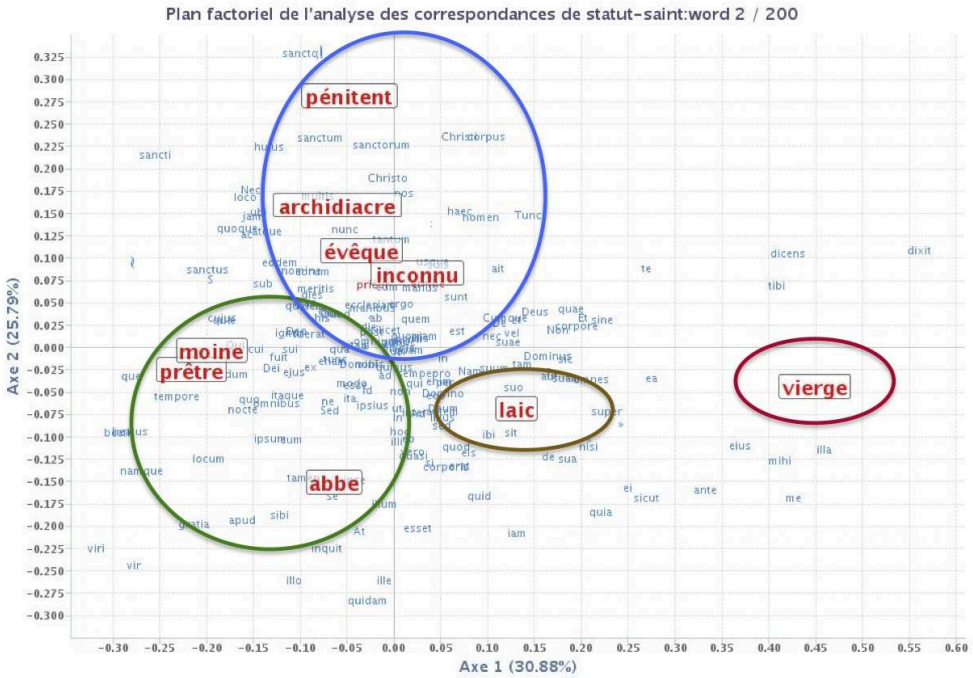
diplomatique) apparaît à part et montre ainsi un développement particulier par rapport à la production écrite. En marron sont regroupés les textes non localisés (sl = sans lieu), qui apparaissent proches de Langres et qui seraient donc à comparer avec les textes du Nord de la Bourgogne pour proposer une localisation.

Fig. 4. AFC - Textes hagiographiques par lieu de rédaction (diocèse).



18 Une autre analyse possible concerne le statut et le genre des saintes et des saints. On observe deux oppositions principales dans la Fig. 5. D'un côté les saints « ecclésiastiques », entourés de bleu et vert à gauche, par rapport aux laïcs, entourés de marron à droite. Surtout, la principale opposition se situe par rapport au genre des saints, hommes et femmes, car ce sont les saintes vierges qui se trouvent complètement à l'écart à droite, entourées de rouge.

Fig. 5. AFC - Statut et genre des saintes et des saints.



Rétro-développements du corpus épigraphique

19 Nous souhaitons terminer par deux exemples de rétro-développement du corpus épigraphique. Il s'agit de rappeler comment l'exploitation même du corpus peut contribuer à son enrichissement[15].

20 Dans les années 1970 le *Corpus des Inscriptions de la France médiévale* à Poitiers a regroupé les inscriptions répertoriées d'après les circonscriptions territoriales françaises actuelles, les départements. Même si ces circonscriptions sont commodes pour organiser la publication des volumes, elles sont anachroniques par rapport au Moyen Âge. Pour sa part, le projet CBMA a pris comme principal dénominateur commun spatial les circonscriptions ecclésiastiques médiévales, les diocèses. Même si ces entités ont varié au cours du Moyen Âge, il était important de pouvoir proposer une répartition du corpus sur des critères médiévaux. Le corpus épigraphique n'était pas renseigné par rapport aux diocèses d'origine des inscriptions, mais il contenait déjà l'indication importante du lieu de production, de provenance ou de conservation des inscriptions. À ces informations géographiques ont été associées automatiquement les coordonnées de géolocalisation (WGS 84 - EPSG 4326), donnant lieu à la réalisation d'un fichier shapefile (.shp) utilisable dans les logiciels SIG.

21 À partir de ce fichier une série d'analyses SIG ont été réalisées. Depuis la place de chaque inscription sur les cartes, il a été extrait un tableau avec leur situation dans le diocèse correspondant. Le diocèse de chaque inscription a pu ainsi été rétroinjecté depuis les cartes dans les métadonnées du corpus. Par exemple, les localités de l'actuel département de l'Yonne se situaient dans les anciens diocèses de Sens ou d'Auxerre. Cette information était indispensable pour l'exploitation croisée de différents types de sources de CBMA.

22 Un dernier exemple de retro-développement concerne la lemmatisation et la détection automatique des langues. Ce n'était pas prévu au départ, mais avec les inscriptions de la fin du Moyen Âge, plusieurs textes en langue vernaculaire, en ancien français, ou avec les deux langues mélangées, sont entrés dans le corpus. Or, pour la lemmatisation, il a fallu utiliser deux paramètres linguistiques. Ces paramètres ont été testés sur l'ensemble du corpus et selon le nombre de non-reconnaissances il a été possible, d'une part de tester les paramètres existants et d'autre part de proposer des voies possibles de traitement automatisé. Après la formalisation du corpus, l'enjeu était d'observer dans quelle mesure le jeu d'étiquettes linguistiques attribuées par les épigraphistes/philologues aux inscriptions (latin, français, lat_français = prédominance du latin, fr_latin = prédominance du français) correspondrait ou pas à la granularité issue d'une analyse des textes par l'ordinateur dont les paramètres sont établis par langue. Sans entrer dans le détail de ce travail, en fin de parcours, l'une des solutions envisageables pour lemmatiser les inscriptions et surmonter les difficultés posées par les différents degrés de mélange linguistique serait de laisser les lemmatiseurs départager les langues en comparant le nombre de termes non identifiés (*unknows*) pour ne retenir, *in fine*, que la version la plus « efficace ».

23 Plusieurs autres expériences sont en cours, mais il est temps de conclure.

Conclusion

24 En guise de conclusion, nous soulignerons d'abord que l'acquisition d'un corpus documenté est le premier produit « fini » de la recherche en tant que résultat des nombreuses opérations et choix réalisés par les chercheurs, en tant aussi que démarche heuristique. Ensuite, la pratique montre qu'un corpus peut être constamment réinvesti, tant son exploitation, grâce à l'intelligence artificielle, génère des nouvelles données propres à être réinjectées dans le corpus et à leur tour exploitées, constituant une sorte de cercle vertueux. Nous pouvons ainsi tirer tout le parti d'une dynamique où l'acquisition, la documentation et l'exploitation des corpus ne sont pas des étapes successives à connecter mais des opérations co-actives sans cesse réinvesties et réalimentées.

NOTES

1. E. MAGNANI, « Des *chartae* au corpus : la plateforme des CBMA - *Chartae/Corpus Burgundiae Medii Aevi* », in C. BALOUZAT-LOUBET (dir.), *Digitizing Medieval Sources. Challanges and Methodologies /L'édition en ligne de documents d'archives médiévaux. Enjeux, méthodologie et défis*, Turnhout, 2019 (Atelier de Recherches sur les Textes Médiévaux, 27), p. 57-67 - https://doi.org/10.1484/M.ARTEM-EB.5.117328

2. E. MAGNANI, « Un corpus structuré et hétérogène de textes latins médiévaux (Bourgogne, V^e^-XV^e^ siècle) », *Bulletin du CERCOR - Centre Européen de recherches sur les congrégations et ordres*

religieux, 41, 2017, p. 59-65. (ISSN 1243-3217) - https://halshs.archives-ouvertes.fr/halshs-01529451

3. E. MAGNANI, « Les CBMA en corpus structuré. Atelier 2. Le corpus hagiographique bourguignon. Débats et recherches », *Bulletin du centre d'études médiévales d'Auxerre, BUCEMA*, Collection CBMA, Les journées d'études, mis en ligne le 28 juillet 2018, URL : http://journals.openedition.org/cem/15493

4. E. MAGNANI, E. INGRAND-VARENNE, « Le corpus épigraphique bourguignon (VIII[e]-XV[e] siècle). Des catalogues aux applications numériques », , *Bulletin du centre d'études médiévales d'Auxerre, BUCEMA*, Collection CBMA, Les journées d'études, mis en ligne le 15 novembre 2018, consulté le 06 décembre 2018. URL : http://journals.openedition.org/cem/15591

5. Plus exactement les départements des Alpes-de-Haute-Provence (04), Hautes-Alpes (05), Alpes-Maritimes (06), Ardèche (07), Bouches-du-Rhône (13), Drôme (26), Gard (30), Lozère (48), Var (83) et Vaucluse (84).

6. A. MENUDIER, « Le corpus épigraphique provençal : premier bilan et comparaison avec le corpus bourguignon », *Bulletin du centre d'études médiévales d'Auxerre, BUCEMA*, Collection CBMA, Les journées d'études, mis en ligne le 19 mai 2020. URL : http://journals.openedition.org/cem/17076

7. Pour comparaison, la *Patrologie Latine* contient environ 95 931 178 mots.

8. http://www.cbma-project.eu/

9. https://gitlab.huma-num.fr/cbma/hagiographie/ ; https://gitlab.huma-num.fr/lamop/cbma-epigraphie

10. http://textometrie.ens-lyon.fr/

11. https://artfl-project.uchicago.edu/philologic4

12. https://philologic.lamop.fr/cbma/ ; https://philologic.lamop.fr/epigraphie/ ; https://philologic.lamop.fr/hagiographie/

13. https://nosketch-engine.lamop.fr/run.cgi/first_form . Nous remercions Krzysztof Nowak pour son aide dans la mise en place de cette plateforme.

14. https://manuscrits.cbma-project.eu/ . Les dix dernières incorporations, cartulaires et livres de comptes sont : Premier cartulaire de Saint-Étienne de Dijon (XII[e] siècle), Archives départementales de la Côte-d'Or, Cart. 21 (G 125), 71 folios ; Cartulaire de Cîteaux pour les domaines de Saint-Jean de Losne et Losne (XVIII[e] siècle), Archives départementales de la Côte-d'Or, Cart. 198-01 (11 H 1055), 185 folios ; Ordre cistercien : comptes. Secundum registrum monasteriorum ordinis Cisterciensis (1354), Archives départementales de la Côte-d'Or, Cart. 170 (11 H 1159) ; Ordre cistercien et abbaye de Cîteaux : comptes pour l'ordre (1337-1347), comptes du boursier (1337-1402), Archives départementales de la Côte-d'Or, 11 H 1160 ; Ordre cistercien : comptes (1403-1423), Archives départementales de la Côte-d'Or, 11 H 1161 ; Ordre cistercien : comptes (1499-1515), Archives départementales de la Côte-d'Or, 11 H 1162 ; Boursier de Cîteaux : comptes (1489-1493), Archives départementales de la Côte-d'Or, 11 H 1166 ; Grenetier de Cîteaux : comptes (1497-1504), Archives départementales de la Côte-d'Or, 11 H 1172 ; Grenetier de Cîteaux : comptes (1499-1503), Archives départementales de la Côte-d'Or, 11 H 1173 ; Comptes du Petit-Cîteaux (1477-1491), Archives départementales de la Côte-d'Or, 11 H 1179.

15. E. MAGNANI, N. PERREAUX, « A Medieval Epigraphic Corpus and its Retro-Developments (CIFM-CBMA). The Exploratory Research of the COSME[2] Consortium » (à paraître).

INDEX

Index géographique : France/Bourgogne
Mots-clés : corpus, diplomatique, épigraphie, hagiographie, humanités numériques
Index chronologique : Moyen Âge

AUTEUR

ELIANA MAGNANI

CNRS – LAMOP