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

24.2 | 2020
Varia

La stylométrie comme outil pour la recherche de l'élaboration des chartes médiévales. Le cas de Cambrai au douzième siècle (1131-1200)

Eveline Leclercq

Édition électronique
URL : http://journals.openedition.org/cem/17772
DOI : 10.4000/cem.17772
ISSN : 1954-3093

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

Référence électronique
Eveline Leclercq, « La stylométrie comme outil pour la recherche de l'élaboration des chartes médiévales. Le cas de Cambrai au douzième siècle (1131-1200) », *Bulletin du centre d'études médiévales d'Auxerre | BUCEMA* [En ligne], 24.2 | 2020, mis en ligne le 19 décembre 2020, consulté le 17 janvier 2021. URL : http://journals.openedition.org/cem/17772 ; DOI : https://doi.org/10.4000/cem.17772

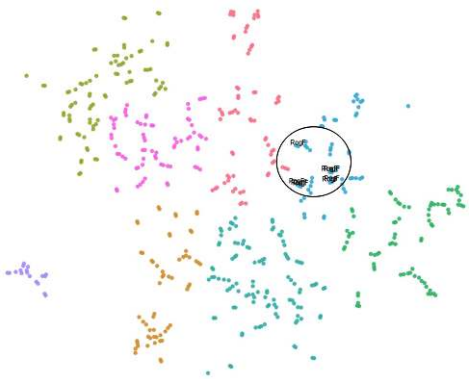
Ce document a été généré automatiquement le 17 janvier 2021.

Les contenus du *Bulletin du centre d'études médiévales d'Auxerre (BUCEMA)* sont mis à disposition selon les termes de la Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Partage dans les Mêmes Conditions 4.0 International.

La stylométrie comme outil pour la recherche de l'élaboration des chartes médiévales. Le cas de Cambrai au douzième siècle (1131-1200)

Eveline Leclercq

Introduction : le développement de la méthodologie



1 La démarche numérique récente dans les humanités a mené au développement de plusieurs logiciels et méthodes. Dans les domaines de l'histoire médiévale et de la diplomatique en particulier, ceux-ci s'intéressent à la paléographie numérique[1], ce qui a soulevé des questions sur les possibilités et les limites des ordinateurs[2]. Alors que la paléographie numérique a évolué vers une discipline presque indépendante, l'étude du dictamen s'est développée à un rythme plus lent. L'étude des formules s'est étendue en allant de l'analyse des phrases protocolaires à l'examen des expressions et des combinaisons de mots plus courtes dans le texte complet (méthode De Paermentier[3]). Dans cette dernière méthode, l'approche numérique s'est principalement limitée à l'usage de bases de données en ligne.

2 Pourtant, dans la littérature médiévale, plusieurs chercheurs ont adopté des méthodes venant de la *stylométrie*[4]. Ce domaine utilise des méthodes numériques pour analyser des textes, afin de déterminer le style d'un rédacteur. Ainsi, la stylométrie pourrait s'avérer intéressante dans le développement des humanités numériques notamment pour l'étude du *dictamen* des chartes médiévales. Idéalement, la recherche pourrait en partie s'automatiser ; néanmoins, l'analyse par l'ordinateur n'est pas toujours fiable à cent pour cent et les résultats présentés nécessitent toujours l'interprétation du chercheur. Comme il s'agit de textes à caractère particulier, quelques réserves doivent être prises en considération avant de pouvoir déterminer dans quelle mesure la stylométrie peut s'appliquer aux chartes médiévales.

La stylométrie et les chartes médiévales

3 Premièrement, la taille limitée des chartes, contenant entre 200 et 400 mots, pourrait poser problème. Les recherches en stylométrie s'appliquent le plus souvent à l'étude des textes littéraires nettement plus longs ; des échantillons comprenant un minimum de 2 000 mots y sont considérés comme étant courts[5]. Plus récemment, des textes plus réduits, comme les blogues, les courriels électroniques, les coupures de presse et les *tweets* ont été pris en compte[6]. K. Luyckx et W. Daelemans, en particulier, ont testé l'influence de la taille des échantillons utilisés sur les résultats des techniques statistiques du type *Memory Based Learning*[7]. Pour des textes très courts (140 mots), K. Luyckx et W. Daelemans ont découvert que les trigrammes – à base de groupes de trois lettres – et les techniques concernant les caractéristiques lexicales obtiennent les meilleurs résultats. Par ailleurs, si le nombre de rédacteurs possibles devient trop grand (quarante ou plus), la performance des logiciels baisse considérablement. Les textes plus courts contenant moins d'informations pour servir de base à l'étude, il n'est pas étonnant que la taille des mises par écrit constitue un défi pour certaines méthodes statistiques[8]. G. Hirst et O. Feiguina ont ainsi proposé de mieux utiliser les informations des textes courts en étudiant leur syntaxe[9], car celle-ci est moins facile à manipuler et varie donc peu dans des textes différents, produits par un même rédacteur[10] :

> High accuracies are achieved even with relatively small fragments of text (little more than 200 words), though the smaller the fragment, the greater the accuracy is boosted by the use of additional lexical features, including (unigram) part-of-speech frequencies[11].

4 Le second défi particulier des chartes concerne leur caractère formulaïque[12]. Cet aspect ne risque-t-il pas de dissimuler le style personnel des chartes pour les logiciels ? Il s'agit là d'une nouvelle interprétation de la question de l'analyse du *dictamen*, concernant la voix personnelle d'un rédacteur. Dans le domaine de l'étude du dictamen, un certain degré de personnalisation est accepté :

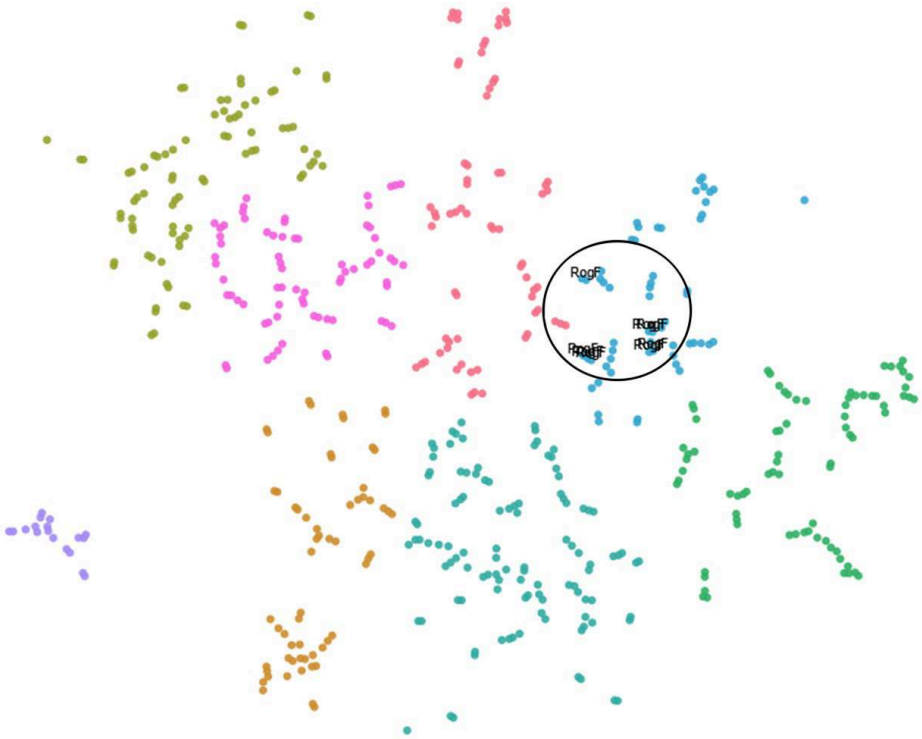
> La première supposition est que le rédacteur avait le libre choix du formulaire disponible. [...] D'autre part, ces règles impliquent que chaque rédacteur avait des préférences, de sorte qu'il n'utilisait que quelques tournures de phrase faisant partie d'un formulaire plus étendu[13].

5 Au lieu de confirmer ou de réfuter la rédaction d'un texte par tel ou tel rédacteur, nous voudrions dans un premier temps contrôler s'il est possible d'utiliser la stylométrie pour l'analyse du *dictamen*, même sans identifications préalables. Le diplomatiste est notamment confronté à un grand groupe de rédacteurs dont le nombre exact est souvent inconnu. Par conséquent, il serait plus efficace de permettre à l'outil informatique de mettre en lumière les textes qui sont proches les uns des autres afin d'être comparés ensuite pour déterminer s'il s'agit du même *dictamen*.

Un premier test pour la stylométrie comme outil du diplomatiste

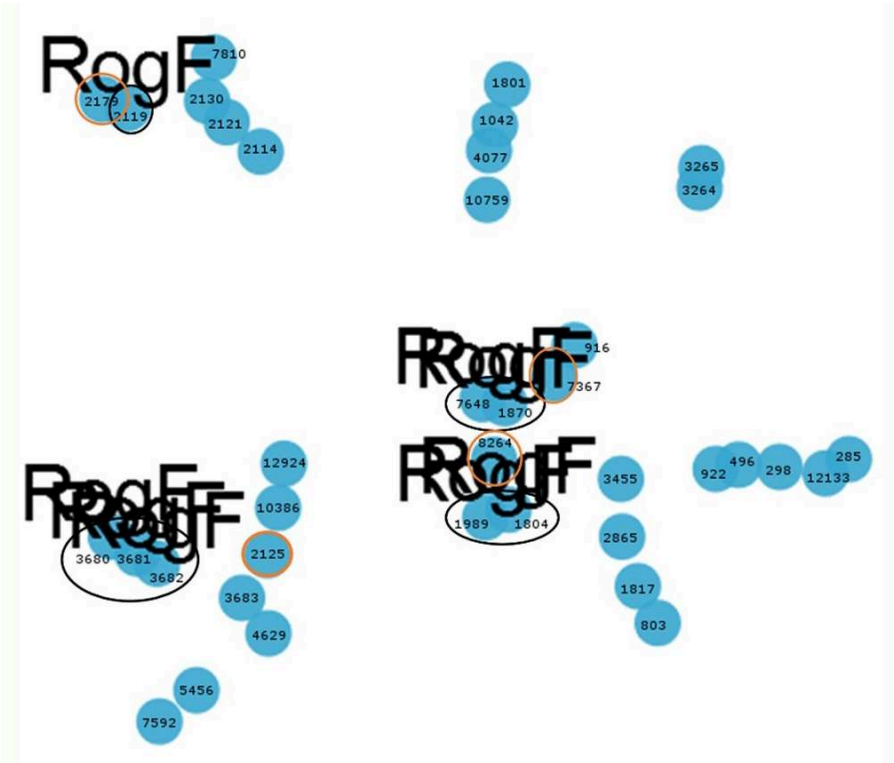
6 La méthode a été appliquée pour la première fois aux chartes médiévales dans le cas d'un rédacteur cambrésien anonyme, opérant pendant les épiscopats de Roger de Wavrin (1179-1191) et de Jean II d'Antoing (1192-1196)[14]. Pour le test, et sur la base de l'analyse paléographique, ce rédacteur est désigné par le nom RogF/JeanE (1187-1196)[15]. Un corpus de 546 textes disponibles dans les bases de données *Diplomata Belgica*[16] et *Chartae Galliae*[17] a été encodé en texte brut. Ce corpus comprend les chartes des épiscopats de Roger de Wavrin et de Jean II d'Antoing, dont l'évêque était l'auteur, ainsi que les chartes émanant des 23 destinataires présents dans ce corpus. Un code adapté pour les chartes médiévales a été développé par M. Kestemont, appliquant deux méthodes à la fois[18] : une représentation visuelle des chartes en nuage de points selon le principe du « voisin le plus proche » – t-SNE ou *T-distributed Stochastic Neighbor Embedding*[19] –, dont la figure 1 donne un exemple (fig. 1), et *Simple Text Reuse Detection*, qui détecte automatiquement les parties identiques des deux textes (fig. 2). Celle-ci révèle souvent des parties de texte substantielles, des formules ou des tours de phrases réutilisés. Pour demander efficacement à l'ordinateur de rechercher l'apport personnel du rédacteur dans le texte, il revient au chercheur de choisir la meilleure méthode statistique afin d'obtenir des résultats fiables. Les techniques apportées par la stylométrie se sont avérées très prometteuses puisqu'elles nous ont permis d'ajouter des chartes déterminées comme des produits de RogF/JeanE à notre liste initiale (cf. fig. 2).

Fig. 1 – Cette figure montre la représentation visuelle de l'analyse stylométrique du corpus de l'étude de cas de RogF/JeanE.



L'ordinateur regroupe les actes ayant des caractéristiques stylistiques similaires en une même couleur, chaque point correspondant à un acte. Nous voyons clairement que les chartes déjà attribuées à RogF/JeanE dans l'analyse du dictamen préliminaire (selon la méthode De Paermentier) sont regroupées.

Fig. 2 – Cette figure montre les chartes déterminées comme rédigées par RogF/JeanE pendant l'analyse du *dictamen* par la méthode De Paermentier (entouré de noir).



Les bulles entourées d'orange désignent les chartes qui, après l'analyse détaillée des textes – méthode De Paermentier et *Simple Text Reuse Detection* – se sont aussi avérées comme étant des rédactions de notre scribe-rédacteur RogF/JeanE[20]. L'analyse a donc révélé quatre actes supplémentaires.

7 Bien que nous ayons prouvé l'utilité de la stylométrie dans le domaine de la diplomatique médiévale, notre travail n'est pas achevé. Le chercheur ne peut pas perdre de vue la nécessité de l'expertise et de la connaissance de son corpus dans son analyse ; l'ordinateur ne fait pas le travail complet. Le plus grand défi pour l'examen de ces méthodes réside dans le côté technique. Le code développé par M. Kestemont est sujet aux évolutions rapides du monde numérique et risque d'être vite dépassé, mais actuellement il reste un instrument utile pour conduire d'autres tests. Des analyses continues peuvent également inclure d'autres éléments à contrôler. En premier lieu, nous considérons l'influence du type de charte. Dans le domaine de la littérature médiévale, il est devenu clair que le style d'un rédacteur pouvait être fortement influencé par le genre de texte[21]. De même, l'influence de la composition du corpus reste à être prise en compte, étant donné que la présence des chartes des grands auteurs comme le pape, les ducs de Brabant et les comtes de Flandre influence les résultats jusqu'à un certain degré. D'autre part, cette même présence a permis de lier certaines chartes à d'autres. On pourrait ainsi étudier le corpus comme un réseau d'échanges de formules.

L'approfondissement de la méthode : les chartes comme un réseau

8 La reprise d'autres études de cas et du corpus complet des chartes épiscopales de Cambrai (1131-1200) pourrait peut-être éclairer le lien entre le scribe et le rédacteur au sein de cette chancellerie, en particulier par l'ajout des résultats d'E. Van Mingroot et de N. Barré[22]. Actuellement, nous travaillons à une communication intitulée « L'usage personnel de formules de rédaction : une analyse stylométrique des chartes épiscopales de Cambrai à la recherche des rédacteurs individuels (1131-1200) », initialement prévue pour le colloque international sur la formule au Moyen Âge à Paris en juin 2020, mais reportée à juin 2021. Pour cette recherche en cours, nous avons encodé et collationné 4 063 textes sur un corpus d'à peu près 6 000 chartes répertoriées dans les bases de données *Diplomata Belgica* et *Chartae Galliae.* Notre approche reste identique à celle adoptée pour l'étude du rédacteur cambrésien RogF/JeanE. Le code développé par M. Kestemont reste en usage et les chartes sont représentées par nuages de points. Au lieu de partir d'un rédacteur présumé (RogF/JeanE), nous analyserons le regroupement général des chartes et leur *dictamen* individuel, en nous focalisant premièrement sur les chartes épiscopales afin de tenter de dévoiler d'autres rédacteurs actifs dans la chancellerie cambrésienne pour cette période. Les vingt mains déterminées pendant l'étude paléographique de ce bureau d'écriture serviront à contrôler dans quelle mesure on pourrait lier les rédacteurs aux scribes. Durant cette analyse, nous tenons compte d'éventuelles adaptations nécessaires afin de pouvoir mieux intégrer cette méthode dans la diplomatique médiévale et ainsi, peut-être, contribuer au développement des humanités numériques.

Reçu : 25 août 2020 – Accepté : 24 septembre 2020

NOTES

1. M. KESTEMONT, G. FIELDING et R. CONN, « Writer Identification by Professional Document Examiners », *Journal of Forensic Sciences*, 42 (1997), p. 778-786. F. CLOPPET *et al.*, « New Tools for exploring, Analysing and Categorizing Medieval Scripts », *Digital Medievalist*, 7 (2011), en ligne [https://journal.digitalmedievalist.org/articles/10.16995/dm.44/], consulté le 26 novembre 2020.

2. P. A. STOKES, « Palaeography and Image-Processing : Some Solutions and Problems », *Digital Medievalist*, 3 (2007-2008), en ligne [https://journal.digitalmedievalist.org/articles/10.16995/dm.15/], consulté le 26 novembre 2020. P. A. STOKES, « Computer-aided Palaeography, Present and Future », *in* M. REHBEIN, P. SAHLE et T. SCHASSEN (dir.), *Kodikologie und Paläographie im digitalen Zeitalter*, Norderstedt, 2009, p. 309-338.

3. E. DE PAERMENTIER, *In cuius rei testimonium et firmitatem. Oorkonden en kanselarijwerking in de entourage van de graven en gravinnen van Vlaanderen en Henegouwen (1191-1244). Een diplomatische en paleografische studie*, thèse de doctorat, université de Gand, 2011, p. 48-67.

4. M. KESTEMONT, « Stylometry for Medieval Authorship Studies. An Application to Rhyme Words », *Digital Philology : A Journal of Medieval Cultures*, 1/1 (2012), p. 42-72. M. KESTEMONT, « What

Can Stylometry Learn From Its Application to Middle Dutch Literature ? », *Journal of Dutch Literature*, 2/2 (2011), en ligne [https://www.journalofdutchliterature.org/index.php/jdl/article/view/21], consulté le 26 novembre 2020. M. KESTEMONT, S. MOENS et J. DEPLOIGE, « Stylometry and the Complex Authorship on Hildegard of Bingen's Œuvre », *Digital Humanities 2013. Conference Abstracts*, p. 255-258, en ligne [https://biblio.ugent.be/publication/3119482], consulté le 26 novembre 2020.

5. G. FRANZINI, M. KESTEMONT *et al.*, « Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm », *Frontiers in Digital Humanities*, 5 avril 2018, p. 2, en ligne [https://www.frontiersin.org/articles/10.3389/fdigh.2018.00004/full], consulté le 26 novembre 2020.

6. K. LUYCKX et W. DAELEMANS, « The effect of author set size and data size in authorship attribution », *Literary & Linguistic Computing*, 25 (2010), p. 35. M. KOPPEL, J. SCHLER et S. ARGAMON, « Authorship Attribution in the wild », *Language Resources and Evaluation*, 45/1 (2011), p. 83-94. Un bon aperçu des travaux menés dans ce cadre se trouve dans G. HIRST et O. FEIGUINA, « Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts », *Literary & Linguistics Computing*, 22/4 (2007), p. 405-406. J. DIEDERICH *et al.,* « Authorship attribution with support vector machines », p. 109-123, *Applied Intelligence*, 19/1-2 (2003), p. 109-123, en ligne [https://www.researchgate.net/publication/43456924_Authorship_Attribution_with_Support_Vector_Machines], consulté le 26 novembre 2020. C. SANDERSON et S. GUENTER, « Short Text Authorship via Sequence Kernels, Markov Chains and Author Unmasking : An Investigation », p. 482-491, *2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), University of Technology, Sydney, Australia, 22-23 July 2006*, p. 482-491, en ligne [https://dl.acm.org/citation.cfm?id=1610142], consulté le 26 novembre 2020.

7. *Memory Based Learning* [MBL] est une méthode simple d'approximation. On entraîne ce genre de programme en mettant toutes les données dans une base de données et en laissant le logiciel trouver les points communs. Ces derniers forment le modèle à partir duquel le logiciel fait ses classifications. Plusieurs logiciels de la stylométrie fonctionnent ainsi. K. LUYCKX et W. DAELEMANS, « The effect of author… », *op. cit.*, p. 35-55.

8. M. KESTEMONT, « Stylometry for Medieval Authorship Studies : An Application to Rhyme Words », *Digital Philology : A Journal of Medieval Cultures*, 1/1 (2012), p. 61.

9. G. HIRST et O. FEIGUINA, « Bigrams of Syntactic Labels… », *op. cit.*, p. 405-417.

10. H. BAAYEN, H. VAN HALTEREN et F. TWEEDIE, « Outside the cave of shadows : Using syntactic annotation to enhance authorship attribution », *Literary & Linguist Computing*, 11/3 (1996), p. 121.

11. G. HIRST et O. FEIGUINA, « Bigrams of Syntactic Labels… », *op. cit.*, p. 415.

12. Nous utilisons ce terme du domaine de la littérature pour exprimer que nos textes sont presque entièrement composés de formules. H. CAZES, « Chapitre 2 : Démonstrations d'amitié et d'humanisme : *alba,* adages et emblèmes chez les petits-enfants d'Érasme », *in* A. STEINER-WEBER et K. ENENKEL, (dir.), *Proceedings of the Fifteenth International Congress of Neo-Latin Studies (Münster 2012)*, 2005, Leiden, p. 25 : « voire, seule la partie formulaïque de ces inscriptions semble personnelle : citations, devises, proverbes occupent le corps même de ces déclarations publiques. »

13. Traduction personnelle d'E. C. DIJKHOF, *Het Oorkondenwezen van enige kloosters en steden in Holland en Zeeland, 1200-1325*, Leuven, 2003, p. 50 : « De eerste veronderstelling is dat een dictator vrijelijk een keuze kon maken uit de voorhanden sjablonen. […] Impliciet veronderstellen de regels echter ook dat iedere dictator een voorkeur bezat en daarom slechts enkele zinswendingen gebruikte van de vele die hem ter beschikking stonden. »

14. Actuellement, un article joint avec M. Kestemont, qui détaille le cas de RogF/JeanE et la méthode utilisée dans ce bêta-test, est en voie de publication dans le journal *Interfaces : A Journal of Medieval European Literatures*, en ligne [https://riviste.unimi.it/interfaces], consulté le 26 novembre 2020.

15. Nous avons, en effet, retrouvé une même écriture pendant les épiscopats de Roger de Wavrin (RogF, chronologiquement la sixième main dans les chartes) et de Jean d'Antoing (JeanE, chronologiquement la cinquième main dans les chartes).

16. T. DE HEMPTINNE, J. DEPLOIGE, J.-L. KUPPER et W. PREVENIER (dir.), *Diplomata Belgica. Les sources diplomatiques des Pays-Bas méridionaux au Moyen Âge*, en ligne [http://www.diplomata-belgica.be], consulté le 26 novembre 2020.

17. P. BERTRAND, B.-M. TOCK, *et al., Chartae Galliae*, en ligne [http://telma.irht.cnrs.fr/chartes/chartae-galliae/index/], consulté le 26 novembre 2020.

18. Il faut bien souligner qu'il n'existe pas de méthode unique dans le domaine de la stylométrie. Une myriade de techniques statistiques à base de plusieurs éléments de la langue (la syntaxe, les bi- et trigrammes, la fréquence des mots-outils, etc.) est disponible. Un état de la question des méthodes exhaustif n'est pas possible dans le cadre de cet article ; nous renvoyons à E. LECLERCQ, *L'élaboration des chartes médiévales. L'exemple des évêchés d'Arras, Cambrai et Liège (XI^e^-XII^e^ siècles)*, thèse de doctorat, université de Strasbourg, 2019, p. 317-331.

19. L. VAN DER MAATEN, « Visualizing Data using t-SNE », *Journal of Machine Learning Research*, 9 (2008), p. 2579-2605, en ligne [https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf], consulté le 26 novembre 2020.

20. Les numéros accompagnant les bulles renvoient aux numéros des chartes dans les *Diplomata Belgica* et les *Chartae Galliae*.

21. H. VAN HALTEREN *et al.*, « New Machine Learning Methods Demonstrate the Existence of a Human Stylome », *Journal of Quantitative Linguistics*, 12/1 (2005), p. 65-77.

22. E. VAN MINGROOT, *De bisschoppelijke kanselarij te Kamerijk 1057-1130*, thèse de doctorat, Département d'Histoire, K. U. Leuven, 1969. N. BARRÉ, « Chancellerie épiscopale, chancellerie canoniale : unicité ou pluralité des institutions à Cambrai au XII^e^ siècle ? », *Bulletin de la Commission royale d'Histoire*, 176/2 (2010), p. 129-146.

INDEX

Mots-clés : diplomatique, chartes épiscopales, stylométrie, méthodologie
Index géographique : Cambrai

AUTEUR

EVELINE LECLERCQ

Université de Strasbourg