Université Libre de Bruxelles

Faculté de Lettres, Traduction et Communication

Étude stylométrique de l'œuvre de Gerbert d'Aurillac

Guillaume Quintin

Mémoire présenté sous la direction de Sébastien De Valeriola en vue de l'obtention du titre de Master en Sciences et Technologies de l'Information et de la Communication

Année académique 2022–2023

Résumé

Informations

- Nom et prénom : Quintin Guillaume ;
- Filière : Master en Sciences et Technologies de l'Information et de la Communication ;
- Année académique : 2022-2023 ;
- Titre du mémoire : Étude stylométrique de l'œuvre de Gerbert d'Aurillac.

Mots-clés

Gerbert d'Aurillac, Sylvestre II, stylométrie, attribution de paternité, réseaux de co-occurrence de mots, *De informatione episcoporum*, *De utilitatibus astrolabii*.

Brève description

Gerbert d'Aurillac, considéré encore aujourd'hui comme une figure centrale de l'histoire des sciences, a donné son nom à des traités théologiques et scientifiques que l'historiographie tend désormais à lui contester. La stylométrie, qui malgré son apparence mathématique s'ancre dans une tradition d'études de paternité longtemps qualitatives, part du postulat que le style d'un auteur est mesurable statistiquement et que son empreinte stylistique se retrouve dans tout texte qu'il compose. Même si l'attribution de paternité quantitative a déjà été éprouvée sur du latin médiéval, jamais l'œuvre de Gerbert n'a été étudiée à l'aune des techniques stylométriques. Équipé d'une méthode qui transforme les textes en réseaux de co-occurrence de mots et utilise leurs mesures topologiques pour établir des profils stylistiques, nous questionnons la paternité de deux œuvres : *De informatione episcoporum* et *De utilitatibus astrolabii*. Malgré une efficacité relative de la méthode mise en œuvre, cette dernière s'est révélée incapable de répondre de façon satisfaisante aux questions de paternité que nous nous étions posées au début de notre recherche. Nous tentons de comprendre les difficultés rencontrées par les modèles et proposons de nouvelles pistes de recherche.

Table des matières

Table des figures

Chapitre 1

Introduction

Gerbert d'Aurillac : moine, écolâtre, scientifique, philosophe, archevêque non pas une mais deux fois, pape de l'an mil et… suppôt de Satan ? Ce personnage fascinant de la deuxième moitié du X^{e} siècle n'a pas manqué d'intriguer ses contemporains et, dès sa mort, le légendaire prend le pas sur la réalité historique et l'emmène jusqu'à la cour de Cordoue en al-Andalus. Quelle est la part de véridique dans la réputation qui l'entoure ? Difficile de le savoir définitivement, et les chercheurs n'ont pas fini de débattre sur ce qu'il faudrait attribuer à Gerbert et ce qui n'est que rumeur.

Le grand débat historiographique au sein duquel Gerbert se trouve plongé est celui de la transmission de la science arabe en direction de l'Occident latin. Certains ont pu attribuer à Gerbert un rôle prépondérant dans cette transmission, voyant notamment en lui le premier latin à utiliser les chiffres indo-arabes et l'imaginant comme introducteur d'outils d'origine arabes comme l'astrolabe. C'est dans ce contexte qu'un petit traité sur l'usage de l'astrolabe lui a parfois été attribué, parfois véhémentement refusé. Mais n'oublions pas que Gerbert n'est pas qu'un scientifique, il est avant tout un ecclésiastique et c'est au sein et autour de l'Église qu'il a effectué sa carrière. Il n'est guère surprenant, dans ces conditions, que des écrits à portée théologique et religieuse lui aient été également attribués à tort ou volontairement.

Un tel terrain de jeu, nous en sommes persuadé, éveillerait l'intérêt de n'importe quel stylométriste digne de ce nom. La stylométrie est l'étude quantitative et statistique du style d'un auteur. Tout individu possède des caractéristiques stylistiques qui lui sont propres et, pour le plupart, inconscientes. Le postulat de la stylométrie, c'est que ces caractéristiques peuvent être mesurées et comparées efficacement avec les caractéristiques d'autres individus. S'étant développé depuis le milieu du XX^{e} siècle et de façon particulièrement active depuis les années 1990, cette discipline s'ancre dans le domaine plus large des études de paternité, qui remontent jusqu'à l'antiquité et connaissent d'illustres ancêtres dans les personnes de Lorenzo Valla ou d'Abélard. La stylométrie s'appuie sur des éléments linguistiques qui semblent anodins, comme la fréquence des mots-outils (*function words*) ou les bigrammes de caractères que nous

employons. Le développement, depuis une vingtaine d'années, des méthodes d'apprentissage automatique au sein de la discipline a permis aux chercheurs de multiplier les caractéristiques linguistiques qu'ils prenaient en compte, en laissant de puissants algorithmes classer et mettre de l'ordre dans les données récoltées.

Notre objectif dans ce mémoire sera de nous familiariser avec la recherche sur Gerbert d'Aurillac et sur la stylométrie et de tenter de mettre en place une méthode d'attribution de paternité. Les ouvrages sur lesquels notre intérêt se porte tout particulièrement sont au nombre de deux. Le premier, intitulé *Sermo de informatione episcoporum*, est un traité sur la formation des évêques qui, notamment, critique la simonie. Même si la paternité de Gerbert a été largement contestée et son véritable auteur identifié, il nous a semblé pertinent d'atteindre quantitativement une conclusion que d'autres ont pu construire qualitativement. Le second, intitulé *De utilitatibus astrolabii* mais souvent plus connu sous le nom de *Liber de astrolabio*, est un traité d'astronomie empreint de culture et de connaissance scientifique arabes. Contrairement au premier texte, la paternité de Gerbert est encore largement débattue et aucun consensus ne semble se profiler à l'horizon, puisque les deux camps avancent des arguments très pertinents.

Nous avons découvert, plus tôt cette année, l'existence des réseaux de co-occurrence de mots et leur application à des questions stylométriques. Bien différente des techniques traditionnelles qui reposent la plupart du temps sur le concept de fréquence, l'analyse de réseaux linguistiques postule que leurs mesures topologiques permettent de capturer le style d'un auteur, et que les informations stylistiques ainsi recueillies sont différentes de celles obtenues au moyen de méthodes plus traditionnelles. Des techniques d'apprentissage automatique permettent de faire sens de toutes ces mesures ainsi obtenues, et en particulier de classer des textes en fonction de leur auteur. Nous mettrons en œuvre deux d'entre elles : KNN et SVM.

Notre mémoire se divisera en quatre parties principales. Premièrement, nous présenterons Gerbert d'Aurillac en tant que personnage historique, ainsi que ses œuvres et les débats qui animent la communauté scientifique à leur sujet. Nous évoquerons brièvement la notion d'auteur, et dans quelle mesure elle doit être conçue différemment à l'époque médiévale. Deuxièmement, nous découvrirons la stylométrie et les études de paternité dans lesquelles elle s'insère. Nous évoquerons les jalons qui ont construit cette discipline, en pointant également des critiques qui ont pu lui être adressées. Nous nous plongerons en particulier dans l'application des techniques stylométriques à la langue latine et à l'utilisation de l'analyse de réseaux pour modéliser la langue et capturer le style d'un auteur. Troisièmement, nous détaillerons notre méthodologie, depuis l'acquisition des textes jusqu'à l'obtention de résultats par le truchement d'algorithmes d'apprentissage automatique. Le code R qui nous a permis de mettre en pratique notre méthode, attaché en annexe de notre mémoire, sera

également commenté et explicité dans cette partie de notre travail. Quatrièmement, nous mettrons en œuvre notre méthodologie et discuterons des résultats obtenus, tant à propos de l'efficacité générale de la méthode que des conclusions qu'il serait possible de tirer au sujet des deux écrits disputés que nous avons mis en évidence ci-dessus.

Chapitre 2

Gerbert d'Aurillac : sa vie, son temps, son œuvre

Dans ce premier chapitre, nous présenterons la figure historique que représente Gerbert d'Aurillac. Nous commencerons par évaluer les sources qui nous parlent de Gerbert, tant celles de son époque et des siècles ultérieurs que la littérature scientifique contemporaine. Nous évoquerons ensuite sa vie, en la découpant en six périodes que nous considérons comme possédant leur identité propre et distinctes les unes des autres. La production littéraire de Gerbert sera elle aussi abordée, en dressant une liste de ses œuvres mais surtout en détaillant les débats historiographiques sur deux ouvrages dont la paternité est mise en doute. Nous terminerons ce chapitre avec quelques considérations théoriques sur la notion d'auteur, en particulier durant la période médiévale.

2.1 Les sources

La vie de Gerbert nous est connue par quelques sources contemporaines, avant que ne se développe une légende autour de sa personne, très vite après sa mort. La littérature scientifique tente de démêler le vrai du faux dans cette réputation que Gerbert a acquis contre son gré.

2.1.1 Sources primaires

La vie de Gerbert nous est connue d'une part grâce à un corpus épistolaire conséquent,[1] qui s'étend de 983 à 997, d'autre part grâce aux écrits de son disciple Richer de Saint-Rémy.[2]

[1] Gerbert d'Aurillac. *Correspondance*. Sous la dir. de Pierre Riché et Jean-Pierre Callu. 2 t. Paris : Les Belles Lettres, 1993.

[2] Richer de Saint-Rémy. *Histoire de France (888-995)*. Sous la dir. de Robert Latouche. 2 t. Paris : H. Champion, 1930.

Très rapidement après sa mort, des rumeurs ont circulé sur Gerbert d'Aurillac. Son contemporain Adhémar de Chabannes, que nous reverrons par la suite, lui prête un séjour en al-Andalus, à la cour du calife omeyyade de Cordoue. Plus tard, les légendes ne font qu'enfler et donnent naissance à une véritable légende noire : sa vie exceptionnelle ne serait pas naturelle, mais le résultat d'un pacte avec le Diable. La naissance de cette légende semble se trouver au XII^{e} siècle, sous la plume de Guillaume de Malmesbury.[3]

Cette légende noire n'est pas la seule réputation que Gerbert a obtenu *post mortem*. Il a en effet été considéré comme une personnalité scientifique de premier plan et pour ainsi dire une autorité en la matière. Ses écrits scientifiques ont été recopiés de nombreux siècles après sa mort (jusqu'au début de l'époque moderne) et des traités postérieurs lui ont été attribués, vraisemblablement au nom de la réputation qu'il avait déjà à cette époque. G. Beaujouan, lors du premier colloque de Bobbio (voir infra), formule en quelques mots – repris depuis lors de nombreuses fois dans la littérature scientifique – l'essence du mystère incarné par Gerbert en la matière :

> « Dans le cas de Gerbert, il y a un déconcertant contraste entre sa réputation d'introducteur de la science arabe en Occident et la quasi-inexistence de traces d'influence arabe dans ses écrits indubitablement authentiques. »[4]

Gerbert d'Aurillac est encore considéré aujourd'hui comme une figure clé de l'histoire des sciences, même si les recherches les plus récentes soulignent davantage son rôle d'incitateur et de pédagogue, et moins sa production scientifique à proprement parler. Nous pensons notamment à des articles réalisés par des scientifiques ou publiés dans des revues de sciences naturelles.[5]

2.1.2 Littérature scientifique

Il serait indigeste d'évoquer dans cette partie la masse de contributions à l'étude du personnage de Gerbert d'Aurillac depuis le début du XX^{e} siècle. Nous évoquerons

[3] Marco Zuccato. « Gerbert of Aurillac and a Tenth-Century Jewish Channel for the Transmission of Arabic Science to the West ». In : *Speculum* 80.3 (2005), p. 746-750.

[4] Guy Beaujouan. « Les Apocryphes mathématiques de Gerbert ». In : *"Gerberto : scienza, storia e mito" : atti del Gerberti Symposium (Bobbio 25-27 luglio 1983)*. Sous la dir. de Michele Tosi. Bobbio : Archivi storici bobiensi, 1985, p. 646.

[5] Voir entre autres, durant cette dernière décennie : Costantino Sigismondi. « Gerbert of Aurillac : Astronomy and Geometry in Tenth Century Europe ». In : *International Journal of Modern Physics : Conference Series* 23 (2013), p. 467-471. arXiv : 1201.6094 ; Carlo Bianchini et Luca J. Senatore. « Gerbert of Aurillac (c. 940–1003) ». In : *Distinguished Figures in Descriptive Geometry and Its Applications for Mechanism Science : From the Middle Ages to the 17th Century*. Sous la dir. de Michela Cigola. History of Mechanism and Machine Science 30. Cham : Springer International Publishing, 2016, p. 33-51 ; Marek Otisk. « Gerbert of Aurillac (Pope Sylvester II) as a Clockmaker ». In : *Teorie vědy/Theory of Science* 42.1 (2020), p. 25-49.

la plupart de nos références au cours de notre exposé. Il nous paraît néanmoins opportun de souligner quelques ouvrages, articles et contributions diverses qui nous ont particulièrement nourri.

Il est nécessaire de commencer par la monographie de référence sur Gerbert d'Aurillac, rédigée par Pierre Riché,[6] qui nous sert de référence principale pour dresser la vie de Gerbert. Le même auteur a écrit un ouvrage sur la fin du X^{e} siècle,[7] qui permet d'y voir plus clair dans une période mouvementée. En outre, nous désirons revenir quelque peu sur un ouvrage de vulgarisation qui revient sur le côté scientifique de Gerbert.[8] Nous trouvons en effet que nous avons été dur avec cet ouvrage lors de notre travail préparatoire. Il est vrai que la partie dédiée à la vie de Gerbert est par moment sensationnaliste, mais une relecture de la partie consacrée aux aspects scientifiques après avoir consulté la myriade de références à ce sujet a mis en lumière la qualité de synthèse et de pédagogie de cet ouvrage.

Gerbert et son œuvre ont été au centre de quatre colloques organisés sur un vingtaine d'années, entre 1985 et 2005. Il est par ailleurs impressionnant de constater l'évolution dans le ton et les propos entre le premier colloque et le dernier, et l'éloignement d'une forme d'hypercriticisme qui caractérisait la fin du XX^{e} siècle. Trois de ces colloques se sont tenus à Bobbio[9] et un à Aurillac.[10] Lorsqu'il s'agira de les référencer, nous citerons directement les articles repris dans les actes des colloques.

2.2 D'humble oblat auvergnat au trône de saint Pierre

Dans les lignes qui suivent, nous établissons une brève biographie de Gerbert d'Aurillac, en nous basant comme nous l'avons évoqué sur la biographie de P. Riché. Le titre de cette section est, nous l'avouons, quelque peu trompeur. En effet, il semble laisser entendre que le pontificat représente l'apogée de la vie de Gerbert. C'est par ailleurs le titre que P. Riché décide de donner à son chapitre sur cette période de la vie de Gerbert.[11] Or, même s'il est possible de considérer que, pour un ecclésiastique, l'accession au Saint-Siège soit l'aboutissement de sa vie, ce n'est pas la raison de la

[6] Pierre Riché. *Gerbert d'Aurillac : Le Pape de l'an Mil*. Paris : Fayard, 1987.

[7] Pierre Riché. *Les Grandeurs de l'an Mille*. Paris : Bartillat, 1999.

[8] Alain Schärlig. *Un Portrait de Gerbert d'Aurillac : Inventeur d'un Abaque, Utilisateur Précoce Des Chiffres Arabes, et Pape de l'an Mil*. Lausanne : Presses polytechniques et universitaires romandes, 2012.

[9] Michele Tosi, éd. *"Gerberto : scienza, storia e mito" : atti del Gerberti Symposium (Bobbio 25-27 luglio 1983)*. Bobbio : Archivi storici bobiensi, 1985 ; Flavio G. Nuvolone, éd. *Gerberto d'Aurillac da abate di Bobbio a papa dell'anno 1000 : atti del congresso internazionale, Bobbio, Auditorium di S. Chiara, 28-30 settembre 2000...* Bobbio : Associazione culturale Amici di Archivum Bobiense, 2001 ; Flavio G. Nuvolone, éd. *Gerberto d'Aurillac - Silvestro II, Linee per Una Sintesi : Atti Del Convegno Internazionale, Bobbio, Auditorium Di S. Chiara, 11 Settembre 2004, Sotto La Presidenza Del Prof. Pierre Racine...* Bobbio : Associazione culturale Amici di Archivum bobiense, 2005.

[10] Nicole Charbonnel et Jean-Eric Iung, éd. *Gerbert l'Européen. Actes Du Colloque d'Aurillac, 4–7 Juin 1996*. Aurillac : Société des Lettres, Sciences et Arts "La Haute Auvergne", 1997.

[11] Riché, *Gerbert d'Aurillac*, p. 205-237.

renommée de Gerbert, que ce soit de son temps ou pour la postérité. En effet, Richer nous indique que l'écolâtre avait déjà une renommée internationale et, comme le souligne caustiquement A. Schärlig :

> « Si Gerbert d' Aurillac n'avait été que *le pape de l'an mil* – il l'a été sous le nom de Sylvestre II – sa notoriété ne serait pas plus grande que celle des autres papes de son époque : à peu près nulle. »[12]

Il serait par conséquent, à notre sens, malavisé de lire la vie de Gerbert d'Aurillac comme une trajectoire purement téléologique.

2.2.1 Du milieu du X^e^ siècle à 970 : période de formation

On ne connaît réellement ni la date ni le lieu de naissance de Gerbert d'Aurillac. La plupart des estimations indiquent le milieu du XX^e^ siècle, tandis que P. Riché suppose qu'il est né vers 945-950[13] : il était adolescent en 970 (selon Richer) et vieillard en 997 (selon ses propres lettres). Une fois ces valeurs converties, on obtient qu'il avait une vingtaine d'années en 970, car l'adolescence, concept flou, allait de 14 à 21 ans, voire un peu plus, et qu'il avait la cinquantaine en 997 (un âge où on se trouvait déjà dans la vieillesse à cette période). Richer nous indique qu'il est né en Aquitaine, région très vaste à l'époque, mais l'onomastique permet de le situer plus précisément en Auvergne, dans les environs d'Aurillac. Bref, comme de nombreux individus d'origine modeste, les précisions autour de son naissance sont perdues dans les affres du temps.

Très tôt dans sa vie, il est confié au monastère Saint-Géraud d'Aurillac, ce qui complète le nom sous lequel il est connu aujourd'hui. Gerbert y est un oblat, terme qui désigne un enfant offert à un monastère. Il y apprend le *trivium*, c'est-à-dire la grammaire, la dialectique et la rhétorique. Il est âgé d'environ vingt ans lorsque le comte de Barcelone, Borrell, est reçu au monastère. L'abbé lui propose que Gerbert l'accompagne chez lui, dans la Marche d'Espagne, ce qui deviendra la Catalogne.

C'est à ce moment de sa vie que les tenants de la légende noire de Gerbert insèrent une visite au cœur d'al-Andalus et un pacte avec le Diable. Nous ne discuterons pas de la deuxième accusation, mais soulignerons que la première est très peu vraisemblable. Des contacts étroits entre la Marche d'Espagne et al-Andalus semblent bien attestés,[14] ce qui rend tout à fait vraisemblable un Gerbert introduit aux connaissances scientifiques arabes lors de son séjour catalan, mais une expédition de près d'un millier de kilomètres vers le sud de la péninsule l'est beaucoup moins. Nous reviendrons

[12] Schärlig, *Un Portrait de Gerbert d'Aurillac*, p. 10.

[13] Riché, *Gerbert d'Aurillac*, p. 18.

[14] Voir Marco Zuccato. « Arabic Singing Girls, the Pope, and the Astrolabe : Arabic Science in Tenth-Century Latin Europe ». In : *Viator* 45.1 (2014), p. 99-120, sur lequel nous reviendrons.

sur ces considérations dans la section 2.3.2, car elles sont particulièrement importantes pour déterminer si Gerbert a pu ou non être l'auteur du traité sur les usages de l'astrolabe. Néanmoins, il est indéniable que Gerbert y a complété son éducation en matière de *quadrivium*, c'est-à-dire la musique, l'arithmétique, la géométrie et l'astronomie, puisqu'il surprendra les cours pontificale et impériale lors de son voyage à Rome en 970. À son arrivée en territoire catalan, il est confié à Hatton, évêque de Vich, maître d'une école et savant connu pour sa connaissance du *quadrivium*. Se trouvait également, non loin de Vich, le monastère Santa Maria de Ripoll, à la bibliothèque très riche en traités scientifiques et littéraires. Il est même possible d'envisager des contacts directement avec le monde scientifique arabe, puisqu'il existait à Barcelone au moins un clerc connaissant l'arabe et intéressé par les matières scientifiques : Sunefred Lobet, archidiacre de Barcelone.[15]

2.2.2 De 970 à 981 : écolâtre à Reims et secrétaire de l'archevêque

En 970, en compagnie de Borrell et d'Hatton, Gerbert quitte la Marche d'Espagne en direction de Rome, où il est présenté au pape Jean XIII et l'empereur Otton I[16]. Il les impressionne par ses connaissances, à tel point qu'Otton le désigne tuteur de son fils Otton II, une habitude qui deviendra héréditaire puisqu'il sera également précepteur du fils du deuxième Otton, Otton III. Il noue également des liens d'amitié avec Théophano, princesse byzantine et toute nouvelle épouse d'Otton II, et avec Adélaïde, épouse d'Otton I.

En 972 arrive à Rome Gerannus, archidiacre de Reims et savant en matière de dialectique. Gerbert et lui nouent des liens d'estime intellectuelle et amicale, et notre moine demande à l'empereur l'autorisation d'aller à Reims avec son nouveau compagnon. Le climat rémois lui a semble-t-il convenu, puisqu'il y restera dix ans et y reviendra. Il occupe alors, à partir de 973, la fonction d'écolâtre dans l'école-cathédrale de la ville, dont le rayonnement devient international sous la direction de Gerbert.[17] Ce dernier met en place des pratiques pédagogiques novatrices, en particulier dans le *quadrivium* : il l'enseigne par l'exemple et non sous l'égide d'autorités de l'antiquité classique, qui ne sont par ailleurs pas mentionnées par son disciple Richer. Son enseignement du *trivium* est, quant à lui, beaucoup plus traditionnel.[18] À côté de son activité d'écolâtre, il est également secrétaire de l'archevêque Adalbéron.

[15] Riché, *Gerbert d'Aurillac*, p. 26.

[16] Nous adoptons l'orthographe «Otton» et non «Othon» qui semble relativement répandue dans la littérature. Nous considérons que les trois premiers empereurs du Saint-Empire s'appellent Otton, tandis qu'Othon désigne le général romain brièvement empereur au I^er^ siècle de notre ère

[17] Riché, *Les Grandeurs de l'an Mille*, p. 175-176.

[18] Zuccato, «Arabic Singing Girls, the Pope, and the Astrolabe», p. 111.

2.2.3 De 981 à 984 : voyage en Italie et abbé de Bobbio

Fin 981, Gerbert et Adalbéron voyagent en Italie, à Ravenne. Ils y sont reçus par Otton II, empereur depuis la mort de son père en 973 mais pour la première fois en Italie depuis 972 (et sa rencontre avec Gerbert). La réputation de notre écolâtre le précède : Otric, écolâtre à Magdebourg en Saxe, avait accusé Gerbert d'une erreur à propos des parties de la philosophie et, plus généralement, de « ne rien avoir compris à la philosophie ».[19] L'empereur, après avoir appris cette accusation, a été étonné de cette erreur venant de Gerbert qu'il connaissait bien. Il profite d'avoir les deux écolâtres dans la même ville, à Ravenne, pour organiser un débat contradictoire en janvier 981. Gerbert en ressort clair vainqueur, prouvant sa grande maîtrise des sujets abordés.

Entre la fin de 981 et 982, Otton le nomme abbé de Saint-Colomban à Bobbio, peut-être comme récompense après le débat contradictoire de Ravenne. Ce très prestigieux monastère comportait notamment une vaste bibliothèque estimée à 690 volumes, dont Gerbert a dû profiter pendant la durée de son séjour. Toutefois, le contexte politique de Bobbio n'était pas aussi agréable que son environnement intellectuel. Confronté sans cesse aux nobles locaux qui ne voulaient entendre parler d'aucune des réformes amenées par Gerbert, ce dernier décide de quitter l'abbaye en 984, en laissant derrière lui un orgue construit de sa propre main ainsi que la vaste bibliothèque. Il restera abbé de Bobbio jusqu'à sa mort, sans y remettre toutefois les pieds.

2.2.4 De 984 à 991 : retour à Reims

Gerbert, de retour à Reims, reprend ses anciennes fonctions d'écolâtre et de secrétaire auprès de l'archevêque Adalbéron. Il n'oublie toutefois pas Bobbio, et se considère toujours comme « abbé et écolâtre » (lettre 142). Il reprend ses activités de chercheur, en demandant un traité d'astronomie à un certain Lupitus, identifié comme Sunefred Lobet, archidiacre de Barcelone que nous avons déjà précédemment mentionné. Il demande également un traité d'arithmétique à un autre de ses contacts catalans, l'évêque de Gérone. Gerbert a donc gardé des liens privilégiés avec la Marche d'Espagne, plus de quinze ans après son départ vers Rome.

En outre, notre écolâtre s'intéresse aux affaires politiques, qui ne sont pour le moins pas apaisées dans ce que P. Riché a appelé le « crépuscule carolingien »[20] : cette époque, dans la seconde moitié du X^{e} siècle, qui voit les Carolingiens s'effacer au profit des Robertiens en Francie occidentale et des Ottoniens en Francie orientale. En effet, l'archevêque de Reims est un seigneur autant spirituel que temporel, avec une place privilégiée dans un archidiocèse à cheval entre les deux Francies, et il n'est pas surprenant que son secrétaire s'implique lui aussi dans ces jeux de pouvoir. Gerbert

[19] Riché, *Gerbert d'Aurillac*, p. 59.
[20] Riché, *Les Grandeurs de l'an Mille*, p. 71.

a notamment œuvré aux côtés d'Adalbéron pour faire élire Hugues Capet, duc des Francs, sur le trône français après la mort du dernier carolingien régnant, Louis V. Le duc est élu avec succès par les grands du royaume lors de l'assemblée de Senlis en 987. Gerbert devient le secrétaire du nouveau roi, preuve de son implication dans la politique de son époque, même si P. Riché remet en question la réputation de Gerbert comme « faiseur de roi », comme le pensait Jules Michelet et comme semble le suggérer une lettre 163 de Gerbert, dans laquelle ses ennemis le nomment de la sorte.[21]

Un événement tragique vient bouleverser la vie de Gerbert en 989 : Adalbéron, pris de fièvre, rend son dernier souffle. Gerbert, effondré, ne succède pas à l'archiépiscopat. Le siège revient à Arnoul, clerc de Laon, fils bâtard de Lothaire et frère du dernier roi carolingien Louis V, qui promet de livrer sa ville à Hugues Capet s'il est élu archevêque de Reims. Gerbert prévient le roi du danger qu'un tel choix représente, mais Arnoul est désigné comme successeur d'Adalbéron. Il reprend donc sa fonction de secrétaire auprès du nouvel archevêque.

L'intuition de Gerbert ne tarde pas à se vérifier, puisqu'Arnoul se prépare à livrer Reims à son oncle Charles de Lorraine à peine quelques mois après son élection. Après avoir brièvement suivi Arnoul et Charles de Lorraine, Gerbert se rallie à Hugues Capet. Il est alors fait appel au pape Jean XV, autant par Hugues Capet que par les évêques de France (sans aucun doute dans les deux cas sous la plume de Gerbert). Face au silence du pontife, et après la défaite de Charles de Lorraine en 991, un procès se prépare pour juger l'archevêque félon. C'est le concile de Saint-Basle, lors duquel Gerbert joue un rôle prépondérant et qui s'achève par une condamnation et destitution d'Arnoul et l'élection subséquente de Gerbert à l'archiépiscopat qu'il convoitait tant.

2.2.5 De 991 à 999 : d'archevêque de Reims à archevêque de Ravenne

Le concile de Saint-Basle ne sera pas reconnu par la papauté, pas plus que l'élection de Gerbert. Débute une période de conflit ouvert entre Reims et Rome. En 995, deux synodes (à Mouzon puis à Reims) tentent de résoudre ces difficultés, sans succès. En 996, Gerbert se rend à Rome, où il rencontre le jeune empereur Otton III qui s'y fait couronner. Il prend brièvement une fonction de secrétaire auprès de l'empereur. Entretemps, Jean XV est passé de vie à trépas. Son successeur, Grégoire V, poursuit toutefois la politique de son prédécesseur à l'égard de Gerbert et considère son élection comme illégitime.

En 997, Grégoire V tient un synode à Pavie, lors duquel il est décidé de suspendre de leur fonction tous les évêques ayant pris part à la déposition d'Arnoul. Gerbert est

[21] Riché, *Gerbert d'Aurillac*, p. 100.

abandonné de tous. Le nouveau roi, Robert le Pieux, pourtant son ancien disciple, négocie avec la papauté pour annuler son excommunication en échange du retour d'Arnoul à l'archiépiscopat. Alors qu'il s'apprêtait à partir en direction de Rome pour plaider une nouvelle fois son cas en désespoir de cause, Otton III le convoque et lui demande de devenir son précepteur, comme l'avait été son père Otton II avant lui.

Gerbert rejoint Otton III à Magdebourg et le suit en Italie. En 998, il est nommé archevêque de Ravenne, la deuxième ville d'Italie. Dès son arrivée sur le siège archiépiscopal, il met en œuvre des réformes, comme il l'avait fait à Bobbio mais avec plus de succès. Il s'attaque en particulier à la simonie. Il n'oublie pas qu'il est également abbé de Bobbio, dont sa nouvelle « principauté » est adjacente : il demande à l'empereur de promulguer un privilège pour l'abbaye, confirmant les biens du monastère.

2.2.6 De 999 à 1003 : Gerbert, deuxième Sylvestre

Le pape Grégoire V meurt subitement en 999. Otton III se hâte à Rome pour contrôler l'élection du nouveau pape. Gerbert le suit de près. Après quelques jours de discussion, l'empereur choisit Gerbert en tant que successeur de saint Pierre. Il prend le nom de Sylvestre, deuxième du nom, un choix qui n'est pas anodin puisqu'il renvoie un premier Sylvestre. Ce dernier avait baptisé Constantin, le premier empereur chrétien : ce nom souligne les liens étroits entre le nouveau pape et son empereur.

Au cours de son pontificat court mais mouvementé, Gerbert établit deux nouvelles Églises, celle d'Hongrie et celle de Pologne, établissant la limite orientale de la chrétienté latine encore d'actualité aujourd'hui. Il régularise également la situation de l'archevêché de Reims, confirmant Arnoul dans sa fonction. Le pape et l'empereur sont chassés de Rome par une révolte en 1001, et Otton III meurt avant de pouvoir retourner dans la Ville éternelle, en 1002, alors qu'il n'avait que 22 ans. Gerbert reprend sa place au Latran et ne survit à son disciple et empereur qu'une seule année. Il est pris d'un malaise le 3 mai 1003 dans la basilique Sainte-Croix-de-Jérusalem et meurt au Latran le 12 mai. Une précision spatio-temporelle tout en contraste par rapport à sa naissance !

2.3 Une production littéraire entre légende et réalité

Les éditions des œuvres de Gerbert sont datées, puisqu'elles remontent au XIX^{e} siècle. La Patrologie latine de J.-P. Migne[22] les comprend, bien que certaines œuvres soient attribuées à d'autres auteurs. C'est de ce corpus que nous avons récupéré les

[22] Jacques-Paul Migne, éd. *Patrologiae Cursus Completus. Series Latina.* 217 t. Paris : Frères Garnier, 1841-1855.

textes pour nos analyses. A. Olleris[23] réunit uniquement les textes de Gerbert, avec quelques œuvres manquantes. N. Bubnov[24] se concentre uniquement sur les œuvres à caractère scientifique. Citons également l'édition récente de la correspondance de Gerbert par P. Riché et J.-P. Callu,[25] qui fait figure d'exception parmi ces éditions d'il y a près de deux siècles. De plus, certaines œuvres individuelles font l'objet d'éditions ponctuelles, mais sans volonté d'éditer un corpus cohérent avec tous les écrits de Gerbert.

M. Mostert a réalisé un tableau synthétique des œuvres de Gerbert selon leur genre,[26] que nous reproduisons ci-dessous avec quelques légères modifications en nous inspirant d'autres classifications[27] :

- Traités ecclésiastiques : *Sermo de informatione episcoporum* (traité sur la formation des évêques), *De corpore et sanguine domini* ;
- Dialectique : *De rationale et uti* ;
- Poésie métrique : quatre épitaphes, un éloge de Boèce, quelques vers isolés ;
- Correspondance : une collection de 220 lettres ;
- Chartes : privilèges et actes de Sylvestre II ;
- Hagiographie : *Vita prior sancti Adalberti Pragensis* ;
- Traités scientifiques :
  - Arithmétique : trois lettres à Constantin portant respectivement sur les règles pour l'abaque, à nouveau sur l'abaque (*De norma rationis abaci*) et sur un passage de l'*Institutio airthmetica* de Boèce ;
  - Géométrie : deux lettres, l'une à Constantin sur la construction des sphères (*De sphaera*) et l'autre à Adelbold sur le calcul de la surface d'un triangle, ainsi que les treize premiers chapitres de ce que Bubnov a publié sous le nom d'*Isagoge geometriae* (introduction à la géométrie) ;
  - Musique : deux lettres à Constantin sur deux passages de l'*Institutio musica* de Boèce ;

[23] Alexandre Olleris. *Oeuvres de Gerbert, pape sous le nom de Sylvestre II*. Clermont-Ferrand & Paris : F. Thibaud & Ch. Dumoulin, 1867.

[24] Nicolaus Bubnov. *Gerberti Opera Mathematica (972-1003)*. Berlin : R. Friedländer & Sohn, 1899.

[25] Gerbert d'Aurillac, *Correspondance*.

[26] Marco Mostert. « Gerbert d'Aurillac, Abbon de Fleury et la culture de l'An Mil : étude comparative de leurs oeuvres et de leur influence ». In : *Gerberto d'Aurillac da abate di Bobbio a papa dell'anno 1000 : atti del congresso internazionale, Bobbio, Auditorium di S. Chiara, 28-30 settembre 2000...* Sous la dir. de Flavio G. Nuvolone. Bobbio : Associazione culturale Amici di Archivum Bobiense, 2001, p. 398-431.

[27] Notamment Bianchini et Senatore, « Gerbert of Aurillac (c. 940–1003) » ; Riché, *Gerbert d'Aurillac* ; Schärlig, *Un Portrait de Gerbert d'Aurillac*.

– Astronomie : une lettre à Adam sur la construction d'un calendrier horologique et les deux versions d'un traité sur l'astrolabe (*De utilitatibus astrolabii*, en 19 ou 21 chapitres).

Parmi ces écrits, certains ont été prouvés comme des pseudépigraphes et d'autres ont été mis en doute. Nous en évoquerons rapidement quelques-uns, avant de nous concentrer sur deux d'entre eux : le *Sermo de informatione episcoporum* et le *De utilitatibus astrolabii* :

- P. Riché considère que le *De corpore et sanguine domini* n'a pas été écrit par Gerbert mais par un de ses disciples, Hériger de Lobbes[28] ;
- Même si son caractère pseudépigraphe est reconnu depuis longtemps, H. Fros a définitivement démontré lors du premier colloque à Bobbio[29] que la *Vita prior sancti Adalberti Pragensis* n'avait pas été écrite par Gerbert ;
- Quelques faux se trouvent dans les chartes pontificales.

Le cas des traités scientifiques est quelque peu plus complexe. A. Schärlig, notamment, soutient que seule l'authenticité des lettres est indubitable et que le corpus scientifique associée au nom de Gerbert a été davantage suscité que rédigé par lui.[30] L'auteur suit en cela la position critique de G. Beaujouan,[31] qui met en doute la plupart des traités de Gerbert en soulignant la fragilité des hypothèses sur lesquelles reposent les attributions à notre auteur. Il le considère donc comme un « patron » des sciences dont le nom aurait été apposé sur des œuvres comme un gage de valeur pédagogique et scientifique.

Deux traités en particulier sont au centre des discussions : le traité sur l'astrolabe, que nous traiterons en profondeur, et l'introduction à la géométrie (*Isagoge geometriae*). Nous pouvons souligner que la position de N. Bubnov, qui divise cette introduction en deux parties et attribue les treize premiers chapitres à Gerbert et le reste de l'ouvrage à un auteur inconnu, n'a pas été abandonnée dans la littérature scientifique.[32]

[28] Riché, *Gerbert d'Aurillac*, p. 139.

[29] Henryk Fros. « Les Vies de St-Adalbert - Vojtech, attribuées à Sylvestre II ». In : *"Gerberto : scienza, storia e mito" : atti del Gerberti Symposium (Bobbio 25-27 luglio 1983)*. Sous la dir. de Michele Tosi. Bobbio : Archivi storici bobiensi, 1985, p. 567-576.

[30] Schärlig, *Un Portrait de Gerbert d'Aurillac*, p. 29-36.

[31] Beaujouan, « Gerberto ».

[32] Marco Mostert. « Les Traditions Manuscrites Des Œuvres de Gerbert ». In : *Gerbert l'Européen. Actes Du Colloque d'Aurillac, 4–7 Juin 1996*. Sous la dir. de Nicole Charbonnel et Jean-Eric Iung. Aurillac : Société des Lettres, Sciences et Arts "La Haute Auvergne", 1997, p. 316.

2.3.1 *Sermo de informatione episcoporum*

Puisque Gerbert s'était attaqué vigoureusement à la simonie lorsqu'il était archevêque de Ravenne, le *Sermo* lui a été attribué étant donné qu'il aborde les mêmes thématiques, alors que son réel auteur est Adhémar de Chabannes, qui l'aurait attribué à Gerbert pour lui conférer plus de légitimité.[33]

Le texte du *Sermo* provient d'un seul manuscrit, daté du XI^{e} siècle et trouvé à Limoges par Jean Mabillon en 1690. Depuis lors, il est intégré dans le corpus gerbertien, même si A. Olleris exprime déjà quelques doutes sur son authenticité et l'attribue à un faussaire du X^{e} siècle.[34] Ce manuscrit porte directement le nom de Gerbert dans son *incipit* (*Sermo Giberti philosophi, papae urbis Romae, qui cognominatus est Silvester, De informatione episcoporum*). Il s'agit d'une version modifiée d'un traité attribué à Ambroise de Milan (vraisemblablement à tort[35]), intitulé *Sermo de Dignitate Sacerdotali*. Dans son article sur ce traité, G. Williams envisage que le traité trouvé par J. Mabillon puisse s'agir d'une réfection par Gerbert vers la fin de sa vie.[36]

F. Nuvolone, après avoir étudié la question de la transmission de l'œuvre-« mère » qu'est le *Sermo pastoralis* attribué à Ambroise de Milan, a effectué une étude très détaillée et complète sur le texte du *Sermo de informatione episcoporum*.[37] Il constate que cette version est très remaniée par rapport au texte d'origine (ajouts, omissions, variations), et repère un changement de style qu'il qualifie d'évident au niveau de la rapidité de l'expression, de l'explicitation de certains passages, de l'abondance de discours indirects, etc.[38] Pour vérifier si une attribution à Adhémar tient la route, F. Nuvolone détaille la méthodologie qu'il va mettre en pratique. Il commence par sélectionner les lemmes qui ont été introduits dans cette nouvelle version et sont absents du texte d'origine. Il y en a moins de 300 et il n'y a presque pas d'expressions complètes ou d'idées nouvelles évoquées. Ce n'est donc pas concluant. Il est intéressant de préciser que l'auteur enlève les mots-outils de ses considérations, car ce sont des « éléments à peine importants » (« *componenti difficilmente significative* »).[39] Nous verrons qu'en stylométrie, ces mots-outils sont utilisés depuis les années 1960 pour identifier des auteurs. F. Nuvolone se base alors sur tout un ensemble de critères (en

[33] Riché, *Gerbert d'Aurillac*, p. 201.

[34] Flavio G. Nuvolone. « Il *Sermo pastoralis* Pseudoambrosiano e il *Sermo Giberti philosophi papae urbis Romae qui cognominatus est Silvester de informatione Episcoporum*. Riflessioni ». In : *"Gerberto : scienza, storia e mito" : atti del Gerberti Symposium (Bobbio 25-27 luglio 1983)*. Sous la dir. de Michele Tosi. Bobbio : Archivi storici bobiensi, 1985, p. 388.

[35] George Huntston Williams. « The Golden Priesthood and the Leaden State. A Note on the Influence of a Work Sometimes Ascribed to St. Ambrose : The Sermo de Dignitate Sacerdotali ». In : *Harvard Theological Review* 50.1 (1957), p. 37-64.

[36] Williams, « The Golden Priesthood and the Leaden State. A Note on the Influence of a Work Sometimes Ascribed to St. Ambrose », p. 39-40.

[37] Nuvolone, « Gerberto », p. 430-489.

[38] Nuvolone, « Gerberto », p. 434-441.

[39] Nuvolone, « Gerberto », p. 450.

particulier le vocabulaire employé, mais également les thèmes évoqués, les citations utilisées…) pour établir une similarité avec les écrits d'Adhémar.[40] Il conclut que le *Sermo* a été préparé ou supervisé par Adhémar.[41] Il admet toutefois qu'il ne dispose pas d'une « certitude mathématique de ce fait » (« *Non abbiamo la certezza matematica del fatto.* »[42]), en raison du manque d'études sur Adhémar lui-même. Nous avons trouvé cette formulation surprenante dans la bouche d'un philologue qui travaille qualitativement, et avons trouvé bon de la mettre en évidence.

2.3.2 *De utilitatibus astrolabii*

La littérature relative à ce traité d'astronomie est plus vaste et complexe. La question de la paternité de cet ouvrage entraîne non seulement les doutes sur les écrits scientifiques de Gerbert, que nous avons déjà évoqués ci-dessus, mais également les relations intellectuelles entre le monde musulman et le monde chrétien.

Ce traité est attribué à Hermann Contract dans la Patrologie latine. Dans cette dernière, le titre *De astrolabio* est donné au rassemblement de deux livres, le premier étant le traité qui nous intéresse et le second un rassemblement de conseils qui n'a rien à voir avec le contenu du premier.[43] Cette confusion dans la division des textes et dans l'auteur est certainement due au fait qu'Hermann Contract est à l'origine d'une compilation de trois traités : le *De mensura astrolabii* de lui-même et les deux textes contenus dans le *De astrolabio* de la Patrologie latine. Des doutes sur l'attribution de *De utilitatibus astrolabii* à Hermann Contract ont été formulés très tôt : N. Bubnov le place dans les *opera dubia* de Gerbert, tandis que des ouvrages consacrés à Hermann lui en retire la paternité.[44]

Des désaccords dans la littérature sont évidents dès le premier colloque organisé à Bobbio en 1983. Trois érudits y évoquent le *De utilitatibus astrolabii*. E. Poulle et G. Beaujouan, déjà évoqué, remettent tous deux en question une attribution de ce traité à Gerbert et le conçoivent comme un facilitateur ou un rassembleur de documentation pour la génération suivante,[45] tandis que U. Lindgren part du principe, dans son exposé, que ce traité est bien dans le main de Gerbert.[46] G. Beaujouan évoque des

[40] Nuvolone, « Gerberto », 454-sqq.

[41] Nuvolone, « Gerberto », p. 581-582.

[42] Nuvolone, « Gerberto », p. 487.

[43] Catherine Jacquemard, Olivier Desbordes et Alain Hairie. « Du quadrant vetustior à l'horologium viatorum d'Hermann de Reichenau : étude du manuscrit Vaticano, BAV Ott. lat. 1631, f. 16-17v ». In : *Kentron. Revue pluridisciplinaire du monde antique* 23 (23 2007), p. 89-92.

[44] Joseph Drecker. « Hermannus Contractus Über Das Astrolab ». In : *Isis* 16.2 (1931), p. 201.

[45] Emmanuel Poulle. « L'Astronomie de Gerbert ». In : *"Gerberto : scienza, storia e mito" : atti del Gerberti Symposium (Bobbio 25-27 luglio 1983)*. Sous la dir. de Michele Tosi. Bobbio : Archivi storici bobiensi, 1985, p. 597-617 ; Beaujouan, « Gerberto ».

[46] Uta Lindgren. « Ptolémée chez Gerbert d'Aurillac ». In : *"Gerberto : scienza, storia e mito" : atti del Gerberti Symposium (Bobbio 25-27 luglio 1983)*. Sous la dir. de Michele Tosi. Bobbio : Archivi storici bobiensi, 1985, p. 619-644.

travaux antérieurs qui soit considèrent la Marche d'Espagne comme le foyer de l'influence arabe sur la science occidentale,[47] soit la Lotharingie.[48] Dans les deux cas, le manuscrit 225 de Ripoll est au cœur de la question, et il y restera.

À la suite de ce colloque, les travaux s'enchaînent. W. Bergman attribue solidement le traité à Gerbert, sur base d'un ensemble de critères (les textes contemporains sur l'astrolabe, les connaissance astronomiques de Gerbert, le contexte codicologique...).[49] Il présuppose qu'une traduction inconnue de l'arabe a servi de source à Gerbert. A. Borst, quant à lui, raffermit la position de la Marche d'Espagne comme premier point de contact avec la science arabe.[50] Ce point de vue est suivi par P. Kunitzsch, qui identifie le manuscrit 225 de Ripoll comme le point de départ de la littérature latine sur l'astrolabe.[51] Il doute néanmoins d'une attribution du traité à Gerbert et privilégie plutôt l'un de ses disciples, en raison du manque de vocabulaire arabe qu'il emploie dans sa correspondance (même dans les lettres de nature scientifique). U. Lindgren, lors d'une intervention au colloque en 2000 à Bobbio, confirme le présupposé dont elle était partie en 1983 et défend une attribution du traité à Gerbert, en mettant en évidence son intérêt pour l'astronomie et le fait qu'il fait venir un traité sur ce sujet depuis la Marche d'Espagne en 984.[52] Dans le colloque de 2004, M. Zuccato souligne que les sphères célestes utilisées par Gerbert dans son enseignement sont une preuve de l'influence arabe à Reims à cette époque. M. Otisk considère lui aussi que le traité a vu le jour dans l'entourage de Gerbert, mais que le véritable auteur est l'un de ses disciples.[53]

Ce même colloque de 2004 voit E. Poulle réaffirmer ses doutes sur la paternité de ce traité.[54] Il met en relation les découvertes convergentes de deux chercheurs qui pourtant ne se connaissaient pas. Le premier, C. Burnett, argumente que Reims

[47]Josep Millàs Vallicrosa. *Assaig d'Història de Les Idees Fisiques i Matemàtiques a La Catalunya Medieval*. Barcelone : Institucio Patxot, 1931.

[48]André Van de Vyver. « Les Premières Traductions Latines (X^{e}-XI^{e} s.) de Traites Arabes Sur l'astrolabe ». In : *1er Congrès International de Géographie Historique*. Sous la dir. de Fritz Quicke. T. 2. 1931, p. 266-290 ; A. van de Vyver. « Les plus Anciennes Traductions Latines Médiévales (X^{e}-XI^{e} Siècles) de Traités d'astronomie et d'astrologie ». In : *Osiris* 1 (1936), p. 658-691.

[49]Werner Bergmann. *Innovationen Im Quadrivium Des 10. Und 11. Jahrhunderts : Studien Zur Einführung von Astrolab Und Abakus Im Lateinischen Mittelalter*. Stuttgart : Franz Steiner Verlag, 1985.

[50]Arno Borst. *Astrolab Und Klosterreform an Der Jahrtausendwende : Vorgetragen Am 11. Februar 1989*. Heidelberg : Winter, 1989.

[51]Paul Kunitzsch. « Les Relations Scientifiques Entre Occident et Monde Arabe ». In : *Gerbert l'Européen. Actes Du Colloque d'Aurillac, 4–7 Juin 1996*. Sous la dir. de Nicole Charbonnel et Jean-Eric Iung. Aurillac : Société des Lettres, Sciences et Arts "La Haute Auvergne", 1997, p. 193-203.

[52]Uta Lindgren. « Gerbert et les arts libéraux ». In : *Gerberto d'Aurillac da abate di Bobbio a papa dell'anno 1000 : atti del congresso internazionale, Bobbio, Auditorium di S. Chiara, 28-30 settembre 2000...* Sous la dir. de Flavio G. Nuvolone. Bobbio : Associazione culturale Amici di Archivum Bobiense, 2001, p. 107-125.

[53]Otisk, « Gerbert of Aurillac (Pope Sylvester II) as a Clockmaker ».

[54]Emmanuel Poulle. « Gerbert Homme de Science ». In : *Gerberto d'Aurillac - Silvestro II, Linee per Una Sintesi : Atti Del Convegno Internazionale, Bobbio, Auditorium Di S. Chiara, 11 Settembre 2004, Sotto La Presidenza Del Prof. Pierre Racine...* Sous la dir. de Flavio G. Nuvolone. Bobbio : Associazione culturale Amici di Archivum bobiense, 2005, p. 95-123.

n'était pas le centre des connaissances astronomiques à la fin du X^{e} siècle, mais que c'était la région de Chartes et de Fleury (où se trouve l'abbaye d'Abbon, le grand rival de Gerbert, aussi intéressé par l'astronomie). C'est à cet endroit que les premiers traités latins sur l'astrolabe auraient été réalisés. La deuxième, C. Jacquemard, n'avait pas encore publié au moment du colloque, mais rejoignait ces observations. En 2006, elle a publié un article dans lequel elle invalide un des arguments principaux des tenants de la paternité de Gerbert, à savoir que ce dernier mentionne l'alidade de l'astrolabe dans une de ses lettres. Or, la chercheuse montre que le terme employé, *radius geometricus*, ne convient pas réellement pour désigner cette partie de l'outil.[55] En résumé, selon E. Poulle, les découvertes de ces deux chercheurs « ont mis Gerbert complètement hors de course » en ce qui concerne l'attribution du traité.[56]

Il nous cherche un chercheur à évoquer, M. Zuccato, qui avait déjà proposé une vision différente de l'arrivée de la connaissance astronomique arabe en Occident latin dans son article de 2005,[57] en parlant d'un « canal juif ». Cet article était peu convaincant par le nombre d'hypothèses que cette interprétation demandait de formuler. Dix ans plus tard, le même chercheur a publié un nouvel article dans lequel il revient sur son hypothèse en la détaillant.[58] Pour trouver la source de l'influence arabe sur la connaissance de Gerbert, il fallait envisager des contacts après 978, date de la création d'une école d'astronomie en al-Andalus, centrée autour de la personne de Maslama. Or, il est évident que Gerbert a été imprégné d'un savoir novateur lors de son séjour en Marche d'Espagne, notamment par son utilisation de sphères célestes dans le cadre de son enseignement à Reims. L'auteur propose l'existence d'une « astronomie pratique » en al-Andalus pendant la deuxième moitié du X^{e} siècle, non pas basée sur des écrits mais sur des gestes et des instructions orales. Dans le cadre de relations étroites entre al-Andalus et la Marche d'Espagne, ce qui était le cas à cette époque où cette dernière cherchait à prendre ses distances par rapport au royaume franc, l'auteur propose que des échanges intellectuels et culturels se sont produits, parmi lesquels l'arrivée de cette connaissance astronomique. Gerbert se serait imprégné de cette connaissance entre 967 et 970 et l'aurait appliqué dans l'école-cathédrale de Reims en l'identifiant au *quadrivium* latin, expliquant le côté novateur de la didactique gerbertienne. Au fur et à mesure, ces connaissances auraient été mises par écrit, donnant naissance au premier corpus de textes sur l'astrolabe en Occident latin. Cette interprétation est beaucoup plus satisfaisante que la première que M. Zuccato avait proposée, et explique l'origine des connaissances astronomiques de Gerbert. Toute-

[55] Catherine Jacquemard. « Erectio, inclinatio / erectus, inclinatus : de Vitruve à Gerbert d'Aurillac (à propos de l'expression de la distance angulaire fin Xe - début XIe siècle) ». In : *Collection de l'Institut des Sciences et Techniques de l'Antiquité* 993.1 (2006), p. 157-162.

[56] Poulle, « Gerberto d'Aurillac - Silvestro II, Linee per Una Sintesi », p. 121.

[57] Zuccato, « Gerbert of Aurillac and a Tenth-Century Jewish Channel for the Transmission of Arabic Science to the West ».

[58] Zuccato, « Arabic Singing Girls, the Pope, and the Astrolabe ».

fois, elle ne prouve pas que Gerbert ait bien écrit le traité sur l'astrolabe, puisqu'elle est tout à fait compatible avec l'interprétation d'un Gerbert transmetteur de connaissance.

2.4 Le concept d'auteur au Moyen Âge

Il est facile de se reposer sur notre compréhension contemporaine du concept d'auteur et de l'appliquer de manière anachronique sur le Moyen Âge. La déconstruction de l'auteur, associée au mouvement postmoderniste et à des noms comme Michel Foucault ou Roland Barthes, a entraîné une réflexion au sein de la médiévistique.

2.4.1 Critique du concept

Selon M. Foucault, l'auteur ne doit pas être envisagé comme un individu mais comme une fonction. L'attention ne doit plus se porter sur les individus, mais sur les relations qui existent entre les différents éléments d'un texte, parmi lesquels se trouve la fonction-auteur dont l'objectif est d'apporter du sens.[59] Cette fonction n'apparaît pas, selon lui, avant la fin du Moyen Âge. La période médiévale, quant à elle, n'apportait que peu d'importance aux auteurs de textes littéraires, tandis que les noms attachés aux traités scientifiques donnaient toute leur valeur à ces écrits.[60] Nous avons bien constaté ce processus de validation dans la bibliographie de Gerbert d'Aurillac, bien que nous élargirions volontiers la catégorie de « textes scientifiques » pour y inclure les écrits de nature théologique (voir le cas du *Sermo de informatione episcoporum*).

Les médiévistes se réapproprient cette critique pour poser les bases d'une « nouvelle philologie », associée aux noms de B. Cerquiglini et P. Zumthor. Cette nouvelle philologie, résumée par J. Deploige et J. De Gussem,[61] est fondée toute entière sur les notions de mouvance et de variance : les textes médiévaux sont en état de perpétuelle réécriture, et c'est une de leurs richesses et non de leurs défauts. Le but n'est plus de rechercher la version du texte la plus proche de l'œuvre originelle, comme c'est le cas

[59] Atle Kittang. « Authors, Authorship, and Work : A Brief Theoritical Survey ». In : *Modes of Authorship in the Middle Ages*. Sous la dir. de Slavica Ranković. 22. Toronto : Pontifical Institute of Mediaeval Studies, 2012, p. 17-18, 28.

[60] Michel Foucault. « Qu'est-ce qu'un auteur ? » In : *Bulletin de la Société française de philosophie* 63/3 (1969), p. 799-800.

[61] Jeroen Deploige et Jeroen De Gussem. « Medieval Authorship and Canonicity in the Digital Age – an Introduction ». In : *Interfaces : A Journal of Medieval European Literatures* 8 (2021), p. 113-114 ; voir également la partie dédiée à cette question dans la thèse de Jeroen De Gussem. « Collaborative Authorship in Twelfth-Century Latin Literature : A Stylometric Approach to Gender, Synergy and Authority ». Thèse de doct. Gent & Antwerpen : Universiteit Gent & Universiteit Antwerpen, 2019, p. 3-7.

en philologie lachmanienne. Cette philologie porte également le nom de « philologie matérielle », pour souligner l'importance qu'elle octroie aux manuscrits en eux-mêmes et pour eux-mêmes : leur environnement social, le cadre de leur rédaction, leur aspect matériel. Dans ces conditions, la notion de paternité est plus flexible que dans son acception actuelle. Dans le cadre de ce nouveau paradigme, B. Cerquiglini considère que « [l]'auteur n'est pas une idée médiévale » et que parler d'un auteur médiéval est « un anachronisme fonctionnel ».[62]

2.4.2 Qu'est-ce qu'un auteur (médiéval) ?

Comment rebondir après une formule si définitive ? La plupart des ouvrages que nous avons pu consulter ont été consacrés aux textes littéraires et/ou à la fin de la période médiévale.[63] Rares sont ceux qui mentionnent les autres genres et le début du Moyen Âge.

Dans l'ouverture d'un colloque dont il a édité les actes, M. Zimmermann critique l'anachronisme commis lorsque l'on recherche un auteur au sens romantique dans la période médiévale :

> « [...] influencés par le souci prioritaire de l'attribution des œuvres et des droits de l'auteur, les historiens ont longtemps jugé la création médiévale à l'aune de critères rétrospectifs, réducteurs et peu pertinents ; constatant les phénomènes récurrents de continuité, de glose, d'emprunt et d'anonymat, ils dénoncent plagiaires et faussaires, stigmatisent l'absence d'originalité, se désolent de devoir attribuer telle œuvre à une école ou à un atelier, de devoir relever le travail d'un scriptorium ou d'une chancellerie. Tout se passe à leurs yeux comme si l'individu/auteur, socialement identifiable et professionnellement producteur unique d'une création originale, était une réalité permanente et une valeur intemporelle. Si le Moyen Age n'a pas produit d'auteurs, c'est qu'il n'a pas accédé à la création. »[64]

L'absence d'auteur et d'œuvre au sens actuel du terme ne signifie pas que les médiévaux n'écrivaient pas, ne composaient pas, ne créaient pas. Dans le même colloque, P. Bourgain propose, pour aborder au plus proche ce que pourrait être un au-

[62] Bernard Cerquiglini. *Éloge de la variante. Histoire critique de la philologie*. Paris : Editions du Seuil, 1989, p. 25-26.

[63] Paul Zumthor. *La Lettre et La Voix. De La "Littérature" Médiévale*. Paris : Editions du Seuil, 1987 ; Virginie Greene, éd. *The Medieval Author in Medieval French Literature*. New York : Springer, 2006 ; A. J. Minnis. *Medieval Theory of Authorship : Scholastic Literary Attitudes in the Later Middle Ages*. 2nd ed. with a new preface by the author. Philadelphia : University of Pennsylvania Press, 2010 ; Erik Kwakkel et Stephen Partridge. *Author, Reader, Book : Medieval Authorship in Theory and Practice*. Toronto : University of Toronto press, 2011.

[64] Michel Zimmermann, éd. *Auctor et auctoritas. Invention et conformisme dans l'écriture médiévale. Actes du colloque tenu à l'Université de Versailles-Saint-Quentin-en-Yvelines (14-16 juin 1999)*. Mémoires et documents de l'Ecole des Chartes 59. Paris : Ecole des Chartes, 2001, p. 8.

teur au sens médiéval du terme, d'étudier le vocabulaire utilisé par désigner les réalités en lien avec la notion d'auteur. Elle repère qu'à chaque étape de la création correspondent des verbes différents : la conceptualisation de l'œuvre renvoie à des verbes tels que *componere* et *tractare* (dans le cas d'un commentateur), tandis que l'étape de rédaction à proprement parler invite l'attendu *scribere*, mais également des verbes liés à l'oral pour refléter la place occupée par les scribes et les secrétaires (*dictare*, *dicere*...). Des métaphores liées à des activités artisanales existent également : *cudere* qui fait écho à l'activité du forgeron, *texere* à celle du tisserand, *pangere* à celle du constructeur.[65] Remarquons que ce type de métaphore existe encore actuellement, comme le démontre clairement l'anglais *wordsmith*.

Cette division du concept d'auteur fait écho à la compréhension actuelle de l'auteur, sous la forme d'un ensemble d'activités liées, qui ne sont pas toujours réalisées par une seule et même personne. La fonction-exécuteur représente la réalisation de l'écrit à proprement parler. La fonction-précurseur désigne une influence dont la contribution à l'œuvre créée est substantielle, ce qui est le mode de création usuel au Moyen Âge. La fonction-déclarateur est l'acte de placer une œuvre sous l'autorité d'un nom reconnu et estimé.[66]

M. Zink, dans un article dont le nom fait écho au colloque précédemment mentionné, s'intéresse à la réalité étymologique représentée par le mot *auctor*.[67] À l'époque médiévale, l'*auctor* est celui qui possède l'*auctoritas*, c'est-à-dire l'autorité mais c'est aussi la qualité d'un auteur du passé, sur lequel on peut s'appuyer. Si l'on ne considère que son sens étymologique, le mot *auctor* est bien approprié pour désigner la pratique de l'auteur médiéval : il « augmente », il complète une tradition antérieure « en glosant, en l'assimilant à son propre texte, en dialoguant avec elle ».[68] L'auteur reprend l'expression de Bernard de Clairvaux, qui selon lui reflète parfaitement la conception médiévale de ce que sont des auteurs : « des nains juchés sur les épaules des géants ».

Au-delà des explorations étymologiques, si passionnantes soient-elles, certains chercheurs ont tenté de bâtir de nouveaux modèles pour mieux comprendre l'auteur médiévale. L'ouvrage collectif dirigé par S. Ranković[69] en contient plusieurs, dont celui de l'éditrice, mais nous n'évoquerons que celui proposé par M. Drout.[70] Ce der-

[65]Pascale Bourgain. « Les verbes en rapport avec le concept d'auteur ». In : *Auctor et auctoritas. Invention et conformisme dans l'écriture médiévale. Actes du colloque tenu à l'Université de Versailles-Saint-Quentin-en-Yvelines (14-16 juin 1999)*. Sous la dir. de Michel Zimmermann. Paris : Ecole des Chartes, 2001, p. 361-374.

[66]Harold Love. *Attributing Authorship : An Introduction*. Cambridge : Cambridge University Press, 2002, p. 39-50.

[67]Michel Zink. « Auteur et autorité au Moyen Âge ». In : *De l'autorité. Colloque annuel du Collège de France*. Paris : Odile Jacob, 2008, p. 143-158.

[68]Zink, « Auteur et autorité au Moyen Âge », p. 153.

[69]Slavica Ranković, éd. *Modes of Authorship in the Middle Ages*. 22. Toronto : Pontifical Institute of Mediaeval Studies, 2012.

[70]Michael D.C. Drout. « "I Am Large, I Contain Multitudes" : The Medieval Author in Memetic Terms ». In : *Modes of Authorship in the Middle Ages*. Sous la dir. de Slavica Ranković. 22. Toronto :

nier s'inspire de la notion de *meme* introduite par R. Dawkins, qui lui semble parfaite pour expliquer l'auteur médiéval qui, comme nous l'avons vu, fonctionne sur base de citations et d'augmentations individuelles qui se superposent pour former une production collective. Les *memes* sont définis comme des petites entités qui peuvent se multiplier et qui, ensemble, créent des traditions. L'auteur, quant à lui, est l'endroit où ces entités se combinent et se reproduisent : l'auteur « *is the crucible (mental and physical) in which all the various memes combine to create new memes* ».[71] La beauté de cette théorie est qu'elle reconnaît la place centrale de la tradition dans la pensée médiévale, tout en n'invisibilisant pas l'individu qui est le moteur qui permet à cette tradition de se développer.

Pontifical Institute of Mediaeval Studies, 2012, p. 30-51.

[71] Drout, « "I Am Large, I Contain Multitudes" : The Medieval Author in Memetic Terms », p. 42.

Chapitre 3

La stylométrie et les études de paternité

Cette section propose un développement sur les études de paternité, dont la stylométrie. Nous commencerons par définir cette discipline dans le cadre des études de paternité. Ensuite, nous établirons un historique de la discipline, en évoquant des articles qui sont des témoins de l'évolution des techniques et des méthodologies. Nous exclurons volontairement tout article manipulant du latin de cette partie, puisque nous en consacrerons une aux liens qui unissent cette langue et la stylométrie. Enfin, puisqu'il s'agit de la méthodologie que nous avons décidé de mettre en pratique, nous ferons un point exhaustif sur les réseaux utilisés en stylométrie.

3.1 Les études de paternité

La seule monographie que nous avons trouvée sur le sujet des études de paternité est l'ouvrage de H. Love.[1] Elle a le mérite de contextualiser la stylométrie, aussi appelée études de paternité non-traditionnelles, dans un cadre plus large. Pour reprendre les mots de l'auteur, son objectif est :

> « to mediate from a literary perspective between the impressive computer-based work on attribution studies which has been done over the last four decades and a much older tradition of such studies, which, considered as an organised scholarly enterprise, reaches back as far as the great library of Alexandria and embraces the formation of the Jewish and Christian biblical canons. »[2]

P. Juola définit les études de paternité comme suit : « any attempt to infer the characteristics of the creator of a piece of linguistic data. » Cette définition, volontaire-

[1] Love, *Attributing Authorship*.
[2] Love, *Attributing Authorship*, p. 1.

ment vague, englobe de nombreuses applications, qui sont de trois ordres : déterminer l'auteur d'un texte donné parmi un ensemble d'auteurs-candidats (attribution de paternité), déterminer si un auteur donné a écrit un texte ou pas (vérification de paternité), extraire le plus de caractéristiques d'un texte donné (« profiling »).[3]

3.1.1 Historique

Les premières traces d'études d'attribution remontent à l'érudition alexandrine, qui a cherché à différencier les œuvres authentiquement d'Homère des autres. Un des cas historiques les plus célèbres est la publication par l'humaniste Lorenzo Valla, au XV^{e} siècle, de sa *Declamatio de falsa et ementita donatione Constantini* : il déclare que la déclaration de Constantin est un faux, en se basant sur plusieurs critères tels que des anachronismes dans le vocabulaire latin. Les répercussions historiques ont été bien réelles, puisque ce document permettait au pape de justifier sa possession des États pontificaux. À la même époque, Érasme lui aussi se prête à des études de paternité, avec des prémisses de formalisation : il repère notamment que Quintilien a l'air d'utiliser plus souvent *interim* alors que les autres auteurs préfèrent *interdum*. Il ne réalise toutefois jamais d'étude quantitative systématique.[4]

Pour déterminer l'auteur d'une œuvre, il est possible de chercher des preuves externes ou des preuves internes. Les preuves externes sont toute référence au monde extérieur, ainsi par exemple lorsque le nom de l'auteur est donné en début d'ouvrage. Ce sont ces preuves externes qui nous permettent de donner un nom aux auteurs, plutôt qu'un pseudonyme. Les preuves internes, souvent moins facilement accessibles que les preuves externes, se situent au niveau de chaque mot du texte : non seulement le style, qui sera le point de départ de toute analyse stylométrique, mais les idées évoquées, les allusions, les parallèles entre écrits…[5] Lorsque F. Nuvolone tente de prouver une paternité d'Adhémar de Chabannes pour le *Sermo de informatione episcoporum*, il emploie toutes les possibilités que lui offre une analyse interne du texte.

Le style, nous l'avons vu, est une preuve interne. Lié étymologiquement à *stilus*, le moyen d'écrire puis par métonymie la façon d'écrire, ce concept n'est pour autant pas aisé à définir en prenant en compte toutes les acceptions qui existent pour ce terme. Trois chercheurs, après un état de l'art sur la question dans trois traditions historiographiques différentes, propose une définition qui n'exclut aucun texte, qu'il

[3] Patrick Juola. « Authorship Attribution ». In : *Foundations and Trends in Information Retrieval* 1.3 (2006), p. 328 ; Christian Delcourt. « Stylometry ». In : *Revue belge de philologie et d'histoire* 80.3 (2002), p. 981-983.

[4] Love, *Attributing Authorship*, p. 14-20 ; Hugh Craig. « Stylistic Analysis and Authorship Studies ». In : *A Companion to Digital Humanities*. Sous la dir. de Susan Schreibman, Ray Siemens et John Unsworth. Blackwell Publishing. Malden, Oxford & Victoria : Blackwell Publishing Oxford, UK, 2004, p. 282.

[5] Love, *Attributing Authorship*, p. 51-97 ; Craig, « Stylistic Analysis and Authorship Studies », p. 283.

soit littéraire ou non : « Style is a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively. ».[6] L'existence du style ne fait pas de doute, la meilleure méthode pour le représenter fait toujours débat. Des méthodes de quantification précèdent les techniques s'appuyant sur la puissance de l'ordinateur, mais cette dernière permet d'en améliorer substantiellement la facilité et l'efficacité, ainsi que d'en diminuer la subjectivité basée sur l'« intuition » des érudits. H. Love souligne que le passage à la stylométrie, définie comme la mesure quantitative du style, se passe presque imperceptiblement, au début du XX^{e} siècle.[7] De même, J. De Gussem souligne que la stylométrie ne doit pas être conceptualisée comme en rupture avec les études de paternité qualitatives, mais comme leur prolongement. L'application de méthodes statistiques est une révolution technique, mais n'est pas une révolution littéraire.[8] Elle s'ancre en effet dans une tradition qui remonte plus loin que le XIX^{e} siècle, comme nous avons pu le constater.

Si la stylométrie a du mal à convaincre les chercheurs des sciences humaines, c'est que le système de preuve y est discursif : il faut expliquer nos résultats et convaincre qu'ils sont fiables. Même si à première vue la communication entre la stylométrie et les sciences humaines semble compliquée, il est utile de souligner que le processus stylométrique lui-même est mathématique et objectif, mais que l'interprétation est subjective. Par conséquent, il est nécessaire qu'elle se soumette au système de preuve discursif commun dans les sciences humaines et ne pas se cacher derrière l'aspect « scientifique » que lui procurent les nombres.[9] Pour reprendre les mots de C. Delcourt, qui certes s'exprimait à propos du double péril qui menace toute analyse stylométrique (une faute philologique ou une faute mathématique), « stylometry is a two-headed methodology ».[10]

3.1.2 Stylométrie

L'hypothèse première sur laquelle repose toute étude de paternité, traditionnelle ou non, est que chaque auteur est caractérisé par un style qui lui est personnel et qui est vérifiable.[11] Ce style ne peut pas être imité, puisqu'il relève en partie de l'inconscient.[12] C'est cette « empreinte d'auteur », aussi appelée « stylome »,[13] que la stylomé-

[6] J. Berenike Herrmann, Karina van Dalen-Oskam et Christof Schöch. « Revisiting Style, a Key Concept in Literary Studies ». In : *Journal of Literary Theory* 9.1 (2015), p. 44.

[7] Love, *Attributing Authorship*, p. 98-118, 132-162.

[8] De Gussem, « Collaborative Authorship in Twelfth-Century Latin Literature », p. 80-85.

[9] Love, *Attributing Authorship*, p. 209-227.

[10] Delcourt, « Stylometry », p. 990.

[11] Love, *Attributing Authorship*, p. 12.

[12] Craig, « Stylistic Analysis and Authorship Studies », p. 284 ; David I. Holmes. « The Evolution of Stylometry in Humanities Scholarship ». In : *Literary and Linguistic Computing* 13.3 (1998), p. 111.

[13] Concept introduit par Hans Van Halteren et al. « New Machine Learning Methods Demonstrate the Existence of a Human Stylome ». In : *Journal of Quantitative Linguistics* 12.1 (2005), p. 65-77.

trie cherche à quantifier. Il est évident qu'un texte ne peut se réduire à une série de nombres et garde une valeur esthétique propre, et ce n'est pas ce que la stylométrie prétend réaliser : elle postule que la simplification induite par une réduction d'un texte en une série de mesures n'empêche pas la sauvegarde d'informations linguistiques pertinentes.[14]

De nombreuses mesures statistiques ont été proposées pour quantifier ce stylome : la longueur moyenne des phrases ou des mots, la moyenne des syllabes par mot, la distribution des natures, la richesse du vocabulaire... Aucune ne s'est révélée particulièrement fiable, jusqu'à la proposition d'utiliser les mots-outils (*function words*) : des mots très fréquents dans une langue tout en représentant un tout petit pourcentage du nombre total de mots dans cette langue, dénués de valeur sémantique, indépendants du genre ou du sujet, et dont l'utilisation est inconsciente par les auteurs.[15] D'autres caractéristiques ont été envisagées, sans obtenir de réel consensus sur la meilleure d'entre elles. Citons, en plus du choix et des propriétés du vocabulaire déjà évoqués, des propriétés syntaxiques telles que les natures des mots, la ponctuation... ou des propriétés qui combinent syntaxe et lexique comme les n-grammes de mots.[16] P. Juola insiste que l'objectif des chercheurs ne devrait pas être de trouver de nouvelles caractéristiques, pas uniquement en tout cas, mais de trouver la meilleure manière de les combiner pour obtenir une meilleure modélisation du stylome.[17]

Une fois l'extraction des caractéristiques linguistiques effectuées, nous pouvons passer à la phase d'attribution. Avec l'arrivée de l'intelligence artificielle, la tâche d'attribution est équivalente à un problème de reconnaissance de motif. Les méthodes déjà envisagées sont très nombreuses, mais se divisent en deux espèces. D'une part, les analyses non supervisées, dont l'objectif est généralement de visualiser les données afin de mieux les explorer. Se situe dans cette catégorie l'analyse en composantes principales (ACP), qui permet de réduire un objet hautement dimensionnel en deux composantes qui représentent le plus justement possible les relations entre les variables de l'objet d'origine. D'autre part, l'analyse supervisée, qui permet d'aboutir à une classification (et par conséquent une réponse à une question de paternité). Se trouvent entre autres dans cette catégorie l'analyse discriminante linéaire, proche de l'ACP mais sur des données déjà étiquetées ; les machines à vecteur de support (SVM), qui permettent d'inférer des hyperplans pour couper un espace vectoriel de la façon

[14] Craig, « Stylistic Analysis and Authorship Studies », p. 279-280 ; Delcourt, « Stylometry », p. 987-988.

[15] Mike Kestemont. « Function Words in Authorship Attribution. From Black Magic to Theory ? » In : *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*. Gothenburg, Sweden : Association for Computational Linguistics, 2014, p. 60-62.

[16] Juola, « Authorship Attribution », p. 262-271.

[17] Juola, « Authorship Attribution », p. 318 ; David L. Hoover. « Quantitative Analysis and Literary Studies ». In : *A Companion to Digital Literary Studies*. Sous la dir. de Susan Schreibman et Ray Siemens. Malden, Oxford & Chichester : John Wiley & Sons, 2013, p. 519-522.

la plus efficace possible ; d'autres techniques d'apprentissage automatique comme les arbres de décision ou les réseaux de neurones.[18]

Pour terminer cette section sur les études de paternité et avant de nous plonger dans la bibliographie sur la stylométrie, laissons la parole à P. Juola qui exprime l'intérêt de ces études :

> « Why care about authorship attribution ? And, especially, why care about statistical methods for doing authorship attribution ? Because "style," and the identity underlying style, has been a major focus of humanistic inquiry since time immemorial. Just as corpus studies have produced a revolution in linguistics, both by challenging longheld beliefs and by making new methods of study practicable, "nontraditional" stylometry can force researchers to re-evaluate long-held beliefs about the individualization of language. »[19]

L'intérêt des études de paternité excède bien entendu le seul domaine des sciences humaines, de la littérature, de l'histoire. L'apparition d'articles traitant de stylométrie dans des revues destinées à un public issu du droit en est la preuve.[20]

3.2 Historique de la stylométrie à travers une sélection d'articles

L'article de F. Mosteller et de D. Wallace, deux statisticiens, consacré aux *Federalist papers* et publié en 1963, est souvent considéré comme la date de naissance de la stylométrie moderne. Le retentissement de cet article fut tel que le problème des *Federalist papers* est utilisé comme un « problème-type » dans de nombreuses publications de stylométrie.[21] Cet article prend les *function words* comme éléments à calculer, une méthode qui connaîtra beaucoup d'avenir.[22]

[18] Juola, « Authorship Attribution », p. 272-286 ; José Nilo G. Binongo et M. Wilfrid A. Smith. « The Application of Principal Component Analysis to Stylometry ». In : *Literary and Linguistic Computing* 14.4 (1999), p. 445 ; Holmes, « The Evolution of Stylometry in Humanities Scholarship », p. 115 ; Matthew L. Jockers et Ted Underwood. « Text-Mining The Humanities ». In : *A New Companion to Digital Humanities*. Sous la dir. de Susan Schreibman, Ray Siemens et John Unsworth. Malden, Oxford & Chichester : Wiley-Blackwell, 2016, p. 294-295.

[19] Juola, « Authorship Attribution », p. 322.

[20] Carole E. Chaski. « Who Wrote It ? Steps Toward a Science of Authorship Identification ». In : *National Institute of Justice Journal* 233.233 (1997), p. 15-22 ; Efstathios Stamatatos. « On the Robustness of Authorship Attribution Based on Character N-Gram Features ». In : *Journal of Law and Policy* 21.2 (2013), p. 421-439.

[21] Juola, « Authorship Attribution », p. 242-243.

[22] Frederick Mosteller et David L. Wallace. « Inference in an Authorship Problem : A Comparative Study of Discrimination Methods Applied to the Authorship of the Disputed Federalist Papers ». In : *Journal of the American Statistical Association* 58.302 (1963), p. 275-309.

En 1994, D. Holmes publie un article de synthèse sur les caractéristiques qui ont été utilisées dans le cadre d'analyses stylométriques, en déterminant leur efficacité.[23] Par exemple, la longueur des mots, mesure proposée par Mendenhall au XIX^{e} siècle, est considérée comme trop peu fiable et trop dépendante du genre ou du sujet. Il souligne également que les méthodes stylométriques sont plus efficaces lorsqu'il y a une homogénéité dans le genre et l'époque. Il s'agit, à notre connaissance, de la première synthèse méthodologique de cette nature. L'auteur fait également preuve d'un certain flair en écrivant :

> « On a personal note, I believe that the future for stylometry over the next decade will lie in the development of connectionist approaches entailing the extensive use of Artificial Neutral Networks. »[24]

En 1994, B. Kjell propose, pour la première fois, de prendre en compte les bigrammes de caractères et d'appliquer une méthode d'apprentissage automatique. Cette proposition était particulièrement novatrice pour l'année de sa parution, ce qui peut expliquer l'erreur dans l'application des réseaux de neurones (que D. Holmes met en évidence dans son article de 1998).[25]

En 1998 sont publiés deux compte-rendus sur la discipline, celui de D. Holmes et de J. Rudman.[26] L'objectif du premier est d'esquisser le développement historique de la stylométrie, en soulignant les articles qui ont été des jalons pour la discipline et en évoquant, d'une façon positive, les progrès qui découleront de l'application grandissante de l'intelligence artificielle. Le second est beaucoup plus critique sur la discipline, déclarant de but en blanc que « [t]here is more wrong with authorship attribution studies than there is right. ».[27] Il souligne les problèmes qu'il y constate, tels que le manque de consensus sur les méthodologies et les techniques les plus efficaces, le manque de références à de précédentes études, la course folle aux dernières innovations en statistique (ou, comme il le formule : « Statistics should not be the tail that wags the dog of attribution studies »[28]), ou encore le manque de soin dans la préparation des textes. Il tente de formuler quelques solutions à ces problèmes, mais c'est la discipline toute entière qui doit évoluer pour dépasser ces problèmes. H. Love répond à une des critiques de J. Rudman, celle qui concerne la stabilité de la discipline : il est incongru d'espérer une stabilisation des techniques en stylométrie alors que les

[23] David I. Holmes. « Authorship Attribution ». In : *Computers and the Humanities* 28.2 (1994), p. 87-106.

[24] Holmes, « Authorship Attribution », p. 104.

[25] Holmes, « The Evolution of Stylometry in Humanities Scholarship », p. 115.

[26] Holmes, « The Evolution of Stylometry in Humanities Scholarship » ; Joseph Rudman. « The State of Authorship Attribution Studies : Some Problems and Solutions ». In : *Computers and the Humanities* 31.4 (1998), p. 351-365.

[27] Rudman, « The State of Authorship Attribution Studies », p. 351.

[28] Rudman, « The State of Authorship Attribution Studies », p. 355.

disciplines dont proviennent les techniques que sont l'intelligence artificielle et les statistiques sont en état d'expansion continuelle.[29]

J. Burrows publie en 2002 un article qui résume sa méthode Delta. Cette méthode se base sur la fréquence relative des mots les plus courants dans un texte (pour la plupart des mots-outils), et applique une ACP sur les résultats pour pouvoir visualiser les résultats obtenus.[30] Selon D. Holmes, la méthode de J. Burrows a le mérite de fonctionner et est la méthode par défaut à son époque.[31]

J. Grieve produit en 2007 une évaluation des caractéristiques retenues dans les textes.[32] Ce sont 35 mesures qui sont comparées, parmi lesquelles la longueur des phrases, la longueur des mots, la richesse du vocabulaire, la fréquence des graphèmes, la fréquence des mots, la fréquence de la ponctuation et la fréquence des n-grammes de caractères. L'auteur conclut que la plupart des méthodes sont efficaces et que les n-grammes sont la meilleure caractéristique.

En 2009 sortent à nouveau deux articles de synthèse, qui font le point sur les caractéristiques à privilégier et sur les méthodes d'attribution les plus efficaces.[33] E. Stamatatos souligne une différence d'approche en ce qui concerne les méthodes d'attribution. Il existe les approches basées sur un profil d'auteur, lors desquelles tous les textes d'un même auteur sont mis ensemble, et les approches basées sur le cas, où des paquets contenant un nombre décidé de mots sont classés selon un algorithme de classification.

W. Daelemans, dans son article de 2013, réclame plus d'explications en stylométrie. L'objectif principal est de faire comprendre les techniques aux chercheurs, pas d'optimiser les performances. Elle remarque également que la stylométrie fonctionne bien en théorie, mais dans des conditions réelles (c'est-à-dire avec de nombreux auteurs-candidats et/ou dans une situation de vérification de paternité plutôt que d'attribution), les techniques sont beaucoup moins efficaces.[34] M. Kestemont répond à son appel à davantage d'explications en publiant, l'année suivante, un article expliquant ce que sont les *function words* et pourquoi ils sont efficaces (voir supra).[35]

[29]Love, *Attributing Authorship*, p. 153.

[30]John Burrows. « 'Delta' : A Measure of Stylistic Difference and a Guide to Likely Authorship ». In : *Literary and Linguistic Computing* 17.3 (2002), p. 267-287.

[31]Holmes, « The Evolution of Stylometry in Humanities Scholarship », p. 114.

[32]Jack Grieve. « Quantitative Authorship Attribution : An Evaluation of Techniques ». In : *Literary and linguistic computing* 22.3 (2007), p. 251-270.

[33]Moshe Koppel, Jonathan Schler et Shlomo Argamon. « Computational Methods in Authorship Attribution ». In : *Journal of the American Society for Information Science and Technology* 60.1 (2009), p. 9-26 ; Efstathios Stamatatos. « A Survey of Modern Authorship Attribution Methods ». In : *Journal of the American Society for Information Science and Technology* 60.3 (2009), p. 538-556.

[34]Walter Daelemans. « Explanation in Computational Stylometry ». In : *Computational Linguistics and Intelligent Text Processing*. 14th International Conference, CICLing 2013 Samos, Greece, March 24-30, 2013 Proceedings, Part II. Sous la dir. d'Alexander Gelbukh. Berlin & Heidelberg : Springer, 2013, p. 454-457.

[35]Kestemont, « Function Words in Authorship Attribution. From Black Magic to Theory? »

Une nouvelle méthode de classification non supervisée est proposée en 2014 par M. Koppel et Y. Winter. Cette méthode, celle des imposteurs généraux, a été spécialement conceptualisée pour répondre à une faiblesse de la stylométrie : les problèmes de vérification de paternité. Elle consiste à générer des imposteurs, qui présentent des caractéristiques aléatoires, et à calculer la similarité entre l'imposteur généré et l'auteur évalué.[36]

P. Juola propose, dans un article de 2015, de mettre en place des protocoles standardisés pour les analyses. Selon lui, l'absence de standards nuit à la discipline et à sa crédibilité auprès des sciences humaines, car l'abondance des méthodes couplée avec ce manque de standards rend la comparaison de différentes méthodes difficile, surtout pour un public qui n'est pas familier avec le domaine.[37]

M. Eder, en 2015, propose un article de type méthodologique qui tente de déterminer le nombre de mots minimum pour obtenir des résultats corrects.[38] Il utilise une approche *bag of words*, qu'il juge plus efficace que des séquences de mots, et utilise la méthode Delta de Burrows même s'il suppose que les résultats obtenus sont indépendants de toute méthode. La moyenne des résultats obtenus se situe aux alentours de 5000 mots, tandis que la prose latine semble n'en demander que 2500.

Le même auteur récidive en 2017 avec un article consacrée à la visualisation.[39] Il propose une méthode qui évite toute forme de *cherry picking* en créant un arbre de consensus plutôt que des dendrogrammes individuels.

En 2017, T. Neal et al. réalisent un nouvel article de synthèse sur les caractéristiques et sur les modèles de classification. Le rythme de ces synthèses est tout de même soutenu et permet de garder une vision d'ensemble sur la discipline. Peu de nouveautés apparaissent dans l'article, si ce n'est l'apparition de caractéristiques liées à la structure du document. Désormais, les modèles d'apprentissage automatique tels que SVM ou KNN sont considérés comme ceux qu'il faut utiliser, et les modèles probabilistes sont relégués au second plan.[40]

---

[36]Moshe Koppel et Yaron Winter. « Determining If Two Documents Are Written by the Same Author ». In : *Journal of the Association for Information Science and Technology* 65.1 (2014), p. 178-187.

[37]Patrick Juola. « The Rowling Case : A Proposed Standard Analytic Protocol for Authorship Questions ». In : *Digital Scholarship in the Humanities* 30 (Suppl. 1 2015), p. i100-i113.

[38]Maciej Eder. « Does Size Matter ? Authorship Attribution, Small Samples, Big Problem ». In : *Digital Scholarship in the Humanities* 30.2 (2015), p. 167-182.

[39]Maciej Eder. « Visualization in Stylometry : Cluster Analysis Using Networks ». In : *Digital Scholarship in the Humanities* 32.1 (2017), p. 50-64.

[40]Tempestt Neal et al. « Surveying Stylometry Techniques and Applications ». In : *ACM Computing Surveys* 50.6 (2017), p. 1-36.

3.3 Stylométrie et langue latine

Dans cette brève section, nous mentionnerons quelques études stylométriques qui ont pour corpus un texte latin. L'objectif n'est pas d'y être exhaustif, mais de souligner que le latin a déjà fait l'objet d'analyses stylométriques et que de nombreuses méthodes ont été mises en pratique.

R. Forsyth et al. propose une application de la méthode Delta sur un texte pseudépigraphe de Cicéron, *De consolatione*, que la littérature associe au nom de Carlo Sigonio, un humaniste italien du XVI^{e} siècle. Il calcule la fréquence des mots-outils, applique une ACP pour visualiser ses résultats et un clustering au moyen de dendrogrammes pour les classifier. Il en conclue que le pseudépigraphe en est bien un, et qu'il a très probablement été écrit par Sigonio. Les auteurs en concluent que l'approche élaborée par Burrows est généralisable à une langue fléchie comme le latin, ce qui n'allait pas de soi vu qu'une langue fléchie repose moins sur les mots-outils qu'une langue comme l'anglais.[41]

La Belgique n'est pas en reste en termes d'études stylométriques sur du latin, puisqu'en 2015 M. Kestemont, S. Moens et J. Deploige, des chercheurs des universités d'Anvers et de Gand, proposent une application de la méthode Delta sur un corpus très intriguant, celui d'Hildegarde Bingen, qui entraîne des questionnements sur la paternité collective : la moniale dicte en effet ses écrits à des secrétaires, qui sont autorisés à corriger son latin. Les auteurs préfèrent une méthode qui regarde les mots-outils plutôt que des n-grammes de caractères, car le latin médiéval est caractérisé par une forte variation dans son orthographe et, sans outil de traitement automatique adéquat, ces différences pourraient engendrer des erreurs dans les résultats. Ils regrettent toutefois de pouvoir uniquement utiliser ces mots-outils, car il y a une perte importante des informations contenues dans les terminaisons des mots. Leurs analyses permettent de différencier nettement les œuvres d'Hildegarde de celles de ses secrétaires, en remarquant tout de même un phénomène intéressant au niveau des œuvres où la marque laissée par Guibert de Gembloux était plus présente : ses quelques œuvres semblent présenter un style « combiné ».[42]

Retour à du latin classique et plus précisément au corpus césarien avec un nouvel article de M. Kestemont associé à d'autres chercheurs, notamment M. Koppel qui avait formulé la méthode appliquée dans cet article deux ans plus tôt.[43] Le corpus césarien est un cas typique de vérification de paternité : il s'agit de déterminer si César a écrit ou non les écrits conservés dans le corpus de textes qui couvrent les guerres qu'il

[41] Richard S. Forsyth, David I. Holmes et Emily K. Tse. « Cicero, Sigonio, and Burrows : Investigating the Authenticity of the Consolatio ». In : *Literary and Linguistic Computing* 14.3 (1999), p. 375-400.

[42] Mike Kestemont, S. Moens et Jeroen Deploige. « Collaborative Authorship in the Twelfth Century : A Stylometric Study of Hildegard of Bingen and Guibert of Gembloux ». In : *Digital Scholarship in the Humanities* 30.2 (2015), p. 199-224.

[43] Koppel et Winter, « Determining If Two Documents Are Written by the Same Author ».

a menées depuis la Gaule jusqu'à l'Espagne. Les vérifications de paternité sont plus complexes que les attributions, faute de méthodes adéquates. Les auteurs appliquent une méthode nouvelle, celle des imposteurs généraux, qui consistent à mesurer des distances entre le profil stylistique d'un auteur donné et celui d'un texte généré aléatoirement. Les résultats auxquels ils parviennent sont en accord avec ceux de la philologie traditionnelle, à savoir l'implication d'Hirtius dans le dernier livre de la *Guerre des Gaules* et la participation d'autres auteurs que César ou Hirtius pour les œuvres en fin de corpus, précisément la *Guerre d'Espagne* et la *Guerre d'Afrique*.[44]

L'application de méthodes stylométriques ne sert pas uniquement à identifier des auteurs. M. Eder se base sur la mesure des mots les plus fréquents pour obtenir une vue d'ensemble de la littérature latine sous la forme d'un réseau. Dans ce dernier se distinguent aisément des clusters qui correspondent aux genres littéraires, à la datation, etc.[45]

La thèse de J. De Gussem, sous la direction de J. Deploige, explore plus en profondeur le dossier d'Hildegarde de Bingen effleuré par l'article de 2016 précédemment évoqué, ainsi que plus largement les questions de paternité collective au XII^{e} siècle. Elle explore quelles sont les conséquences sur une quantification du style produites par une culture manuscrite du Moyen Âge qui voit les textes sans cesse filtrés et remaniés à travers la tradition orale, la participation de secrétaires entraînés à imiter le style de leur *dictator* (celui qui dicte), l'insertion de sources d'autorité dans les textes sans le mentionner explicitement. L'auteur souligne également que le latin est une langue qui se prête volontiers à la stylométrie, puisqu'elle possède ses difficultés propres : en particulier sa nature fléchie qui doit avoir des conséquences sur l'utilisation d'une méthode reposant intégralement sur les mots-outils, mais également le fait que la littérature latine qui nous est parvenue doit être considérée comme du *dirty data*, puisque les textes ont été modifiés au cours de leur transmission.[46]

Les études sur du latin médiéval étant assez rares, il nous semble pertinent de mentionner un article qui porte sur deux auteurs du XII^{e} siècle : le moine de Lido et Gallus Anonyme, qui sont selon les résultats des analyses stylométriques une seule et même personne. Soulignons également l'application d'une méthode singulière, la distance de Bray-Curtis. Nous n'entrerons toutefois pas dans les détails de cet article, puisque là n'est pas l'objectif de notre mémoire.[47]

---

[44]Mike Kestemont et al. «Authenticating the Writings of Julius Caesar». In : *Expert Systems with Applications* 63 (2016), p. 86-96.

[45]Maciej Eder. «A Bird's-Eye View of Early Modern Latin : Distant Reading, Network Analysis, and Style Variation». In : *Early Modern Studies After the Digital Turn*. Sous la dir. de Laura Estill, Diane K. Jakacki et Michael Ullyot. Toronto : Iter Academic Press, 2016, p. 63-90.

[46]De Gussem, «Collaborative Authorship in Twelfth-Century Latin Literature», p. 4-6, 27, 76-77.

[47]Jakub Kabala. «Computational Authorship Attribution in Medieval Latin Corpora : The Case of the Monk of Lido (ca. 1101–08) and Gallus Anonymous (ca. 1113–17)». In : *Language Resources and Evaluation* 54.1 (2020), p. 25-56.

Un volume du journal *Interfaces*, édité par J. De Gussem et J. Deploige en 2021, est dédié aux questions de stylométrie au Moyen Âge. J. De Gussem revient sur le dossier d'Hildegarde de Bingen, sans réelle innovation par rapport aux résultats de sa thèse (l'article porte par ailleurs le même nom que le chapitre de sa thèse dédié au dossier d'Hildegarde).[48] E. Leclercq et M. Kestemont, quant à eux, y contribuent un article consacré à des analyses stylométriques sur des chartes médiévales. Ce matériau pose des difficultés particulières, puisque sa transmission est particulièrement étagée : la volonté de celui qui émet la charte, la patte du secrétaire qui la copie et des scribes qui l'ont recopiée, l'intervention de l'éditeur moderne. Malgré ces difficultés, une méthode somme toute devenue traditionnelle (mesure de la fréquence lexicale suivie d'un algorithme d'apprentissage automatique non-supervisé pour classifier les données) donne des résultats prometteurs.[49]

Ce dernier article, lui aussi daté de 2021, jette un pont entre cette section et la suivante. Il s'agit de l'application de l'analyse de réseaux pour quantifier le style de l'*Historia Augusta*, ouvrage dont les analyses stylométriques précédentes ont pu provoquer un certain rejet parmi les tenants d'une philologie qualitative.[50] Il s'agit donc d'un terrain miné. Les auteurs obtiennent toutefois des résultats intéressants, tant sur le plan philologique (les analyses semblent confirmer plusieurs auteurs) que méthodologiques (l'analyse de réseaux est efficace sur du latin, tout autant voire davantage que la méthode des imposteurs généraux). Ils mesurent les propriétés des réseaux de co-occurrence de mots, les normalisent, puis les classent avec l'algorithme d'apprentissage automatique supervisé KNN.[51]

3.4 Stylométrie et analyse de réseaux

Cette section est consacrée à un historique des différentes publications au sujet de l'analyse de réseaux appliquée à des données textuelles, qui commence au début des années 2000.

Si l'article de A. Martins et al. est le premier à appliquer l'analyse de réseaux à un corpus latin (cette méthode n'est pas reprise dans l'état de l'art que réalise J. De Gussem dans sa thèse[52] et nos propres recherches bibliographiques n'ont rien produit), ce n'est pas la première fois que les réseaux sont appliqués à des données linguistiques.

[48] Jeroen De Gussem. « Larger than Life ? A Stylometric Analysis of the Multi-Authored Vita of Hildegard of Bingen ». In : *Interfaces : A Journal of Medieval European Literatures* 8 (2021), p. 125-159.

[49] Eveline Leclercq et Mike Kestemont. « Advances in Distant Diplomatics : A Stylometric Approach to Medieval Charters ». In : *Interfaces : A Journal of Medieval European Literatures* 8 (2021), p. 214-244.

[50] Nous pensons à Ian Marriott. « The Authorship of the Historia Augusta : Two Computer Studies ». In : *The Journal of Roman Studies* 69 (1979), p. 65-77.

[51] Armando Martins et al. « Historia Augusta Authorship : An Approach Based on Measurements of Complex Networks ». In : *Applied Network Science* 6.1 (2021), p. 50.

[52] De Gussem, « Collaborative Authorship in Twelfth-Century Latin Literature », p. 78-79.

Ces applications sont appelées « réseaux linguistiques », parmi lesquels le modèle le plus couramment employé est celui de la co-occurrence de mots : les sommets du réseau représentent les mots, tandis que les arêtes sont des liens de proximité entre eux (directement d'adjacence ou légèrement plus éloignés, avec des distances de deux ou de trois espaces). Ces co-occurrences peuvent s'expliquer de diverses manières : les relations syntaxiques entre certaines natures de mot (article et nom par exemple), des expressions stéréotypées... Un des avantages de ces réseaux de co-occurrence : ils sont indépendants de la langue, et ne demandent pas de traitement linguistique lourd au préalable.[53]

Deux articles de 2001 posent les bases en ce domaine. Ils partent tous deux du principe que les langues sont des systèmes complexes, où les mots interagissent de manière non aléatoire. Un de ces articles, moins intéressant pour notre sujet, tente de mesurer l'évolution d'une langue grâce aux changements dans ses réseaux linguistiques.[54] L'autre article repère que des propriétés associées aux réseaux complexes émergent de la modélisation de données linguistiques : l'effet « petit monde », en cela que la distance moyenne entre deux sommets n'est pas de plus de deux ou trois arêtes, et le caractère invariant d'échelle, à savoir que les degrés du réseau suivant une loi de puissance (un petit nombre de sommets possède un grand nombre d'arêtes). En particulier, ces sommets avec un grand degré sont souvent des mots-outils, qui servent de « raccourcis » entre les différents sommets du réseau.[55]

L. Antiqueira et al. est le groupe de chercheurs à appliquer en premier l'analyse de réseaux pour quantifier le style d'un auteur. Leur article de 2007, qui nous sert de référence, a été précédé d'un article de colloque, mais il fait office de synthèse et nous le privilégions par conséquent.[56] Les chercheurs démontrent que les réseaux d'auteurs différents présentent des mesures topologiques différentes (en particulier le coefficient de clustering, le degré sortant des sommets et la corrélation de degré) : cette observation semble confirmer que les réseaux permettent bel et bien de capturer le style d'un auteur. Les auteurs formulent également la méthode à appliquer pour transformer un texte en réseau, une méthode qui sera appliquée par les articles ultérieurs : d'abord pré-traiter le texte en supprimant les mots-outils ainsi que la ponctuation et en lemmatisant les mots, ensuite représenter chaque lemme comme un sommet et chaque arête dirigée comme un lien d'adjacence entre un mot et un autre (ils ne considèrent donc que les bigrammes de mots).[57] Ils soulignent également que

[53] Ramon Ferrer i Ferer-i-Cancho et Richard V. Solé. « The Small World of Human Language ». In : *Proceedings of the Royal Society of London. Series B : Biological Sciences* 268.1482 (2001), p. 2261.

[54] S. N. Dorogovtsev et J. F. F. Mendes. « Language as an Evolving Word Web ». In : *Proceedings of the Royal Society of London. Series B : Biological Sciences* 268.1485 (2001), p. 2603-2606.

[55] Ferer-i-Cancho et Solé, « The Small World of Human Language ».

[56] Lucas Antiqueira et al. « Some Issues on Complex Networks for Author Characterization ». In : *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial* 11.36 (2007), p. 51-58.

[57] Antiqueira et al., « Some Issues on Complex Networks for Author Characterization », p. 54.

l'objectif ne doit pas être de trouver la mesure parfaite (un travers commun en stylométrie, comme nous avons pu le constater), mais d'identifier l'ensemble de mesures qui permet de caractériser le style en se combinant. Enfin, ils ne mettent pas en place une méthodologie de classification, laissant ce soin à des publications futures.

En 2011, A. Mehri et al. reprennent la méthodologie proposée par L. Antiqueira et al., si ce n'est qu'ils modélisent des réseaux non-dirigés. Ils prennent en compte de nouvelles mesures, comme le degré moyen, la distribution des degrés, la moyenne du degré des plus proches voisins. Le coefficient de clustering semble demeurer une mesure importante. Pour classifier les profils d'auteur ainsi obtenus, ils résument les différentes mesures dans des vecteurs et mettent en place un calcul basé sur la distance entre le vecteur du texte à identifier et le vecteur qui correspond au profil de l'auteur. Ils obtiennent une efficacité de 91%.[58]

La même année, D. Amancio et al. proposent une classification basée sur un calcul de la corrélation entre les mesures topologiques (avec KNN), mais obtiennent une efficacité bien inférieure de 65%. Les mesures qu'ils retiennent sont le coefficient de clustering, encore et toujours, mais également le plus petit chemin moyen (plus petit nombre d'arêtes entre deux sommets donnés) et l'intermédiarité (à quel point un sommet est central dans un réseau, c'est-à-dire le nombre de plus petits chemins par lesquels il est traversé).[59]

S. Segarra et al., d'abord dans un article de colloque[60] et puis dans un article de journal plus détaillé,[61] proposent une nouvelle façon de modéliser des textes en réseaux. Ils proposent de ne garder que les mots-outils (à l'exception des pronoms, considérés comme trop dépendants du genre littéraire), de les placer dans un réseau de co-occurrence de mots (en traçant des arêtes lorsque les mots sont proches l'un de l'autre, avec une distance maximale de 10 sommets), et de les interpréter comme des chaînes de Markov, qui peuvent être quantifiées par une mesure d'entropie. Contrairement à des méthodes basées sur la fréquence de ces fameux mots-outils, comme dans le célèbre article de F. Mosteller et D. Wallace ou dans la méthode Delta de J. Burrows, cette modélisation étudie davantage leur structure relationnelle. La méthode se révèle assez efficace, en obtenant entre 80 et 85% de justesse, et la combiner avec une méthode de fréquence permet d'encore en améliorer l'efficacité. Les auteurs soulignent que la justesse diminuent lorsque les textes étudiés sont trop courts et/ou

[58]Ali Mehri, Amir H. Darooneh et Ashrafalsadat Shariati. « The Complex Networks Approach for Authorship Attribution of Books ». In : *Physica A : Statistical Mechanics and its Applications* 391.7 (2012), p. 2429-2437.

[59]Diego Raphael Amancio et al. « Comparing Intermittency and Network Measurements of Words and Their Dependence on Authorship ». In : *New Journal of Physics* 13.12 (2011), p. 123024.

[60]Santiago Segarra, Mark Eisen et Alejandro Ribeiro. « Authorship Attribution Using Function Words Adjacency Networks ». In : *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. Vancouver, Canada : IEEE, 2013, p. 5563-5567.

[61]Santiago Segarra, Mark Eisen et Alejandro Ribeiro. « Authorship Attribution Through Function Word Adjacency Networks ». In : *IEEE Transactions on Signal Processing* 63.20 (2015), p. 5464-5478.

lorsque le nombre d'auteurs-candidats est trop élevé. Suivant le corpus à notre disposition, il faut jouer sur ces deux paramètres.

En 2015, D. Amancio produit une synthèse sur l'approche basée sur les réseaux pour les données linguistiques, plus large que les simples questions de paternité qui sont tout de même évoquées longuement. Plusieurs modélisations sont possibles, suivant le but recherché, mais dans le cas d'une analyse stylistique, les relations syntaxiques (et par conséquent les réseaux d'adjacence de mots) sont les plus efficaces. L'auteur souligne le caractère redondant des mots-outils, qui relient eux-mêmes des mots, et peuvent donc être considérés comme des arêtes. Nous ne sommes pas réellement en accord avec cette affirmation : la présence de mots-outils est ce qui donne aux réseaux linguistiques l'effet petit-monde (précédemment mentionné), et l'importance pour caractériser le style d'un auteur a été longuement explicitée par la littérature stylométrique. Quelques mesures topologiques sont mises en évidence : le degré, l'intermédiarité, le coefficient de clustering, le plus petit chemin moyen, auxquelles l'auteur ajoute l'assortativité (les mots avec des fréquences distinctes qui apparaissent comme voisins dans le graphe) et l'accessibilité (qui est un type de mesure de centralité). Pour la classification, c'est une méthode supervisée qui est mise en œuvre, sans grande surprise. L'auteur conclut en soulignant que les mesures de réseaux se combinent avec des approches basées sur la fréquence et permettent d'en améliorer l'efficacité.[62]

Une nouvelle mesure est proposée par V. Marinho et al. : la fréquence absolue des motifs avec au moins trois sommets qui apparaissent dans le réseau. Les motifs sont des « sous-réseaux » dont les sommets sont interconnectés entre eux. Les auteurs réutilisent des mesures précédemment mentionnées comme le degré moyen des sommets voisins, le plus petit chemin moyen, l'intermédiarité, le coefficient de clustering, l'assortativité... Après classification avec plusieurs méthodes (SVM, KNN, Bayes naïf), les résultats obtenus se révèlent plus élevés que l'aléatoire, mais pas très concluant (57.5%). Les auteurs soulignent la grande importance des mots-outils : l'efficacité de la classification chute drastiquement si on les enlève des réseaux de co-occurrence.[63]

Un an plus tard, C. Akimushkin et al. testent la fiabilité des réseaux de co-occurrence sur des petits morceaux de texte dont ils retirent les mots-outils et qu'ils lemmatisent (100 ou 200 sommets seulement). Avec une classification par KNN, il obtient un taux d'efficacité de 89%, prouvant que l'analyse de réseaux permet de caractériser efficacement le style d'un auteur même avec un petit nombre de mots. Les mesures suivantes

[62]Diego Raphael Amancio. « A Complex Network Approach to Stylometry ». In : *PLOS ONE* 10.8 (2015), e0136076.

[63]Vanessa Queiroz Marinho, Graeme Hirst et Diego Raphael Amancio. « Authorship Attribution via Network Motifs Identification ». In : *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*. 2016 5th Brazilian Conference on Intelligent Systems (BRACIS). 2016, p. 355-360.

sont extraites : le coefficient de clustering, le diamètre et le rayon du réseau, le nombre de cliques, la centralité. La méthode consiste à observer et comparer l'évolution de la topologie des réseaux entre les différents réseaux obtenus à partir des morceaux de textes.[64]

W. Goh et al., dans un article qui met en évidence l'importance des motifs comme mesure topologique, soulignent l'existence de deux approches en analyse de réseaux de co-occurrence de mots : une approche représentée par la plupart des articles qui suppriment les mots-outils, et une approche proposée par Segarra et al. qui ne gardent qu'eux dans leur réseau de co-occurrence. Les deux approches semblent toutefois efficaces.[65]

Nous avons tenu à intégrer cet article de H. de Arruda et al., même s'il ne pourra pas nous être utile pour nos analyses. En effet, les auteurs proposent un nouveau modèle de réseau, basé non plus au niveau des mots, mais des paragraphes. Même si notre corpus n'est pas divisé en paragraphes, cette approche innovante méritait d'être mentionnée.[66]

Dans un article de 2019, T. Stanisz et al. comparent plusieurs mesures topologiques de réseaux de co-occurrence pour déterminer celles qui seraient les plus à même de capturer des informations sur le style d'un auteur. C'est également l'un des rares articles à expliquer à la fois ce que représente une mesure dans un réseau et ce que cette mesure représente dans le cadre d'un réseau linguistique. Les auteurs utilisent le degré des sommets (ce qui représente la fréquence du mot dans le texte), le plus petit chemin moyen, le coefficient de clustering (ce qui représente la probabilité qu'un mot apparaisse à côté d'un autre plus d'un fois dans l'extrait), le coefficient d'assortativité, la modularité (qui désigne la possibilité de diviser le réseau en cliques, ce qui représente des groupes de mots apparaissant ensemble dasn un texte). Les mesures sont normalisées, et sont classifiées grâce à des méthodes d'apprentissage automatique (notamment des dendrogrammes). Les auteurs appliquent cette méthode à un corpus bilingue (anglais et polonais), en obtenant des résultats similaires, ce qui renforce l'hypothèse de l'indépendance vis-à-vis de la langue que posséderaient les réseaux de co-occurrence. Ils repèrent toutefois que les mesures globales du réseau semblent communes à une langue, tandis que les mesures locales (propres à des sommets donnés) permettent de caractériser le style d'un auteur. Deux mesures sont retenues, qui attribuent correctement un texte dans 85-90% des cas : le degré du

[64]Camilo Akimushkin, Diego Raphael Amancio et Osvaldo Novais Oliveira Jr. « Text Authorship Identified Using the Dynamics of Word Co-Occurrence Networks ». In : *PLOS ONE* 12.1 (2017), e0170527.

[65]Woon Peng Goh, Kang-Kwong Luke et Siew Ann Cheong. « Functional Shortcuts in Language Co-Occurrence Networks ». In : *PLOS ONE* 13.9 (2018), e0203025.

[66]Henrique F. de Arruda et al. « Paragraph-Based Representation of Texts : A Complex Networks Approach ». In : *Information Processing & Management* 56.3 (2019), p. 479-494.

sommet et son coefficient de clustering, tous deux normalisés.[67]

L. Quispe et al. proposent d'améliorer les réseaux de co-occurrence de mots en y ajoutant des « arêtes virtuelles » qui représenteraient des liens sémantiques entre sommets. Ces liens sont obtenus grâce à l'application de plongements lexicaux sur les textes étudiés. Les résultats de cette nouvelle approche sont meilleurs que ceux obtenus sans ajouter les « arêtes virtuelles » sémantiques.[68]

Nous terminons cette section avec le même article que la précédente, celui qui nous a inspiré pour appliquer l'analyse de réseaux dans ce mémoire. Leur méthode a la particularité de ne pas appliquer de lemmatisation aux textes, permettant de conserver la flexion des mots. Les mots-outils ont soit été retirés soit conservés, sans réel impact sur les performances de la méthode. Les textes sont découpés en paquets de taille donnée (100, 200, 500). Les mesures utilisées sont le nombre total de sommets (c'est-à-dire le nombre de mots uniques dans l'extrait), le nombre total d'arêtes (c'est-à-dire le nombre de connexions entre mots successifs), le degré total du sommet (comprenant à la fois le degré entrant et le degré sortant), le coefficient de clustering (c'est-à-dire le nombre de voisins mutuels dans des sommets adjacents, ce qui permet d'identifier une tendance à utiliser des mots chargés d'un contenu sémantique spécifique), la longueur du plus petit chemin moyenne (c'est-à-dire le chemin le plus court d'un sommet à un autre, ce qui permet de repérer les mots les plus importants et les plus centraux d'un texte), le diamètre et le rayon du réseau, l'intermédiarité (qui dénote l'utilisation d'un vocabulaire générique en cas de haute intermédiarité, ou de vocabulaires spécifiques reliés par des mots centraux), l'assortativité (qui permet de quantifier à quel point les liens entre sommets sont dûs à des similarités ou des différences), le nombre de cliques (c'est-à-dire de sous-ensemble du réseau dont tous les sommets sont reliés les uns aux autres, chaque sous-ensemble représentant un groupe de mots utilisés ensemble dans le texte). Les auteurs démontrent que les mesures topologiques, combinées avec KNN comme classificateur, permettent d'obtenir des résultats pertinents dès 100 mots dans un paquet.[69]

[67]Tomasz Stanisz, Jarosław Kwapień et Stanisław Drożdż. « Linguistic Data Mining with Complex Networks : A Stylometric-Oriented Approach ». In : *Information Sciences* 482 (2019), p. 301-320.

[68]Laura V. C. Quispe, Jorge A. V. Tohalino et Diego R. Amancio. « Using Virtual Edges to Improve the Discriminability of Co-Occurrence Text Networks ». In : *Physica A : Statistical Mechanics and its Applications* 562 (2021).

[69]Martins et al., « Historia Augusta Authorship ».

Chapitre 4

Méthodologie

Dans ce chapitre, nous expliciterons les méthodes que nous avons employées dans le cadre de notre mémoire. Nous reviendrons sur des éléments que nous avons déjà évoqués auparavant, dans l'état de l'art (section 3.4). Notre méthodologie est divisée en quatre étapes : l'acquisition des données, le pré-traitement des textes, la création des réseaux et la récolte de leurs mesures topologiques, l'application de techniques d'apprentissage automatique pour visualiser et classer ces mesures afin d'obtenir des réponses à nos questions de paternité.

Ces étapes reflètent ce que la littérature a pu appeler un « algorithme d'attribution ».[1] Nous empruntons l'illustration suivante à T. Neal et al.,[2] même si dans notre cas, les *training features* proviennent du même corpus d'origine :

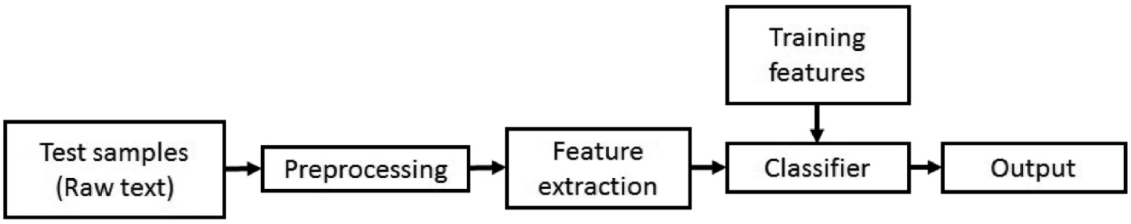


FIG. 4.1 : Algorithme d'attribution

Dans ce chapitre, nous ferons également appel au code R qui figure en annexe de ce mémoire. L'insertion des fichiers R dans le fichier LaTeX entraîne malheureusement la disparition des numéros de ligne. Nous tâcherons tout de même d'être limpide dans nos explications.

[1] GRIEVE, « Quantitative Authorship Attribution », p. 254.
[2] NEAL et al., « Surveying Stylometry Techniques and Applications », p. 6.

4.1 Acquisition et sélection des textes

Comme le rappellent M. Jockers et T. Underwood, la bonne pratique philologique insiste que « *identifying an accurate edition is the first step in responsible research ; here it seems impossible.* »[3] Néanmoins, la littérature demande que les données utilisées doivent être les plus correctes possibles, avec comme objectif de rassembler le plus possible au manuscrit originel.[4] Il semble y avoir un paradoxe entre les données qui sont souhaitables et les données qui sont réalistement accessibles.

Nos données proviennent de la Patrologie latine, qui a l'avantage d'être numérisée sous format texte. Les éditions de référence, bien qu'elles datent également du XIX^{e} siècle, sont celles de A. Olleris et de N. Bubnov. Le défaut de ces dernières se situe dans leur format : les numérisations sont des fichiers PDF, dont l'OCR est par ailleurs imparfait. Après avoir consulté attentivement l'état des fichiers textes de la Patrologie latine et constaté que leur état était respectable (voir la figure 4.2), nous avons décidé d'utiliser cette édition-là.

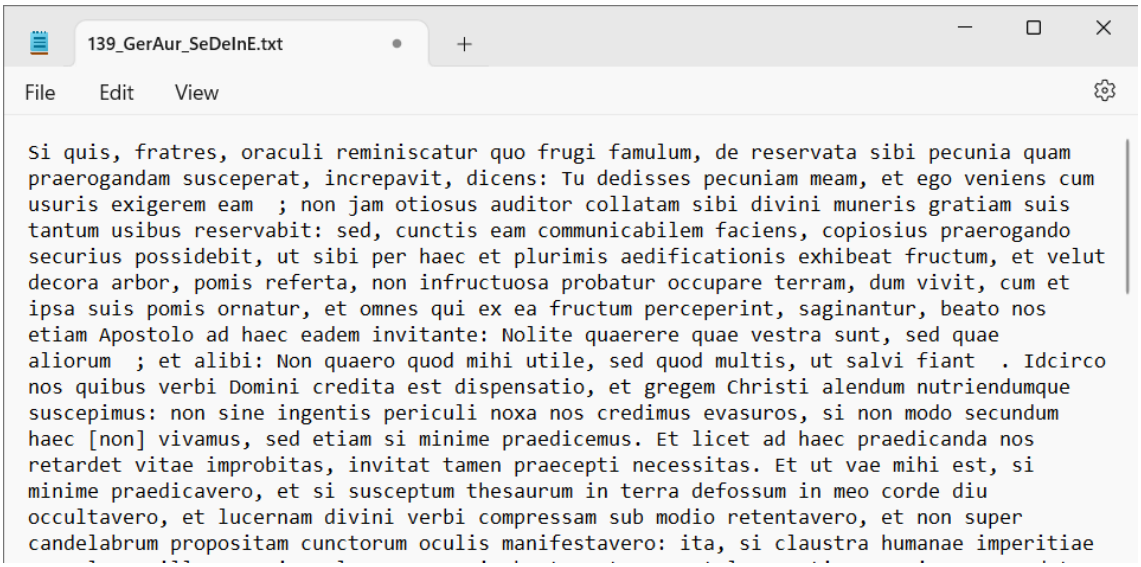


FIG. 4.2 : Le début du *Sermo de informatione episcoporum* dans l'édition numérisée de la Patrologie latine

Nous avons sélectionné un total de quatre auteurs, identifiés sur base des noms rencontrés dans la littérature scientifique que nous avons pu consulter (voir le chapitre 2). Il y a bien sûr Gerbert d'Aurillac (« GerAur »), dont nous avons volontairement exclu les écrits publiés sous le nom de Sylvestre II : la littérature identifie des faux dans ses actes pontificaux, qui représente la plus grande masse de documents produits alors qu'il était pape, et le corpus gerbertien est déjà suffisamment vaste, en particulier grâce à sa correspondance. Nous avons également sélectionné Hermann Contract

[3] JOCKERS et UNDERWOOD, « Text-Mining The Humanities », p. 300.
[4] JUOLA, « Authorship Attribution », p. 320.

(« HerCon ») et Adhémar de Chabannes (« AdeCib ») pour leurs relations avec les deux écrits attribués à Gerbert qui sont au cœur de notre recherche : Adhémar serait le véritable auteur du *Sermo*, tandis que le *De utilitatibus astrolabii* a été attribué à Hermann (c'est encore le cas dans la Patrologie latine, puisque le « DeUtAs » se retrouve dans les écrits de « HerCon ») et que, même si la littérature a prouvé que ce n'était pas le cas, il reste une figure centrale dans l'histoire de l'introduction de l'astrolabe et est à l'origine d'une compilation des traités astrolabiques parmi lesquels notre traité se retrouve. Enfin, nous avons retenu Abbon de Fleury, contemporain et grand rival de Gerbert : nous trouvions tout à fait approprié de le confronter face à ce dernier lors des analyses stylométriques, de la même façon que les deux ecclésiastiques se sont confrontés au cours de leur vie. La figure 4.3 illustre les textes retenus.

Source	Details	Fichier	Auteur	Titre	Type	Tokens
PatrologiaLatina	volume 139	139_AbbFlo_ApAdHuE	Abbo Floriacensis	Apologeticus ad Hugonem et Rodbertum reges Francorum	MISC	4598
PatrologiaLatina	volume 139	139_AbbFlo_CarAcr2	Abbo Floriacensis	Carmen acrostichum	CARM	247
PatrologiaLatina	volume 139	139_AbbFlo_Episto223	Abbo Floriacensis	Epistolae	EPIT	18536
PatrologiaLatina	volume 139	139_AbbFlo_ExDeViR	Abbo Floriacensis	Excerptum de vitis Romanorum pontificum	MISC	15671
PatrologiaLatina	volume 139	139_AbbFlo_PrCoInC	Abbo Floriacensis	Praefatio commentarii in cyclum Victorii	EXEG	533
PatrologiaLatina	volume 139	139_AbbFlo_QuaGra	Abbo Floriacensis	Quaestiones grammaticales	TRAC	5771
PatrologiaLatina	volume 139	139_AbbFlo_ViSEa	Abbo Floriacensis	Vita S. Eadmundi	VITE	5592
PatrologiaLatina	volume 139	139_GerAur_AliEpi	Gerbertus Auriliacensis	Aliae epistolae	EPIT	9337
PatrologiaLatina	volume 139	139_GerAur_DeCaDiA	Gerbertus Auriliacensis	De causa diversitatis arearum in trigono aequilatero	TRAC	433
PatrologiaLatina	volume 139	139_GerAur_DeCoEtS5	Gerbertus Auriliacensis	De corpore et sanguine Domini	TRAC	4991
PatrologiaLatina	volume 139	139_GerAur_DeGeo	Gerbertus Auriliacensis	De geometria	TRAC	20898
PatrologiaLatina	volume 139	139_GerAur_DeNuDi	Gerbertus Auriliacensis	De numerorum divisione	TRAC	1956
PatrologiaLatina	volume 139	139_GerAur_DeRaEtR	Gerbertus Auriliacensis	De rationali et ratione uti	TRAC	4553
PatrologiaLatina	volume 139	139_GerAur_DeSpCo	Gerbertus Auriliacensis	De sphaerae constructione	TRAC	623
PatrologiaLatina	volume 139	139_GerAur_EpScAnS	Gerbertus Auriliacensis	Epistolae scriptae ante summum pontificatum	EPIT	18373
PatrologiaLatina	volume 139	139_GerAur_ExECoMo	Gerbertus Auriliacensis	Excerpta e concilio Mosomensi	REGL	1435
PatrologiaLatina	volume 139	139_GerAur_PrDeFo	Gerbertus Auriliacensis	Profession de foi	MISC	308
PatrologiaLatina	volume 139	139_GerAur_SeDeInE	Gerbertus Auriliacensis	Sermo de informatione episcoporum	SERM	3758
PatrologiaLatina	volume 139	139_GerAur_SuAdEp	Gerbertus Auriliacensis	Supplementum ad epistolas	EPIT	1535
PatrologiaLatina	volume 141	141_AdeCib_AcAdRo	Ademarus Cibardi	Acrostichon ad Rohonem	CARM	339
PatrologiaLatina	volume 141	141_AdeCib_CoAbLeB	Ademarus Cibardi	Commemoratio abbatum Lemovicensium basilicae S. Martialis	CHRO	2338
PatrologiaLatina	volume 141	141_AdeCib_EpDeApS	Ademarus Cibardi	Epistola de apostolatu S. Martialis	EPIT	11653
PatrologiaLatina	volume 141	141_AdeCib_FrSeInC	Ademarus Cibardi	Fragmentum Sermonis in Concilio Lemovicensi	REGL	429
PatrologiaLatina	volume 141	141_AdeCib_Histor4	Ademarus Cibardi	Historiae	CHRO	16836
PatrologiaLatina	volume 143	143_HerCon_CaDeCoO	Hermannus Contractus	Carmen de conflictu ovis et lini	CARM	5354
PatrologiaLatina	volume 143	143_HerCon_Chroni25	Hermannus Contractus	Chronicon	CHRO	69023
PatrologiaLatina	volume 143	143_HerCon_DeMeAs	Hermannus Contractus	De mensura astrolabii	TRAC	3917
PatrologiaLatina	volume 143	143_HerCon_DeUtAs	Hermannus Contractus	De utilitatibus astrolabii	TRAC	9117
PatrologiaLatina	volume 143	143_HerCon_OpuMus	Hermannus Contractus	Opuscula musica	MISC	12930
PatrologiaLatina	volume 143	143_HerCon_SeDeBMa	Hermannus Contractus	Sequentia de B. Maria Virgine	CARM	325

FIG. 4.3 : La liste (provisoire) des textes retenus

Ce choix de textes permet d'obtenir un corpus à peu près de la même époque (Hermann a vécu un demi-siècle après les trois autres auteurs), et diversifié au niveau du genre. Nous ne disposons pas d'assez de textes pour entraîner des modèles uniquement sur des textes du même genre que le texte mis en doute (d'autres traités sur l'astrolabe pour le *De utilitatibus astrolabii* par exemple). Afin d'uniformiser du moins que nous le pouvons ce corpus hétéroclite, nous avons pris la décision d'écarter toute production poétique (que nous reconnaissons grâce à l'abréviation « CARM » dans la colonne « type »). La Patrologie édite encore le *De utilitatibus astrolabii* sous le nom d'Hermann et combiné avec un autre traité qui n'est clairement pas du même auteur. Nous avons donc divisé manuellement le fichier texte en deux parties, intitulées « DeUtAs1 » et « DeUtAs2 » (pour *Liber primus* et *Liber secundus*, noms utilisés dans le texte). Le premier des deux fichiers est notre traité mystère. Nous avons repéré, en-

fin, la présence d'autres textes douteux dans la liste de la Patrologie, en l'occurrence le *De corpore et sanguine Domini* et le *De geometria*, tous deux de Gerbert. Ils sont conservés dans le pré-traitement et l'analyse de réseaux, mais sont exclus de l'étape de classification. Il en va de même pour les deux textes auxquels nous nous intéressons particulièrement, qui sont mis à l'écart de cette même étape. Au terme de ces différentes sélections, chaque auteur conserve chacun au moins 30.000 tokens (Gerbert et Adhémar), Abbon en compte 50.000 et Hermann 85.000.

4.2 Pré-traitement

La littérature stylométrique le caractère essentiel de l'étape de pré-traitement. Il est nécessaire de retirer tout ajout qui n'est pas le fait de l'auteur du texte lui-même.[5] C'est un des objectifs du fichier `01_pretraitementTexte.R`, l'autre étant de tokéniser et lemmatiser les textes.

Les fonctions `monTreeTagger` et `tokenisationPerl` permettent de faire appel à TreeTagger et au script Perl sans sortir de R. Ils prennent en variables les éléments qui changent. Dans le cas de TreeTagger, le nom du fichier en entrée et celui en sortie. Dans le cas du script Perl, uniquement le nom du fichier en entrée puisque le fichier tokénisé est automatiquement nommé par le script.

Une première boucle itère dans la liste des fichiers textes pour les nettoyer. Les éléments que nous considérons comme des impuretés sont capturés par des expressions régulières et supprimés. Que considérons-nous comme des impuretés ? Les balises HTML qui apparaissent parfois en début et en fin de fichier (notamment <p>) ainsi que les chiffres arabes (qui sont assurément des ajouts des éditeurs, et de toute manière n'ont pas beaucoup d'importance une fois l'étape de lemmatisation passée) et les espaces blancs excédentaires. Nous avons longtemps hésité à ne pas retirer le contenu des crochets et des parenthèses, puisqu'elles contiennent des éléments qui sont le fait de l'auteur. Néanmoins, une inspection manuelle des fichiers m'a indiqué que c'était très régulièrement des notes de l'éditeur ou des liens mal résolus. Nous sommes partis du postulat que les quelques éléments qui étaient bien le fait de l'auteur et qui ont été retirés de la sorte n'auront, au vu de la masse des textes, pas un grand impact.

La fin du code inclut deux nouvelles boucles, qui exécutent les fonctions de tokénisation et de lemmatisation précédemment déclarées sur l'ensemble des fichiers. Les fichiers temporaires sont sauvegardés dans un dossier `tmp`. Une fois l'étape de lemmatisation effectuée, les fichiers désormais prêts à la transformation en réseaux sont copiés manuellement du dossier temporaire vers le dossier `dataLemmatisees`.

[5] Juola, « Authorship Attribution », p. 247-248.

Pour l'illustration des étapes de tokénisation et de lemmatisation, voir respectivement les figures 4.4 et 4.5.

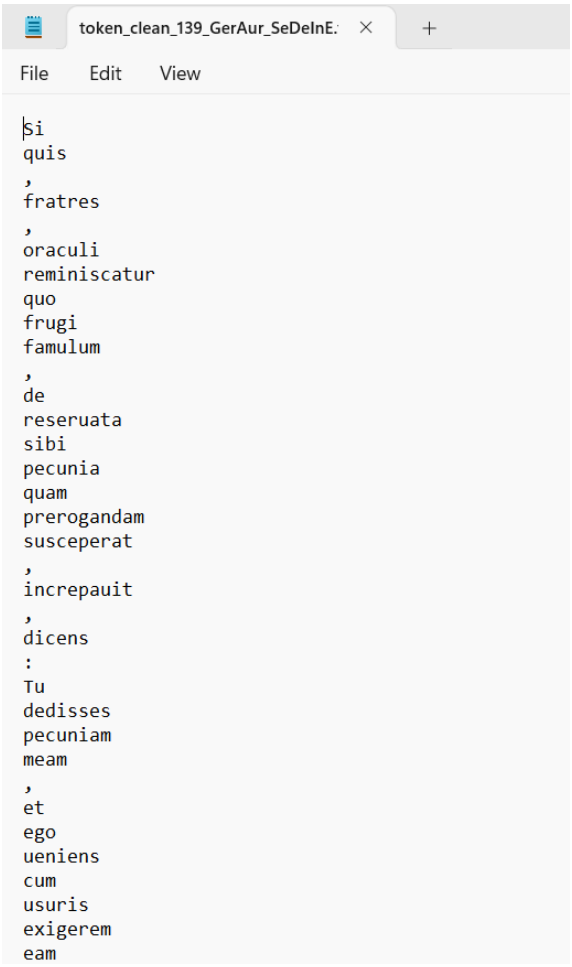


FIG. 4.4 : Le même texte après tokénisation

4.3 Réseaux de co-occurrence de mots

Le fichier `02_matriceFeatures.R` est consacré à la transformation des données textuelles, désormais pré-traitées, en réseaux de co-occurrence de mots et au calcul des mesures topologiques des réseaux ainsi créés.

La variable `K` permet de contrôler le nombre de mots par paquets. Nous avons créé des paquets de 100, 200, 500 mots et 1000 mots. Ensuite, un data frame est créé pour abriter les futures mesures. La majorité du fichier est occupée par une double boucle, qui parcourt l'ensemble des fichiers et, au sein de ceux-ci, l'intégralité des paquets à créer selon la variable indiquée en haut du fichier.

Chaque texte présent dans le dossier `dataLemmatisees` est divisé en quatre colonnes (correspondant au mot du texte, à la nature du mot ou POS, au lemme et à la traduction) et est transformé en data frame pour facilité d'usage. Ce data frame per-

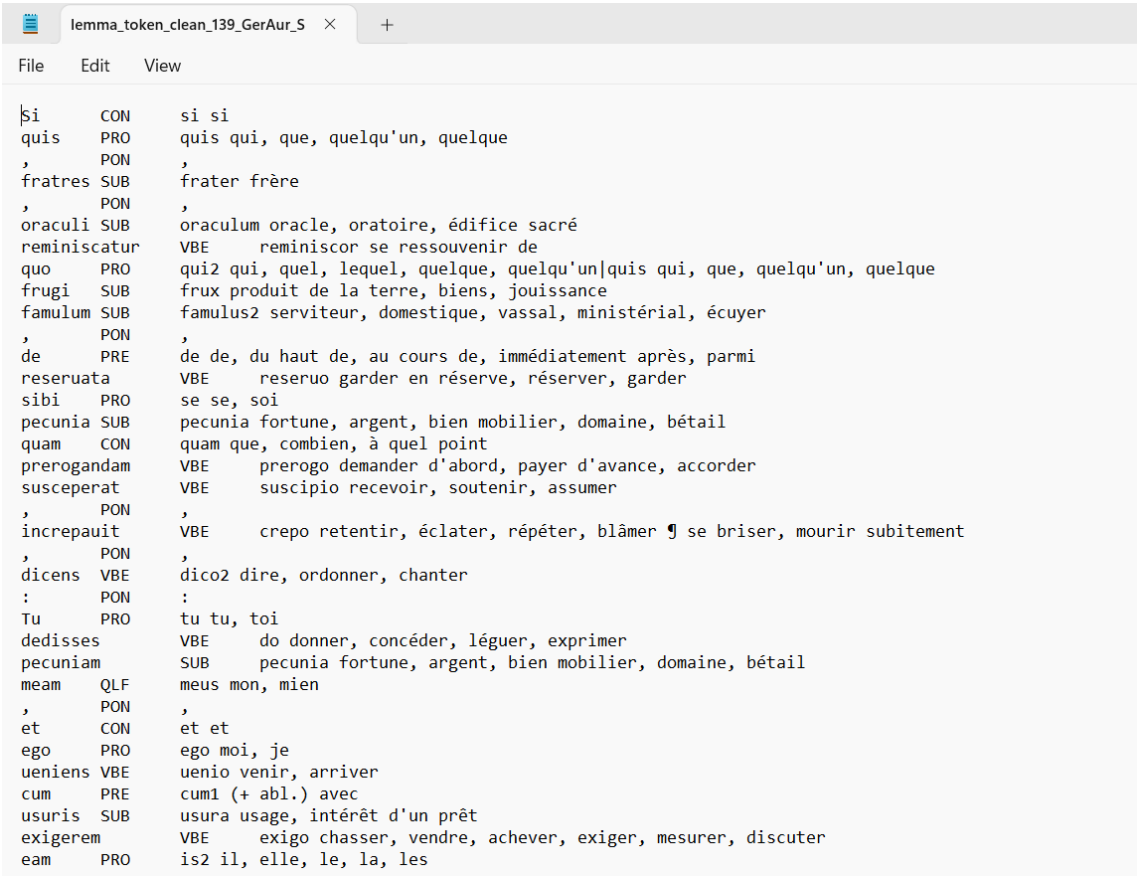


FIG. 4.5 : Le même texte après lemmatisation

met de trier facilement dans le texte lemmatisé pour ne garder que les POS qui nous intéressent. Nous avons envisagé deux cas de figure. Dans le premier cas, nous ne sommes intéressés que par des mots riches sémantiquement : nous supprimons dès lors tous les lemmes qui sont des mots-outils (les conjonctions, les interrogatifs, les pronoms). Dans le deuxième cas, nous ne sommes intéressés que par les mots-outils : dans ce cas-là, nous filtrons le data frame pour ne garder que les POS précédemment mentionnées. Ces deux alternatives reflètent les deux approches qui ont lieu dans les articles utilisant l'analyse de réseaux.[6] Notre objectif en créant cette double possibilité est de vérifier si une approche est plus efficace que l'autre.

Chaque texte est divisé en le plus grand nombre de paquets possible d'après la variable `K`. Chaque paquet comprend un nombre égal de mots. Si le texte est plus petite que cette variable, le reste du code ne s'enclenche pas, évitant une erreur assez fréquente quand les seuls lemmes considérés sont les mots-outils et que la variable est fixée à 500. Le nom de l'auteur et le titre de l'œuvre sont sauvegardées grâce à des expressions régulières : les variables ainsi peuplées seront utilisées lorsque chaque paquet du texte sera traité dans la seconde boucle. Avant d'arriver à cette dernière toutefois, nous créons un nouveau data frame de même structure que le précédent,

[6] GOH, LUKE et CHEONG, « Functional Shortcuts in Language Co-Occurrence Networks », p. 2-3.

mais dont la longueur n'est plus nulle mais de la taille du nombre de paquets.

La seconde boucle itère dans la liste de paquets, dont le numéro est sauvegardée dans le data frame. Les indices de début et de fin du paquet sont calculés à partir de la variable `K` et de l'incrémenteur de la boucle `i`. Ces indices permettent d'isoler les mots du paquet (à savoir les lemmes). Une simple modification du numéro qui indique la colonne des lemmes dans le texte lemmatisé permet de créer des réseaux de mots qui gardent leur flexion. La langue latine étant une langue fléchie, il nous semble pertinent de vouloir conserver au maximum les terminaisons des verbes, noms, adjectifs et pronoms. Une fois les mots voulus identifiés, il nous suffit de créer une liste d'arêtes pour le futur réseau, composées de tous les mots directement adjacents. Le réseau est créé sous le nom de `grPaquetsMots`, avec l'option `directed` puisque nous désirons des réseaux dirigés. Les mesures topologiques suivantes sont alors mesurées (systématiquement au niveau du réseau, en insérant une fonction `mean` pour généraliser les mesures locales) et sauvegardées dans le data frame :

- Le nombre de sommets : le nombre de lemmes uniques dans l'extrait ;
- Le diamètre : le plus grand chemin possible entre deux sommets (n'importe lesquels) ;
- Le rayon : cette mesure l'inverse du diamètre, puisqu'elle représente le plus petit chemin possible entre deux sommets (n'importe lesquels) ;
- Le degré sortant moyen : le nombre d'arêtes qui prennent leur origine dans un sommet, dont on fait la moyenne à l'échelle de tous les sommets du réseau ;
- La distribution des degrés moyenne : la probabilité que les autres sommets du graphe aient le même degré qu'un sommet, dont on fait la moyenne à l'échelle de tous les sommets du réseau ;
- Le degré pondéré moyen : le nombre d'arêtes entrantes ou sortantes pour un sommet, , dont on fait la moyenne à l'échelle de tous les sommets du réseau ;
- Le plus petit chemin moyen : le plus petit chemin possible entre deux sommets donnés, dont on fait la moyenne à l'échelle de tous les sommets du réseau ;
- L'intermédiarité moyenne : le nombre de plus petits chemins qui passe par un sommet donné, dont on fait la moyenne à l'échelle de tous les sommets du réseau ;
- Le coefficient de clustering : cette mesure correspond à la probabilité que, si deux sommets donnés sont liés, ils soient tous deux liés à un même troisième sommet ;

- Le coefficient d'assortativité : cette mesure correspond à la probabilité que les sommets possédant un degré similaire soient reliés entre eux ;
- Le nombre de motifs (de taille 3), c'est-à-dire de sous-réseaux avec une forme donnée ;
- Le nombre de cliques : le nombre de sous-réseaux où tous les sommets sont connectés les uns aux autres ;
- La taille de la plus grande clique : le nombre de sommets dans la plus grande clique du réseau.

Le nouveau data frame est alors ajouté à la matrice globale qui reprend toutes les mesures (la `matriceFeatures`). Pour terminer, un graphique représentant le nombre de paquets par auteur est généré. Voir la figure 4.6 qui représente les situations où seuls les mots avec une valeur sémantique sont considérés (pas de mots-outils) avec `K` qui vaut 100, 300, 500 et 1000. Un fichier .RData sauvegarde également la matrice de mesures.

Deux brèves lignes de code permettent également de générer le graphique d'un réseau donné, mais le résultat n'est pas très visuel. La figure 4.7 représente, à titre d'illustration, un réseau du plus petit `K` que nous avons envisagé (100).

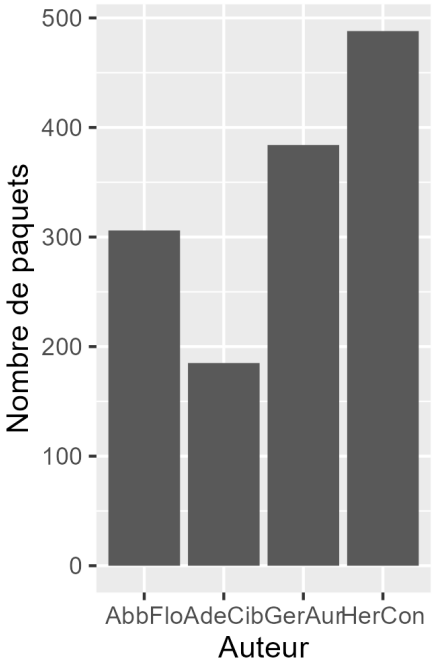
4.4 Application de méthodes d'apprentissage automatique

Deux fichiers comportement le préfixe «03» : `03_ACP` et `03_Classification`. Ils relèvent tous deux de la dernière étape de notre méthodologie : l'exploitation des mesures sauvegardées lors de l'analyse de réseaux au moyen de techniques provenant de l'apprentissage automatique.

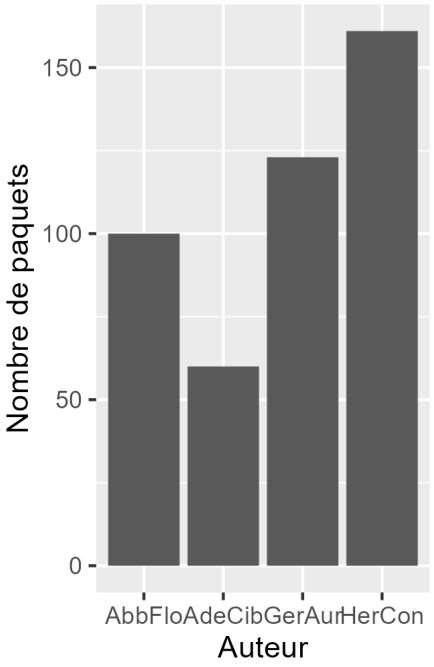
4.4.1 Visualisation

L'analyse en composantes principales (ACP) est une forme d'apprentissage automatique non supervisé. Cette technique consiste à résumer des données complexes (car hautement dimensionnelles) en une série de composantes principales. Souvent, les deux premières sont sélectionnées pour les visualiser sur un plan cartésien. L'intérêt de l'ACP réside non pas dans sa complexité, mais par son caractère intuitif et facile à comprendre. Toutes les informations sont directement visibles sous nos yeux, les données font l'explication à notre place.

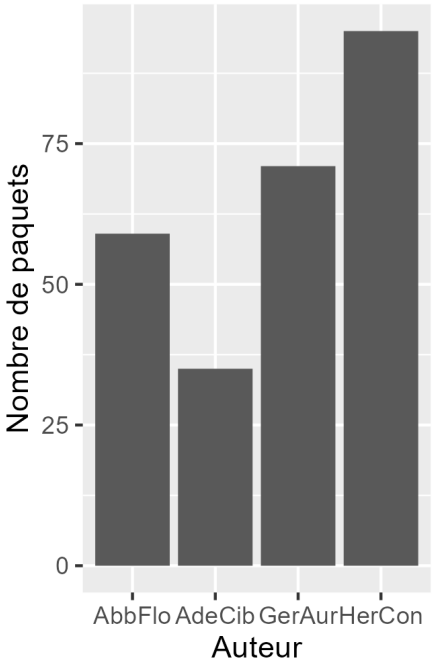
Avant de calculer les composantes principales, il est nécessaire de supprimer les textes identifiés comme douteux de la matrice, pour ne pas fausser les résultats. Une



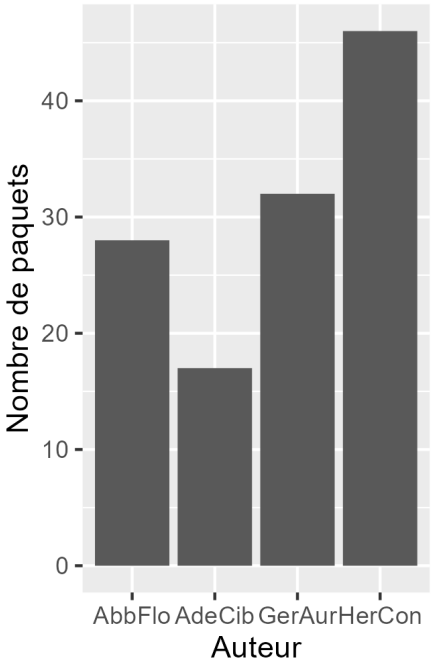
(a) K = 100



(b) K = 300



(c) K = 500



(d) K = 1000

FIG. 4.6 : Nombre de paquets par auteur selon K

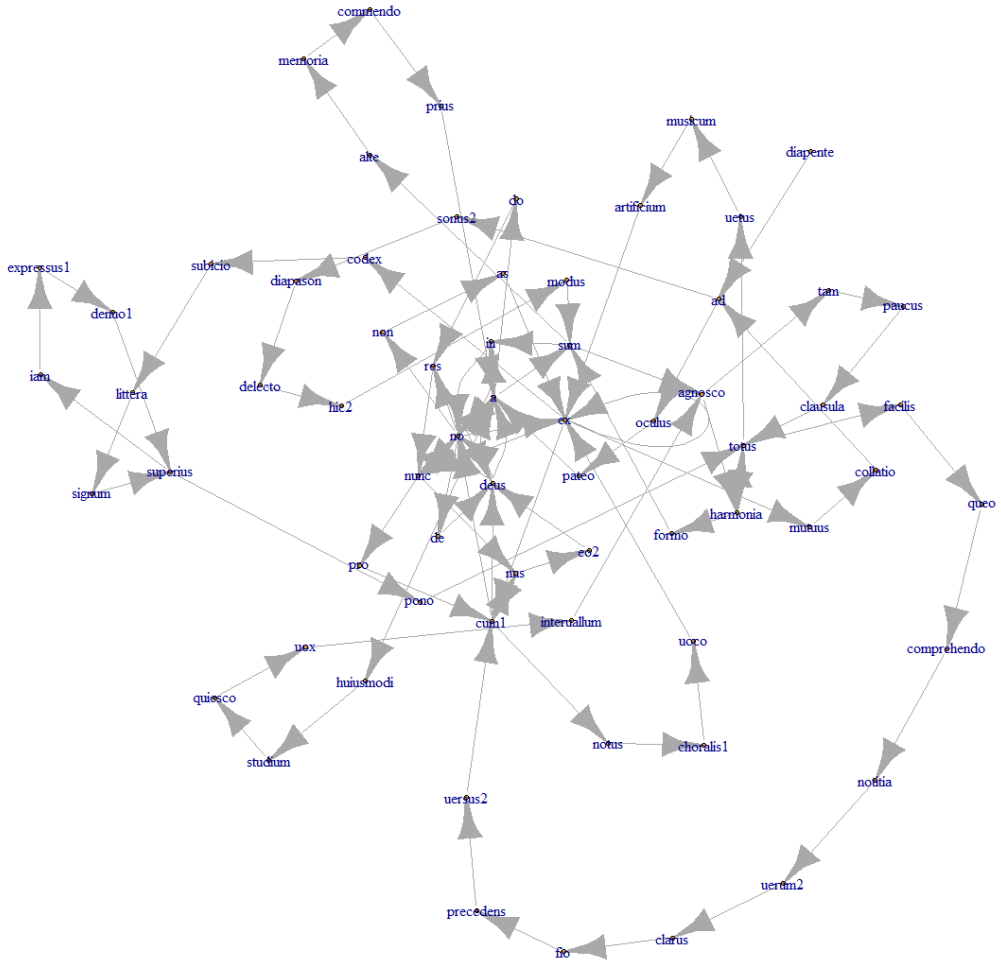


FIG. 4.7 : Un réseau de co-occurrence de mots (100 sommets)

fois l'ACP réalisée, il est possible de visualiser la proportion de variance représentée par chacune des composantes. Dans le cas de la figure 4.8, qui concerne une ACP qui résume les données de la matrice des mesures topologiques (réseaux de 500 sommets), nous pouvons lire que la première composante représente plus de 60% de la variance totale et la seconde un peu moins de 20%. Quand on visualise l'ACP sur un plan cartésien, nous voyons donc 80% de la variance originelle des données. Il est également possible de vérifier les mesures topologiques qui contribuent le plus à la variance. Pour la même matrice que la dernière figure, les mesures qui contribuent le plus aux composantes principales sont, pour les premières et dans l'ordre descendant, l'intermédiarité, le coefficient d'assortativité, le rayon, le diamètre et le coefficient de clustering.

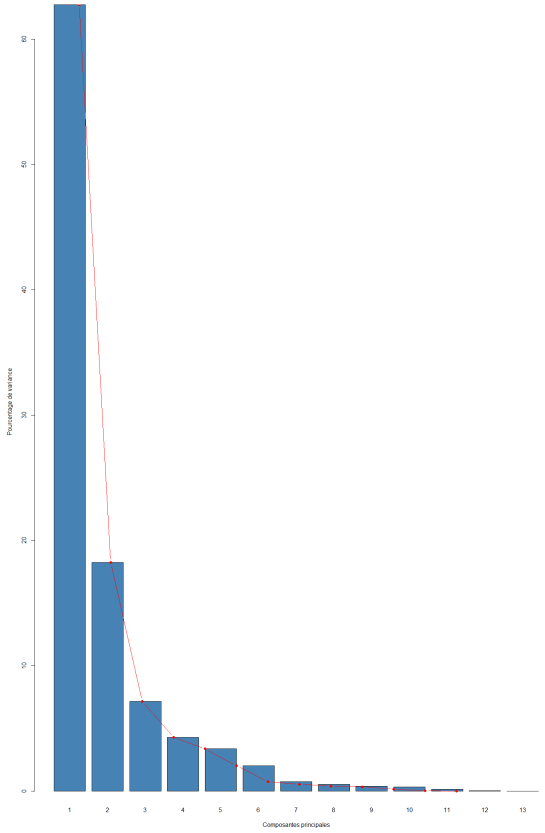
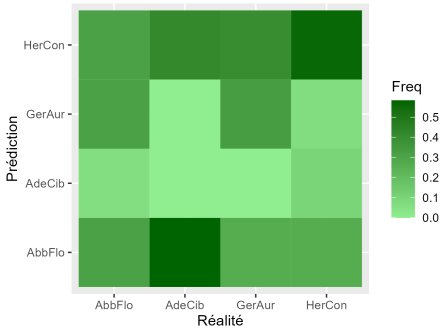
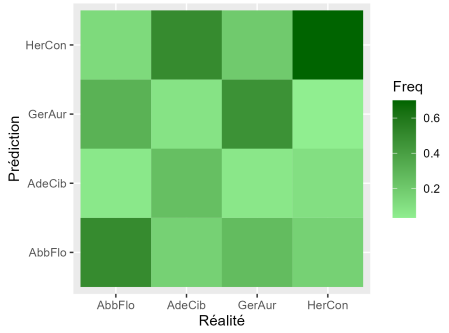


FIG. 4.8 : Pourcentage de la variance représenté par chaque composante principale

Pour la visualisation à proprement parler, il convient de séparer les données relatives aux deux œuvres douteuses afin de pouvoir leur faire ressortir par rapport aux autres points dans le graphique. En l'occurrence, elles obtiennent chacune une couleur différente et, parce que nous trouvions cela plus lisible, nous avons augmenté la taille de leurs points.



(a) KNN sans normalisation (taux d'erreur de 63%)



(b) KNN avec normalisation (taux d'erreur de 46%)

FIG. 4.9 : L'amélioration de KNN grâce à la normalisation

4.4.2 Classification

Bien que l'ACP soit une manière visuelle et intuitive d'explorer nos données, il est plus solide de faire appel à des techniques d'apprentissage automatique cette fois-ci supervisés. La littérature souligne qu'il est préférable d'utiliser des techniques qui peuvent gérer un grand nombre de caractéristiques stylistiques, c'est-à-dire donner automatiquement (sans intervention humaine) la priorité aux caractéristiques qui permettent d'atteindre la meilleure classification. KNN n'est pas une de ces techniques, tandis que SVM l'est.[7] Nous avons décidé de les implémenter toutes les deux. Elles reposent sur des principes similaires, et nous les expliquerons donc de front.

Les mesures contenues dans la matrice sont normalisées : la différence entre la valeur de la mesure et la moyenne de cette mesure, divisée par l'écart-type de celle-ci. Cette transformation permet d'obtenir des valeurs qui rendent compte de leurs relations les unes avec les autres, et non une valeur absolue. Cette normalisation est en particulier utile pour KNN : SVM est capable de sélectionner les mesures les plus pertinentes pour la classification, alors que KNN ne l'est pas. Comme l'illustrent les matrices de confusion à la figure 4.9, la normalisation permet de diminuer le taux d'erreur de KNN de 63% (à peine plus bas que le hasard) à 46%.

Toute méthode supervisée nécessite deux corpus : un corpus d'entraînement et un corpus de test. Un index est créé de façon aléatoire (mais reproductible grâce à `set.seed(1)` : celui-ci marque un tiers de nos données, qui iront dans le corpus de test. Le corpus d'entraînement est donc créé à partir de la matrice de mesures topologiques, à l'exception de la colonne concernée par la classification (dans notre cas la colonne `auteur`) et d'éventuelles colonnes à retirer (les colonnes `titre` et `paquet`). La colonne à classifier est enregistrée quant à elle dans `training.class`.

La méthode KNN (*K-nearest neighbors*) repose sur un paramètre que nous pouvons déterminer subjectivement, mais il est également possible de vérifier par une

[7] JUOLA, « Authorship Attribution », p. 300.

boucle rapide quelle valeur donne le meilleur résultat de classification. Le retour de KNN est une matrice de confusion (qu'il faut encore pondérer), que nous visualisons grâce à `ggplot`. Nous calculons également le taux d'erreur en déduisant le total des prédictions correctes de l'ensemble des prédictions, et en divisant cette différence par le total des prédictions. Un sans-faute obtiendra donc la valeur de 0 et, plus la valeur monte vers 1, plus le modèle de prédiction est imparfait. Pour que KNN classifie les écrits douteux que nous avions mis de côté au début du fichier, il suffit de remplacer le corpus de test par les données du texte douteux. La visualisation des résultats peut désormais se réaliser par un simple graphique en barres, dont l'échelle compte les attributions faites à un auteur.

La méthode SVM (*support vector machine*) repose elle aussi sur la création d'un modèle d'entraînement. Le retour de SVM est lui aussi une matrice de confusion. L'attribution des textes douteux se réalise par une fonction `predict`, dont le fonctionnement est sensiblement identique à KNN, et le rendu est également réalisé au travers d'un graphique en barres.

Puisque l'ensemble du corpus ne produisait pas vraiment de résultats satisfaisants, nous avons également subdivisé la matrice de mesures topologiques en deux sous-groupes qui représentent un genre littéraire homogène : d'une part la correspondance avec les lettres de Gerbert et celles d'Abbon de Fleury, d'autre part le genre historique avec l'*Excerptum de vitis Romanorum pontificum* d'Abbon, les *Historiae* d'Adhémar de Chabannes et le *Chronicon* d'Hermann Contract.

Chapitre 5

Résultats

Dans ce chapitre, nous mettrons en pratique la méthodologie explicitée dans le chapitre précédent, en particulier la dernière étape de visualisation et de classification des mesures topologiques obtenues lors de la transformation en réseaux des paquets textuels. Nous testerons la robustesse du modèle basé sur l'analyse des réseaux linguistiques en considérant l'ensemble de nos données. Ensuite, nous aborderons trois cas spécifiques. Le premier sera de vérifier les résultats obtenus par des œuvres dont la paternité ne pose aucun doute, que ce soit pour Gerbert, Abbon, Hermann ou Adhémar. Le deuxième est celui du *Sermo de informatione episcoporum* et le troisième du *De utilitatibus astrolabii*, déjà longuement évoqué au sein des précédentes pages de ce mémoire.

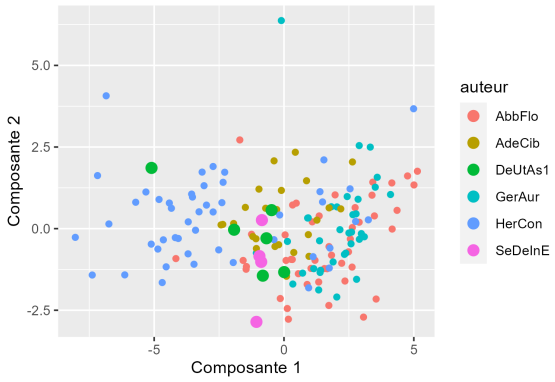
5.1 Robustesse du modèle

Avant d'appliquer notre modèle à des cas particuliers, il convient de l'évaluer de façon globale. Pour ce faire, nous commencerons par nous faire une première idée des résultats en les visualisant sous la forme d'analyses en composantes principales.

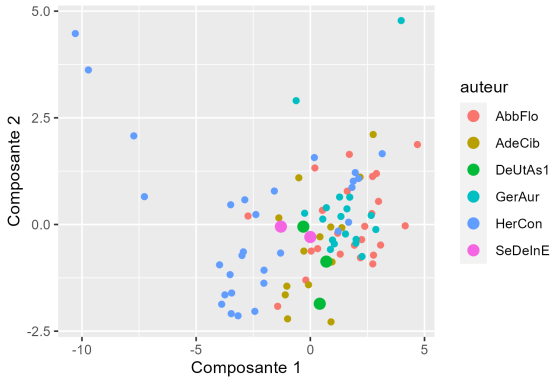
5.1.1 Visualisation avec des ACP

Pour commencer à explorer les données, nous les avons visualisées au moyen d'une ACP. Nous avions créé plusieurs types de matrices : des matrices avec des paquets de 100, 300, 500, 1000 mots ; des matrices dont les données textuelles n'avaient pas été lemmatisées ; des matrices ne comprenant que des *function words*.

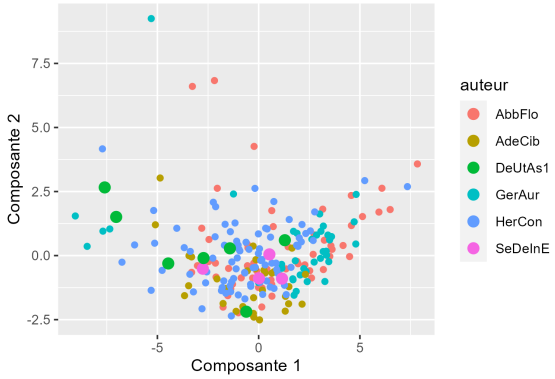
Nous avons commencé par les matrices des données lemmatisées. La visualisation ne permettait pas d'y repérer quelque chose de particulier (voir la figure 5.1). Dans la version avec les paquets de 300 mots, les points nous semblaient un peu plus distingués, sans que cela soit flagrant.



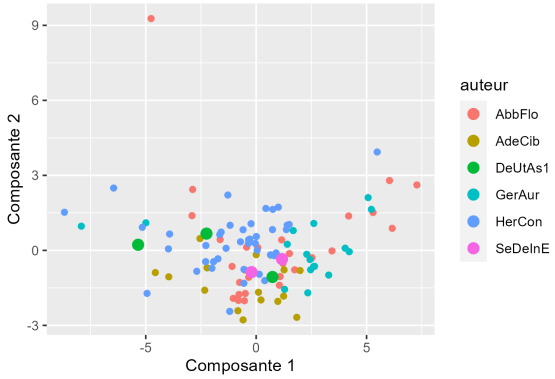
(a) K = 100



(b) K = 300



(c) K = 500



(d) K = 1000

FIG. 5.1 : ACP des données lemmatisées

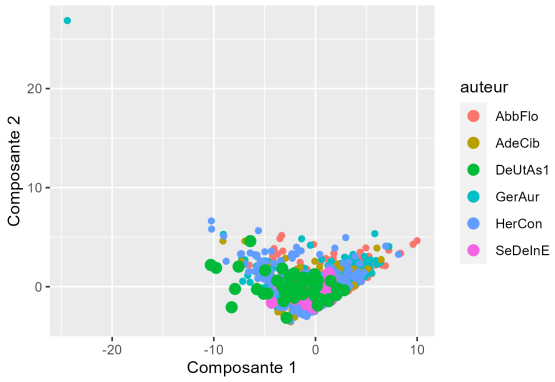
Nous avons reproduit la même expérience sur les données non lemmatisées (voir figure 5.2). Nous n'avons absolument rien pu tirer de ces quatre ACP : tous les points semblent les uns sur les autres. Contrairement à ce qu'un aspect philologique aurait pu me laisser penser, la flexion ne semble pas apporter plus de caractéristiques stylistiques dans les mesures topologiques, au contraire !

Il ne nous reste que les ACP des données ne regroupant que les mots-outils (voir la figure 5.3). Celles-ci n'existent qu'en trois versions : vu que les mots-outils sont plus rares que l'ensemble des mots, faire des paquets de taille `K` = 1000 risquait de faire disparaître intégralement des textes de nos mesures topologiques. Par exemple, le *Sermo de informatione episcoporum* que nous voulons attribuer. Le voir disparaître de la matrice serait contre-productif. Après lecture de ces graphiques, nous avons l'impression que ce sont ces réseaux qui permettent le plus de distinguer des groupes de points, en particulier lorsque `K` = 300 ou 500.

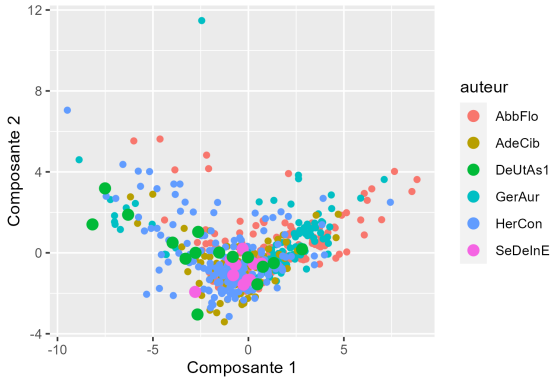
Que conclure de ces explorations ? Premièrement, la méthode mise en œuvre ne semble pas très efficace pour distinguer plusieurs profils stylistiques, que ce soit au niveau des quatre auteurs ou des deux écrits disputés (repérables, pour rappel, par leur couleur différente et leur taille légèrement plus grande). Deuxièmement, il ne s'agit pas d'un trop grand manque de la variance originelle dans les composantes principales : la première composante contient systématiquement entre 50 et 60% de la variance, et la seconde entre 10 et 20% (en l'occurrence, l'ensemble des graphiques générés ressemble à celui reproduit dans le chapitre 4, figure 4.8). Troisièmement, ce sont systématiquement les mêmes mesures qui arrivent en tête de contribution à la première composante principale : le nombre de sommets, le plus petit chemin moyen, le nombre de cliques, le diamètre et dans une moindre mesure la distribution du degré ; tandis que, pour la deuxième composante, les mesures d'intermédiarité et de coefficient d'assortativité sont les plus grand contributeurs. La coefficient de clustering, pourtant souligné dans la littérature comme étant une des mesures principales par son efficacité, n'est que loin dans le classement.

5.1.2 Application des algorithmes de classement

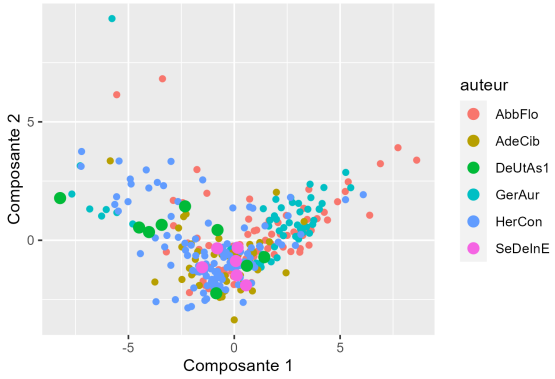
Les observations effectuées grâce aux ACP ne sont pas très encourageantes, mais restent subjectives car très visuelles. Pour avoir des résultats plus objectifs et concrets, les méthodes supervisées SVM et KNN sont particulièrement bien placées. Elles génèrent toutes deux une matrice de confusion et un taux d'erreur moyen. Par conséquent, elles permettent de quantifier la meilleure façon de capturer le style dans un réseau linguistique : formes lemmatisées, non lemmatisées, uniquement mots-outils ; 100, 300, 500 ou 1000 mots par paquet. Une fois que les performances de chaque modèle sera identifiée, nous pourrons en sélectionner un plus petit nombre pour ef-



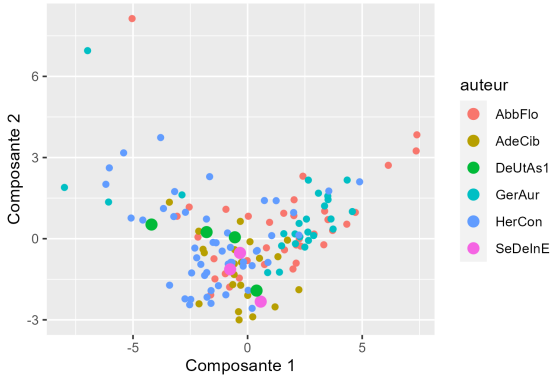
(a) K = 100



(b) K = 300

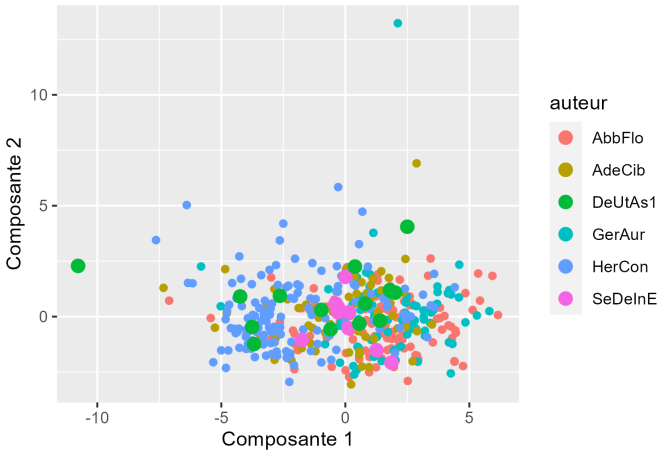


(c) K = 500

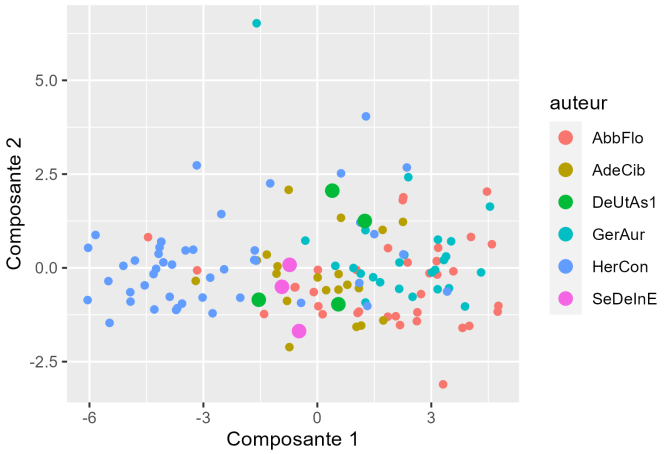


(d) K = 1000

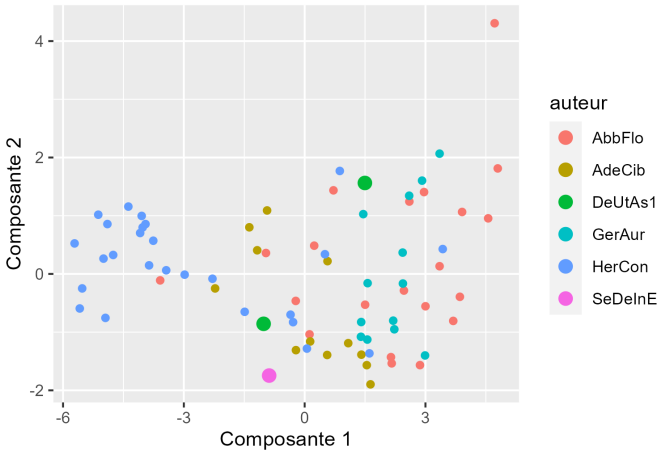
FIG. 5.2 : ACP des données non lemmatisées



(a) K = 100

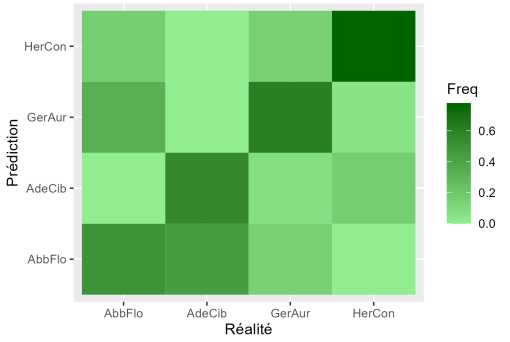


(b) K = 300

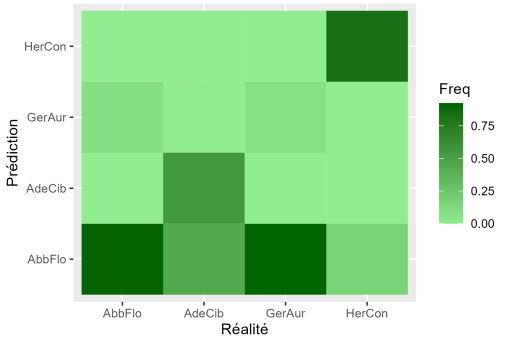


(c) K = 500

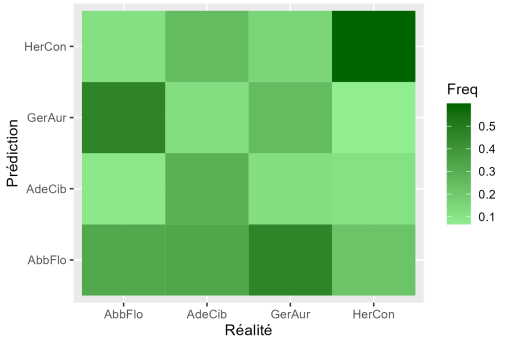
FIG. 5.3 : ACP des données ne regroupant que les mots-outils



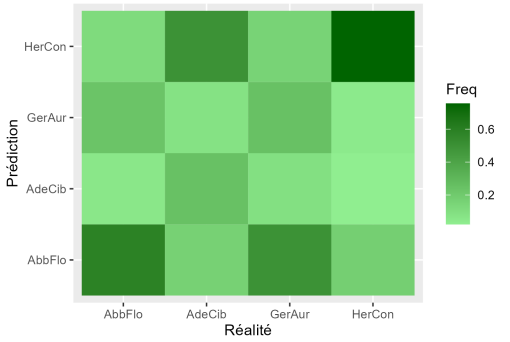
(a) KNN (L) – 0.33



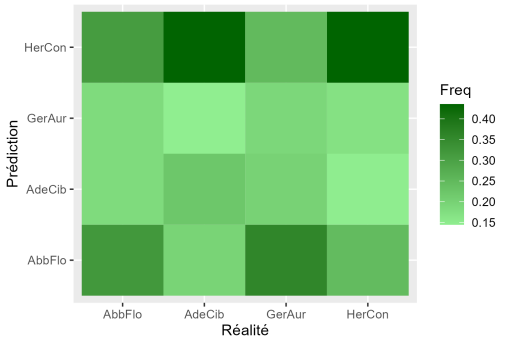
(b) SVM (L) – 0.38



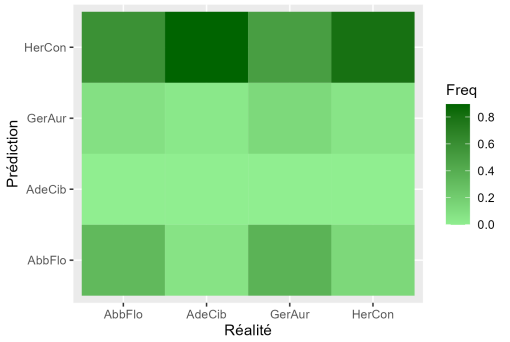
(c) KNN (FW) – 0.6



(d) SVM (FW) – 0.49



(e) KNN (NL) – 0.6

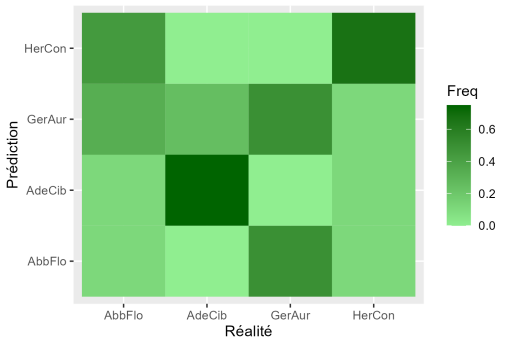


(f) SVM (NL) – 0.59

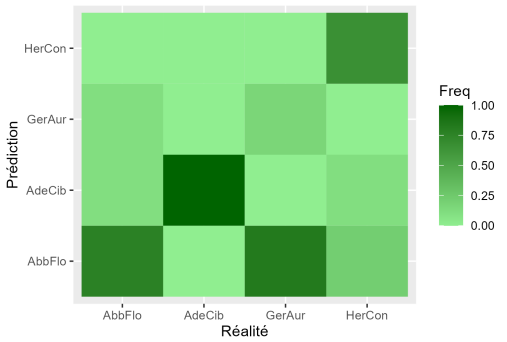
FIG. 5.4 : Matrices de confusion pour KNN et SVM sur des paquets de 100 mots

fectuer nos prochaines analyses sur les cas particuliers. Le titre de chaque matrice de confusion (voir figures 5.4, 5.5, 5.6 et 5.7) doit être lu de la sorte : type de prétraitement (L = lemmatisé, NL = non lemmatisé, FW = mots-outils parce que ***function words***) – taux d'erreur global (de 0 à 1 : plus le chiffre est élevé, plus le taux d'erreur est important).

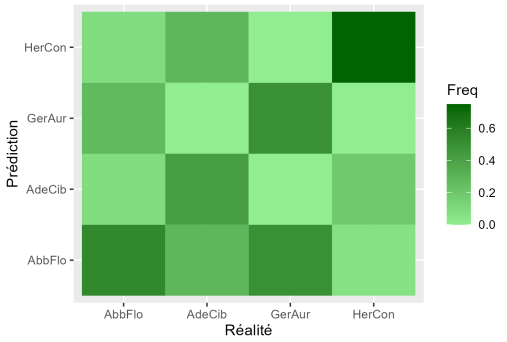
Quelles conclusions tirer des ces nombreuses matrices de confusion ? Premièrement, comme le laissait suggérer l'ACP, les réseaux de mots non lemmatisés ne sont vraiment pas efficaces. Leur fiabilité augmente avec le nombre de mots par paquet (taux d'erreur de 0.44 pour 1000 mots), tout en restant décevante. Deuxièmement, que tous les taux d'erreur restent supérieurs à l'aléatoire (qui se situerait autour de 0.75, comme il y a quatre auteurs). Troisièmement, qu'Hermann Contract semble



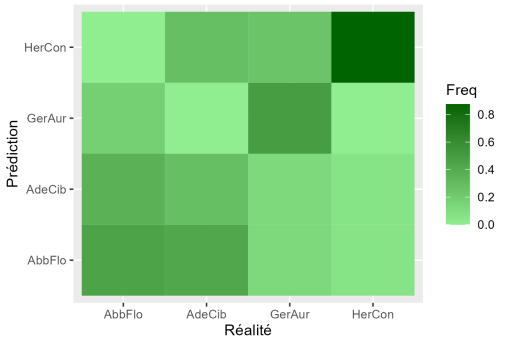
(a) KNN (L) – 0.59



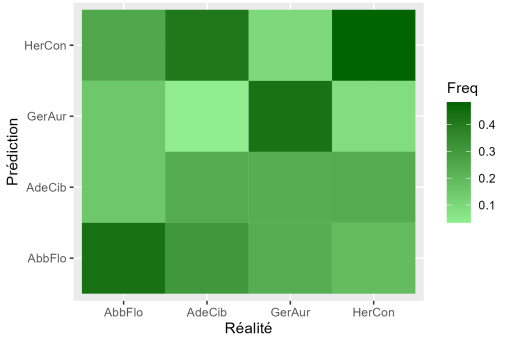
(b) SVM (L) – 0.38



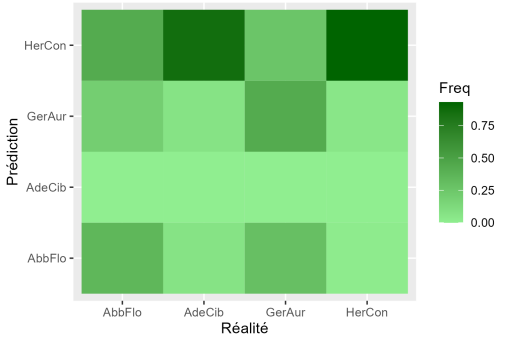
(c) KNN (FW) – 0.4



(d) SVM (FW) – 0.4

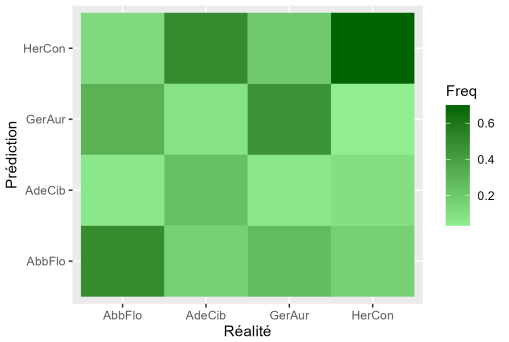


(e) KNN (NL) – 0.58

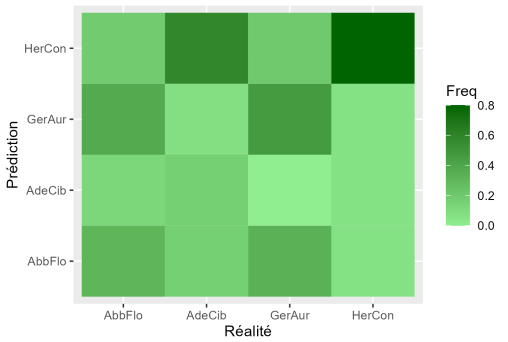


(f) SVM (NL) – 0.48

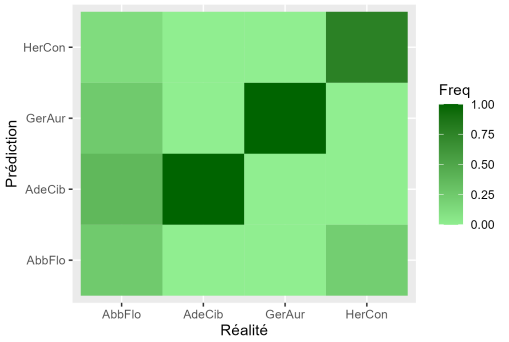
FIG. 5.5 : Matrices de confusion pour KNN et SVM sur des paquets de 300 mots



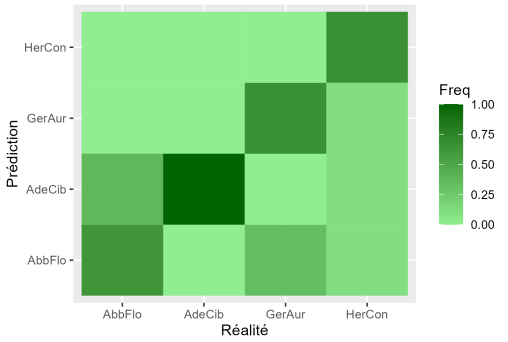
(a) KNN (L) – 0.46



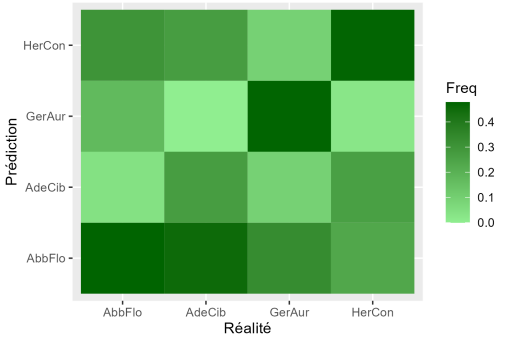
(b) SVM (L) – 0.47



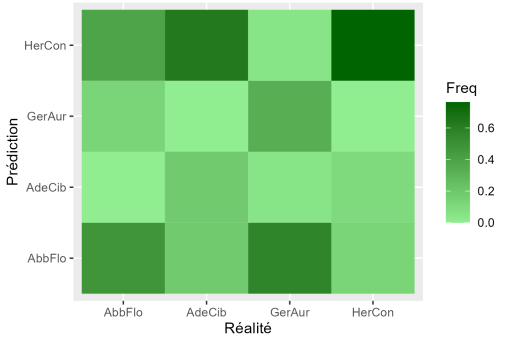
(c) KNN (FW) – 0.33



(d) SVM (FW) – 0.29

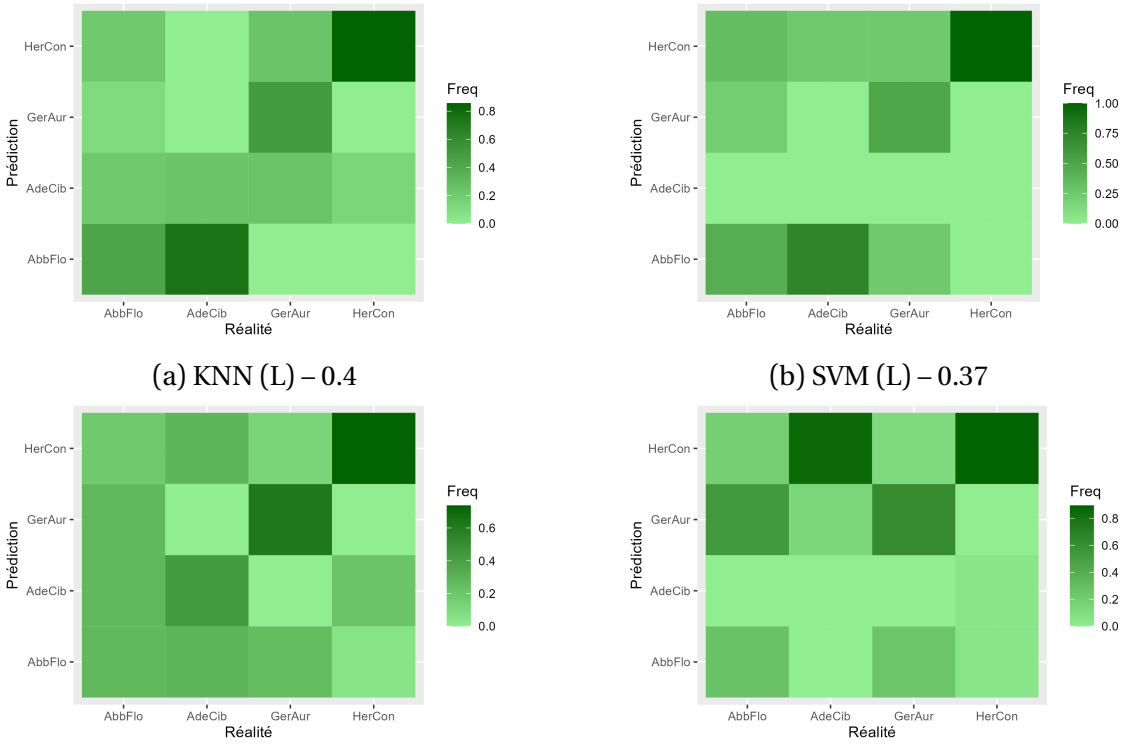


(e) KNN (NL) – 0.54



(f) SVM (NL) – 0.47

FIG. 5.6 : Matrices de confusion pour KNN et SVM sur des paquets de 500 mots



(a) KNN (L) – 0.4

(b) SVM (L) – 0.37

(c) KNN (NL) – 0.44

(d) SVM (NL) – 0.44

FIG. 5.7 : Matrices de confusion pour KNN et SVM sur des paquets de 1000 mots

avoir un style bien distinct, tandis qu'Adhémar de Chabannes, Abbon de Fleury et notre Gerbert ont plus rapidement tendance à se confondre les uns avec les autres. Faudrait-il y voir une influence de sa datation, plus tardive que les trois autres qui sont pour ainsi dire contemporains ? Il serait trop risqué de le postuler, mais nous ne pouvions pas ne pas l'évoquer. Quatrièmement, l'efficacité des réseaux de mots lemmatisés diminuent avec le nombre de mots par paquets, tandis que celle des réseaux de mots-outils augmentent avec ce nombre de mots.

Pour la suite des analyses, nous prenons la décision totalement arbitraire de sélectionner les trois meilleures modélisations, dans l'ordre descendant d'efficacité : des réseaux de mots-outils par paquets de 500 et classés par SVM (0.29), des réseaux de mots-outils par paquets de 500 et classés par KNN (0.33), des réseaux de mots lemmatisés par paquets de 100 et classés par KNN (0.33). S'il faut traduire ces chiffres comme les taux de réussite sont généralement présentés dans la littérature scientifique, il s'agit respectivement de 71%, 67% et 67%. Bien loin des 80 à 90% souvent atteints dans la littérature, mais bien au-dessus de l'aléatoire et côtoyant d'autres méthodes moins efficaces que celles évoquées à l'instant.[1]

[1]Voir notamment AMANCIO et al., « Comparing Intermittency and Network Measurements of Words and Their Dependence on Authorship » ; MARINHO, HIRST et AMANCIO, « Authorship Attribution via Network Motifs Identification », qui revendiquent respectivement 65% et 57% d'efficacité.

5.2 Des œuvres qui ne font pas de doute

Avant d'appliquer nos modèles aux œuvres disputées qui sont le cœur de notre recherche, il est pertinent de les appliquer à des écrits pour lesquels le doute est invraisemblable. Ainsi, nous pourrons juger si les résultats obtenus peuvent faire l'objet d'une interprétation historique ou s'il est préférable de s'abstenir avant de créer un nouveau modèle plus efficace.

Dans un premier temps, nous nous intéresserons à la correspondance de deux rivaux politiques et intellectuels, Abbon et Gerbert, dont les corpus épistolaires sont à peu près égaux. Ensuite, nous traiterons de genre historique avec le *Chronicon* d'Hermann Contract, les *Historiae* d'Adhémar de Chabannes et l'*Excerptum de vitis Romanorum pontificum* d'Abbon de Fleury. Pour une fois, Gerbert restera en arrière-plan.

5.2.1 Les correspondances d'Abbon de Fleury et de Gerbert d'Aurillac

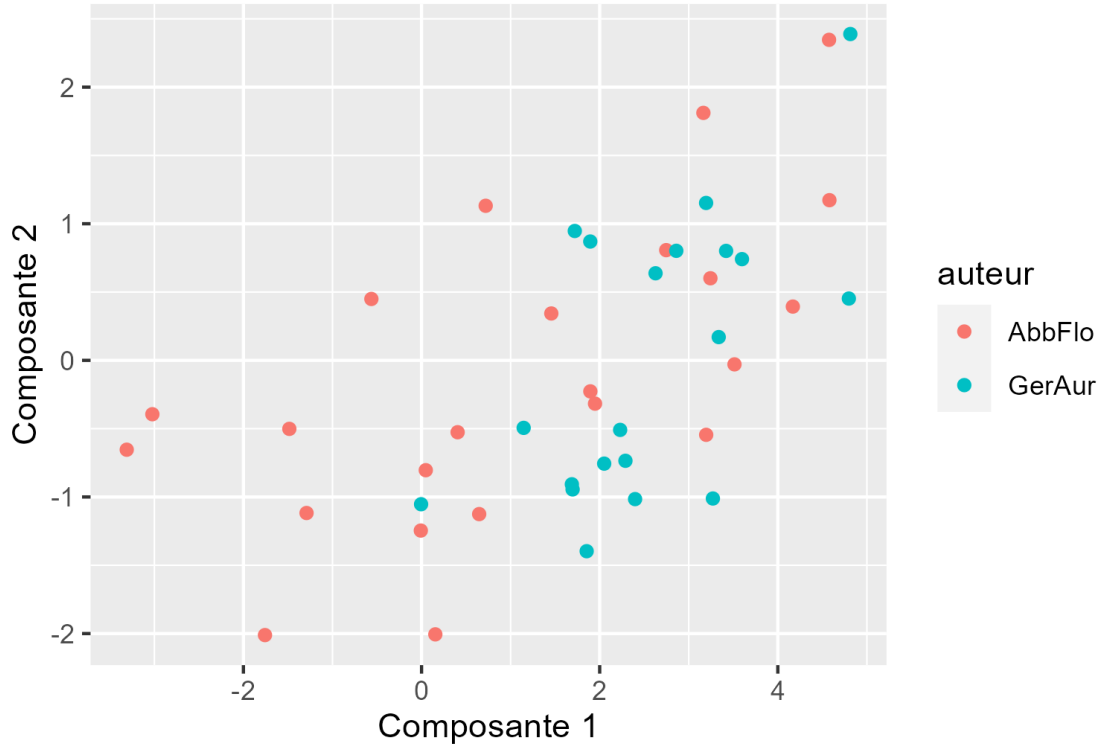
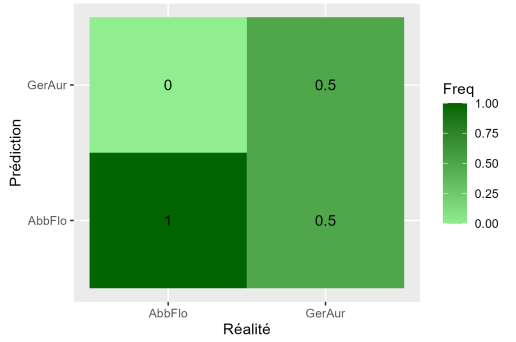
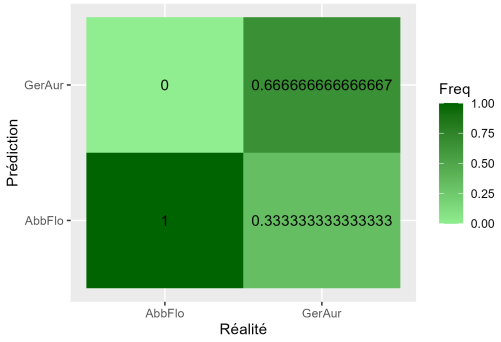


FIG. 5.8 : ACP des correspondances de Gerbert et d'Abbon

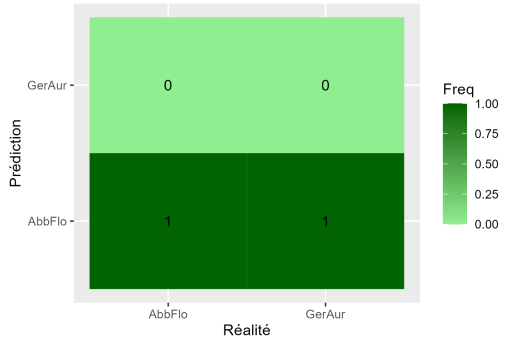
Commençons, comme nous en avons désormais l'habitude, par observer l'ACP de ces deux œuvres épistolaires (figure 5.8). Une fois n'est pas coutume, nous trouvons le graphique informatif. Même si certains points restent mélangés (au milieu à droite), les points d'Abbon semblent être en grande partie dans le quart en bas à



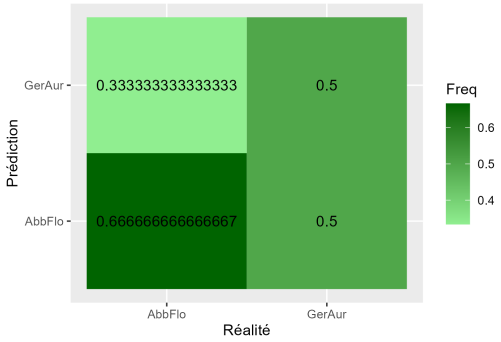
(a) KNN (L) 100 – 0.25



(b) SVM (L) 100 – 0.16



(c) KNN (FW) 500 – 0.4



(d) SVM (FW) 500 – 0.4

FIG. 5.9 : Matrices de confusion pour les correspondances de Gerbert et d'Abbon

gauche du graphique, tandis que ceux de Gerbert se trouvent presque exclusivement dans la moitié de droite.

Les résultats sont moins bons que ne le laissait présager l'ACP (voir les matrices de confusion à la figure 5.9). Pour les réseaux de mots-outils, qui étaient les plus performants à la section précédente, nous obtenons un résultat à peine plus haut que l'aléatoire, qui se situe à 50%. Les réseaux de mots lemmatisés par paquets de 100 mots avec KNN obtiennent 75% d'efficacité, ce qui n'est pas mal. Et, le plus étonnant : c'est la classification de ces mêmes réseaux par SVM, lancée par mégarde en même temps que KNN, qui obtient le meilleur score : 84% ! Ce n'est pas une surprise si importante, étant donné que ce modèle présentait déjà 62% d'efficacité dans la section 5.1. Toutefois, une telle variation dans les résultats ne me rassure pas quant à la fiabilité du modèle dans son ensemble.

5.2.2 Les écrits historiques d'Abbon de Fleury, Hermann Contract et Adhémar de Chabannes

L'ACP produite pour les trois écrits historiques (voir 5.10) est illisible, tous les points se chevauchent à l'exception de quelques points pour chaque auteur à gauche, en haut et à droite du graphique respectivement. Nous ne voyons pas ce que nous pourrions tirer comme informations supplémentaires de cet ACP, mais nous avons appris

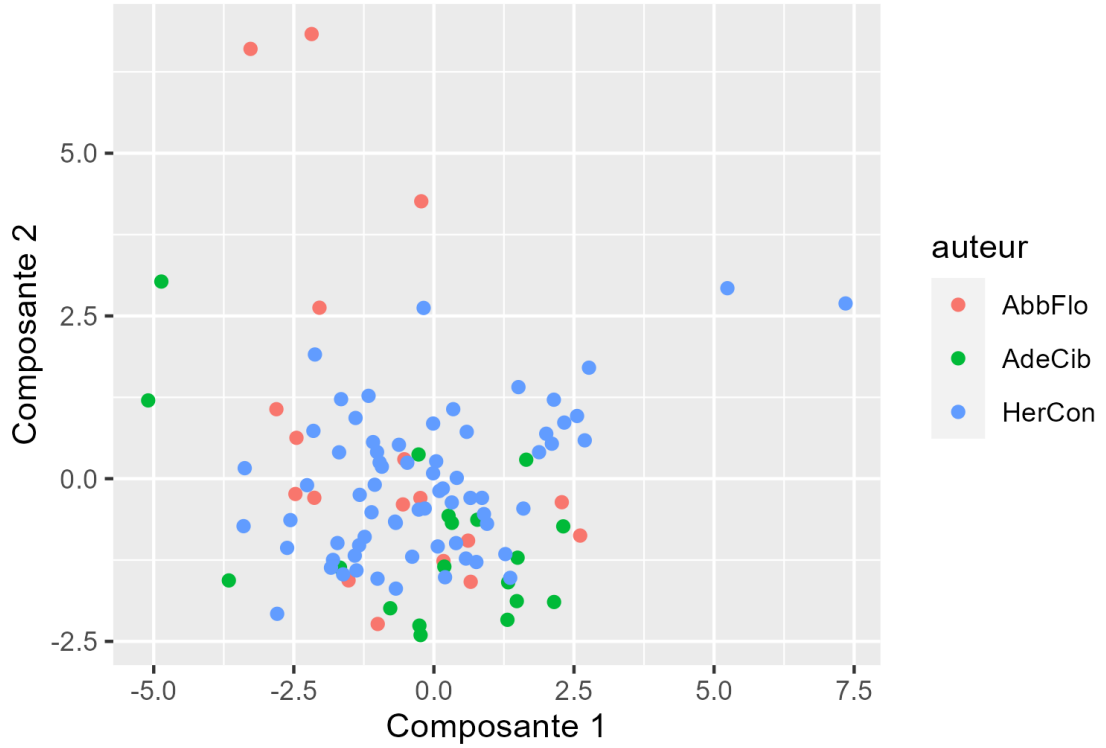


FIG. 5.10 : ACP des écrits historiques d'Abbon, d'Hermann et d'Adhémar

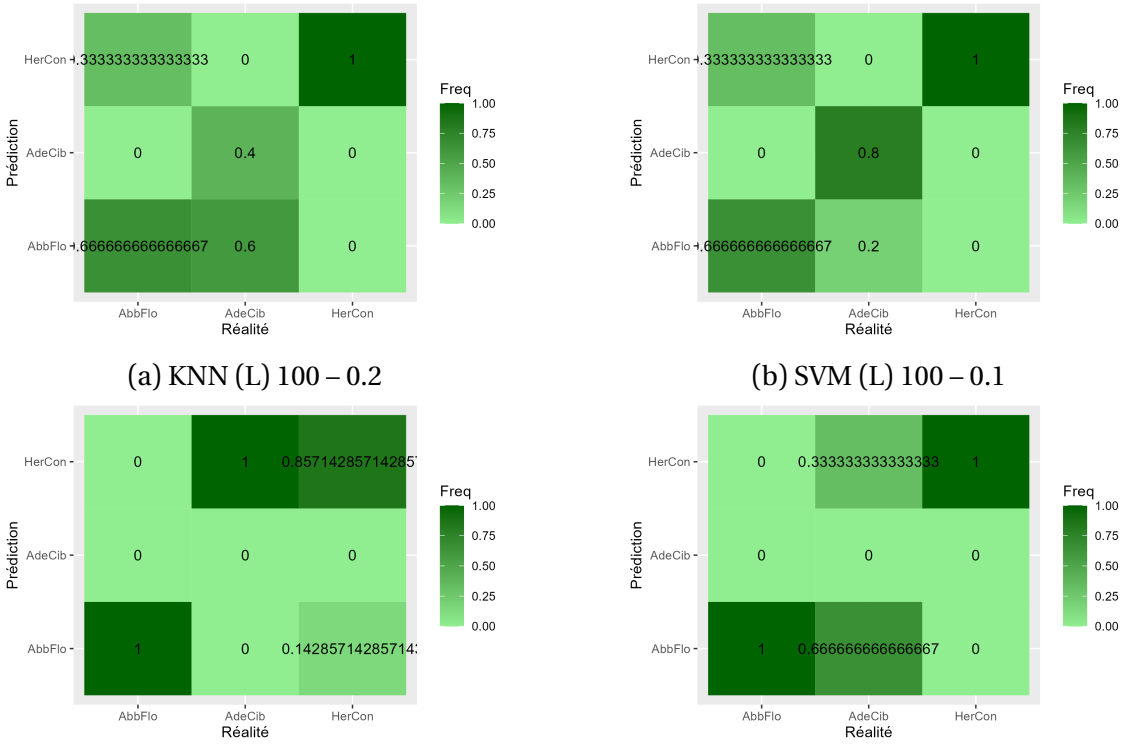
à nous méfier de l'allure arborée par l'ACP lors des résultats de la section précédente.

Les réseaux de mots lemmatisés (100 mots) confirment leur efficacité : KNN est efficace à 80% (avec quelques difficultés pour distinguer Adhémar et Abbon), tandis que SVM l'est à 90%. Heureusement que nous l'avions intégré par mégarde lors de l'analyse précédente, puisque ce modèle continue à être très performant. Le modèle basé sur les mots-outils, qu'il soit classé par KNN (64% d'efficacité) ou par SVM (73% d'efficacité), reste correct tout en étant décevant par rapport à ce qu'il avait montré lors de l'analyse globale. Adhémar a des difficultés à y être reconnu : il est dans un cas entièrement attribué à Hermann (KNN), dans l'autre réparti entre Abbon et Hermann (SVN).

5.3 Des œuvres dont la paternité est contestée

Nous voici arrivés à l'objectif premier de ce mémoire : l'analyse stylométrique va-t-elle être capable de résoudre les débats sur la paternité de ces deux œuvres disputées ? Rappelons que nous avons déjà pu nous rendre compte que l'ACP ne nous était d'aucun recours pour essayer de rattacher l'un ou l'autre texte à un auteur de notre corpus (voir les figures 5.1, 5.2 et 5.3).

Le résultat des prédictions KNN et SVM ne sera plus une matrice de confusion, étant donné qu'on ne considère plus qu'une seule ligne : qui est l'auteur de tel écrit ?



(a) KNN (L) 100 – 0.2

(b) SVM (L) 100 – 0.1

(c) KNN (FW) 500 – 0.36

(d) SVM (FW) 500 – 0.27

FIG. 5.11 : Matrices de confusion pour les écrits historiques d'Adhémar, d'Abbon et d'Hermann

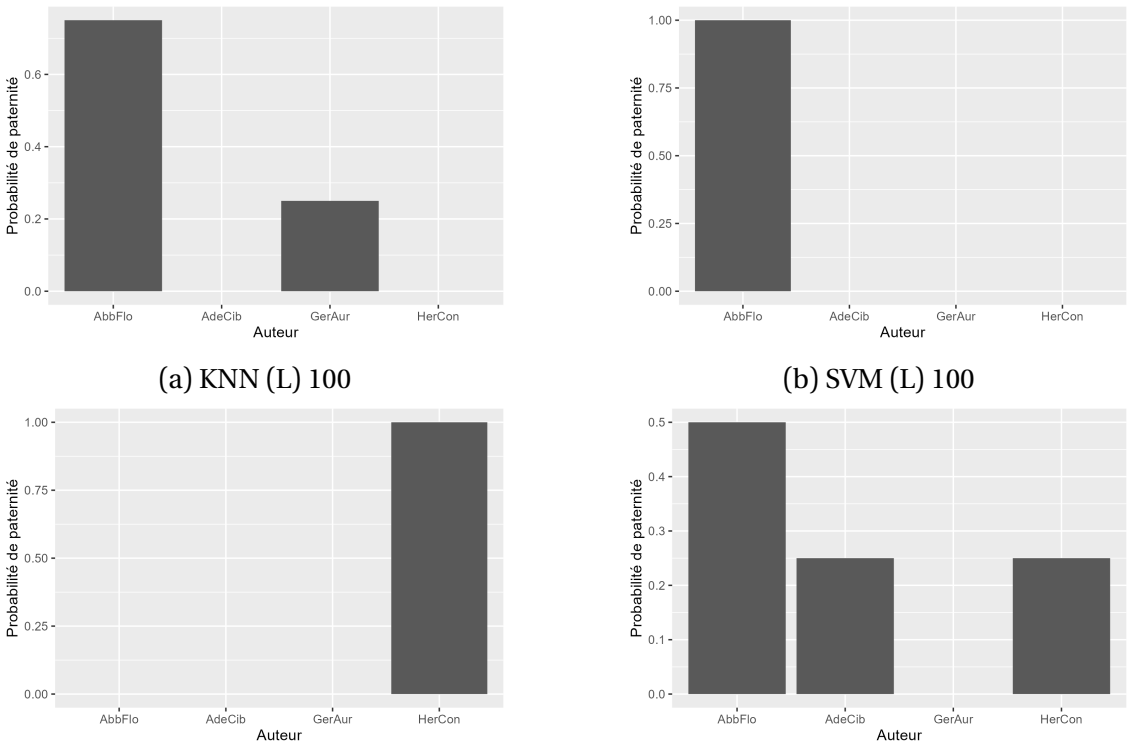
Nous avons donc utilisé des graphiques en barres pour représenter la probabilité qu'un auteur puisse revendiquer la paternité. Attention, ce n'est pas une probabilité absolue mais une probabilité relative, par rapport aux trois autres auteurs du corpus.

5.3.1 Le *Sermo de informatione episcoporum*

Rappelons, avant tout commentaire des graphiques (figure 5.12), que la paternité du *Sermo de informatione episcoporum* est bien connue, comme l'a prouvé F. Nuvolone dans un long article lors du premier colloque organisé à Bobbio : c'est Adhémar de Chabannes.[2] Nous avons décidé de l'intégrer dans une section dédiée à des oeuvres disputées car elle a été attribuée à Gerbert mais n'est sans nul doute raisonnable possible pas de sa main.

Nous pouvons en effet remarquer qu'il n'y a qu'un seul modèle qui l'attribue à Gerbert, avec une probabilité de 25%. C'est là que s'arrête les observations bienvenues. En effet, le *Sermo* est attribué à Adhémar, mais lui aussi à raison d'un seul modèle et pour une probabilité de 25%. Selon toute logique, et en regardant en particulier les deux premiers modèles qui se sont avérés les plus fiables jusqu'à maintenant, nous aurions tendance à attribuer le *Sermo* à Abbon de Fleury et non à son véritable auteur.

[2]NUVOLONE, « Gerberto ».



(a) KNN (L) 100

(b) SVM (L) 100

(c) KNN (FW) 500

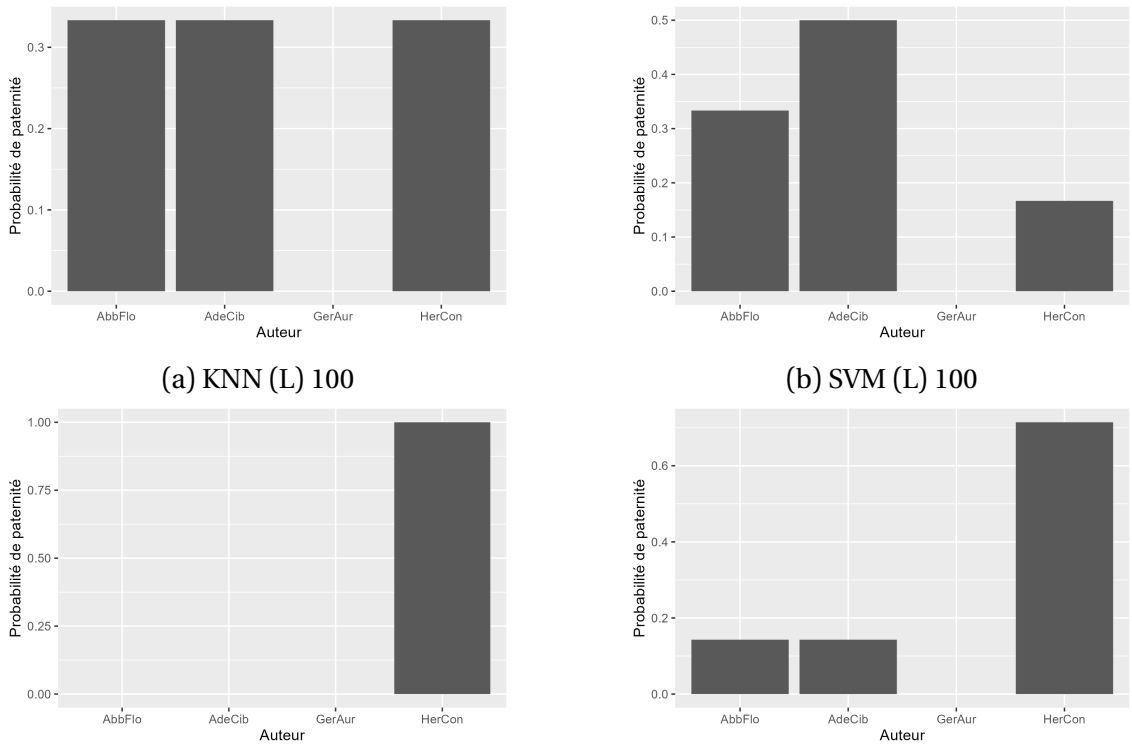
(d) SVM (FW) 500

FIG. 5.12 : Probabilités d'attribution du *Sermo de informatione episcoporum*

Nous avions déjà pu constater précédemment qu'Adhémar était souvent confondu avec Abbon, ce qui pourrait expliquer cette surprenante attribution que réalisent les modèles. De plus, la majorité des textes dont nous disposons de la main d'Adhémar n'appartiennent pas au genre des traités ecclésiastiques, mais davantage au genre historique. La barrière du genre littéraire, reconnue comme problématique par la littérature scientifique, a peut-être également joué un rôle dans l'attribution du *Sermo* à Abbon, qui a rédigé une *Vita* ainsi qu'une exégèse, toutes deux intégrées dans notre corpus de textes.

5.3.2 Le *De utilitatibus astrolabii*

S'il est possible d'observer un seul élément dans les graphiques qui expriment les prédictions vis-à-vis de la paternité du traité sur l'astrolabe (voir la figure 5.13), c'est que Gerbert n'en est pas l'auteur. Une telle affirmation doit bien entendu se prononcer au sein d'une méthode qui, nous l'avons déjà constaté, est très nettement améliorable et pas toujours entièrement fiable. Nos deux meilleurs modèles, les réseaux de mots lemmatisés pris par groupes de 100, rendent compte d'une grande indécision : le traité est attribué aux trois autres auteurs qui accompagnent Gerbert dans notre corpus, avec un nouveau de probabilité entre 30% et 50%. Il ne serait pas injustifié d'y voir l'effet de l'aléatoire.



(a) KNN (L) 100

(b) SVM (L) 100

(c) KNN (FW) 500

(d) SVM (FW) 500

FIG. 5.13 : Probabilités d'attribution du *De utilitatibus astrolabii*

Remarquons une curiosité : les deux modèles basés sur les mots-outils attribuent pour ainsi dire unanimement le traité à Hermann Contract. Rappelons-nous que cet auteur est à l'origine d'une compilation des traités sur l'astrolabe qui circulaient à son époque, et qu'un traité sur l'astrolabe figure dans son propre corpus. Puisqu'il s'agit de réseaux de mots-outils, il est impensable que l'attribution se soit fondée sur un rapprochement thématique, comme nous avons pu le supposer pour le *Sermo* et Abbon de Fleury. Faut-il donc y voir des traces laissées par Hermann lors de sa compilation, qui auraient déteint sur la tradition manuscrite du *De utilitatibus astrolabii* ? Le texte que nous avons utilisé provient également d'une édition (la Patrologie latine) où trois traités, dont un de la main d'Hermann, sont attribués à ce dernier. Ces traces seraient restées au niveau des mots-outils, un niveau inconscient qui garde mémoire de ces transformations ? Ce n'est pas en vain que nous utilisons des questions rhétoriques. Une telle interprétation nous semble tirée par les cheveux et il est bien plus vraisemblable d'y voir une autre imperfection du modèle que nous avons mis en œuvre.

Terminons cette section en remettant en question le choix de la méthode en elle-même. Ce traité sur l'astrolabe, que l'historiographie a tantôt attribué à Hermann, tantôt à Gerbert, tantôt à l'un de ses disciples, tantôt à des disciples des écoles de Fleury et de Chartres, ne serait-il pas plus pertinent de l'envisager comme une véri-

fication de paternité plutôt qu'une attribution de paternité ?[3]. Tout comme l'article de M. Kestemont et al. sur le corpus césarien,[4] il s'agit de déterminer si Gerbert a bel et bien écrit ce traité. Son véritable auteur pourrait être un total inconnu, ou simplement un nom dans les livres d'histoire dont nous n'avons gardé aucun écrit, rendant de la sorte impossible une éventuelle attribution de paternité à proprement parler.

5.4 Synthèse et discussion

Nous avons mis en place une méthode qui fonctionne, du moins partiellement, comme le témoignent les matrices de confusion avec un haut taux d'efficacité dans le début de ce chapitre. Mais, en dépit de ce fait, elle n'est pas suffisamment fiable pour pouvoir nous dire, en cas de résultats surprenants, qu'il doit y avoir quelque chose à creuser au niveau des sources étudiées. Où situer les obstacles qui ont rendu la méthode moins efficace ? Nous percevons plusieurs possibilités.

La qualité du pré-traitement pourrait être une des causes. Nous n'avons pas été chercher la meilleure édition possible pour notre corpus, même si nos textes étaient propres et l'OCR efficace. C. Delcourt et J. Rudman ne seraient pas en accord avec cette décision de se contenter d'un texte qui n'était pas d'une qualité optimale.[5] La suppression forcée de tout le contenu entre parenthèses et crochets a peut-être également quelque peu appauvri le corpus.

Le problème pourrait venir de la sélection des mesures topologique des réseaux. Nos choix se sont basés sur la littérature, qui dit en essence que « plus, c'est mieux ». Comment néanmoins être sûr que certaines mesures ne viennent pas corrompre des résultats qui autrement auraient été meilleurs ? Il est vrai que des algorithmes tels que SVM sont capables de gérer eux-mêmes un grand nombre de caractéristiques et de privilégier celles que sont les meilleures dans la tâche de classification. En effet, SVM avait une meilleure efficacité globale, jusqu'à ce que les mesures topologiques soient normalisées avant d'être traitées par KNN. Les résultats obtenus par l'une et l'autre techniques sont pour ainsi dire similaires, même si SVM conserve un léger avantage.

Les conditions idéales pour une analyse stylométrique sont les suivantes : un long texte à attribuer, un faible nombre d'auteurs potentiels, l'absence de collaborateurs et de réviseurs, l'existence de textes de la main des auteurs, dont la paternité ne fait aucun doute et qui est de genre similaire au texte à attribuer.[6] Il va sans dire que de telles conditions sont rarement réunies dans une situation réelle. En considérant uniquement le cas de Gerbert, nous pouvons remarquer que les textes à attribuer ne brillent

[3] Pour cette distinction, voir la section 3.1

[4] Kestemont et al., « Authenticating the Writings of Julius Caesar ».

[5] Delcourt, « Stylometry », p. 991-992 ; Rudman, « The State of Authorship Attribution Studies », p. 354-358.

[6] Craig, « Stylistic Analysis and Authorship Studies », p. 282-287.

pas par leur longueur (3700 mots pré-traitement pour le *Sermo*) ; que, même s'il n'a pas de collaborateur comme pourrait en avoir une Hildegarde de Bingen, il fut entouré de disciples qu'il a formés et qui pourraient attribuer leurs écrits à Gerbert par déférence vis-à-vis de leur ancien maître ; le nombre d'auteurs-candidats est techniquement limité (ceux dont on a gardé une trace écrite), théoriquement illimité (ou du moins très vaste) et inaccessible ; la disponibilité de textes de même époque et de même genre pose également problème, comme nous avons pu le suggérer dans notre interprétation des résultats obtenus après l'analyse stylométrique du *Sermo*.

Étudier stylométriquement le Moyen Âge est chose possible, comme l'ont déjà démontré de nombreux savants. Il faudrait toutefois, pour s'y essayer fiablement, endurcir les méthodes d'attribution et choisir (ou trouver) une méthode qui permet non seulement d'attribuer un extrait sans doute possible (tâche d'attribution), mais aussi de déterminer si un auteur a écrit un texte ou s'il ne l'a pas fait. Il serait dès lors pertinent, pour une future recherche, d'appliquer une méthode comme celle des imposteurs généraux sur le corpus gerbertien.

Chapitre 6

Conclusion

Notre questionnement d'origine, celui qui nous a initié à la stylométrie et introduit à cette figure fascinante qu'est Gerbert d'Aurillac, était la suivante : est-il l'auteur du *De utilitatibus astrolabii* et du *Sermo de informatione episcoporum*? Si la question portait sur notre appréciation de l'historiographie et notre sensibilité aux arguments avancés par les chercheurs, nous pourrions dire que nous avons notre petite idée. Toutefois, là n'est pas la question, puisqu'implicite dans cette question de recherche se trouvait la suivante : est-il possible de prouver quantitativement que Gerbert est bien l'auteur de ces traités? Dans ce cas, la réponse est limpide : non. La méthode que nous avons mise en place pour y répondre n'est pas suffisamment fiable ou aboutie pour nous permettre d'endosser une si grande responsabilité.

Ce n'est cependant pas en vain que nous avons creusé cette question. Elle nous a permis de découvrir une période méconnue, des centres intellectuels florissants, des ecclésiastiques au cœur d'intrigues politiques, des liens non pas de sang et d'armes mais d'idées et de livres entre le monde arabe et l'Occident chrétien. Plus encore, ce voyage intellectuel nous a mené à la rencontre d'une discipline dont nous ne connaissions même pas le nom et qui pourtant s'ancre dans une longue tradition humaniste et littéraire. Comme de nombreux auteurs l'ont exprimé, la stylométrie n'est que le prolongement de questionnements qui hantent des générations d'érudits depuis les rivages du lac Maréotis jusqu'à nos universités contemporaines.

Des limites techniques encombrent parfois la stylométrie, auxquelles nous nous sommes aussi confronté. Le flou qui entoure le postulat même de la stylométrie, ce « stylome » qui serait le condensé de nos caractéristiques stylistiques inconscientes, bien que son existence ne fasse aucun doute à nos yeux, n'est jamais réellement défini. Transcende-t-il le genre littéraire? Transcende-t-il la langue? Transcende-t-il le temps? S'intéresser à des corpus anciens implique d'être confronté à un manque de sources. Le souci de l'historien devient celui du stylométriste lorsqu'il manque de textes pour comparer efficacement le style d'un auteur à celui d'un autre, et il se retrouve obligé de faire fi des recommandations de sa discipline, qui fonctionne bien

mieux sur des textes de genre et de sujet similaires, pour espérer pouvoir répondre à sa question de recherche.

Le cas de figure dans lequel nous nous sommes trouvé plongé est loin d'être exceptionnel, comme en a témoigné notre exploration de la littérature scientifique. Il existe des méthodes qui permettent de mieux gérer des difficultés de cette espèce que celles que nous avons exploitées, qui par ailleurs possèdent également de nombreux mérites. Pouvoir s'adapter à une langue, le latin médiéval, qui n'avait jamais été étudié sous ce prisme-là auparavant sans requérir autre chose qu'un lemmatiseur n'en est pas le moindre.

Nous sommes convaincu que la discipline continuera à se développer en proposant de nouvelles techniques stylométriques qui pourront venir en complément des méthodes qui sont déjà à la disposition des chercheurs. Gerbert d'Aurillac est, répétons-le, une figure fascinante, de part sa personnalité et de part son époque, une figure qui est encore trop peu connue dans l'histoire médiévale. Il mériterait, ne serait-ce que pour ses pratiques pédagogiques innovantes et les liens qu'il incarne entre religion et sciences, entre science latine et science arabe, ou encore pour son rôle politique de premier plan aux côtés des rois et empereurs de son temps, des études nouvelles et novatrices à son sujet pour renouveler l'historiographie le concernant, elle qui n'a que trop l'habitude de ressasser les mêmes idées reçues dépassées et les mêmes débats – aussi passionnants soient-ils.

Bibliographie

[1] Camilo Akimushkin, Diego Raphael Amancio et Osvaldo Novais Oliveira Jr. « Text Authorship Identified Using the Dynamics of Word Co-Occurrence Networks ». In : *PLOS ONE* 12.1 (2017), e0170527.

[2] Diego Raphael Amancio. « A Complex Network Approach to Stylometry ». In : *PLOS ONE* 10.8 (2015), e0136076.

[3] Diego Raphael Amancio et al. « Comparing Intermittency and Network Measurements of Words and Their Dependence on Authorship ». In : *New Journal of Physics* 13.12 (2011), p. 123024.

[4] Lucas Antiqueira et al. « Some Issues on Complex Networks for Author Characterization ». In : *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial* 11.36 (2007), p. 51-58.

[5] Guy Beaujouan. « Les Apocryphes mathématiques de Gerbert ». In : *"Gerberto : scienza, storia e mito" : atti del Gerberti Symposium (Bobbio 25-27 luglio 1983)*. Sous la dir. de Michele Tosi. Bobbio : Archivi storici bobiensi, 1985, p. 645-658.

[6] Werner Bergmann. *Innovationen Im Quadrivium Des 10. Und 11. Jahrhunderts : Studien Zur Einführung von Astrolab Und Abakus Im Lateinischen Mittelalter*. Stuttgart : Franz Steiner Verlag, 1985.

[7] Carlo Bianchini et Luca J. Senatore. « Gerbert of Aurillac (c. 940–1003) ». In : *Distinguished Figures in Descriptive Geometry and Its Applications for Mechanism Science : From the Middle Ages to the 17th Century*. Sous la dir. de Michela Cigola. History of Mechanism and Machine Science 30. Cham : Springer International Publishing, 2016, p. 33-51.

[8] José Nilo G. Binongo et M. Wilfrid A. Smith. « The Application of Principal Component Analysis to Stylometry ». In : *Literary and Linguistic Computing* 14.4 (1999), p. 445-466.

[9] Arno Borst. *Astrolab Und Klosterreform an Der Jahrtausendwende : Vorgetragen Am 11. Februar 1989*. Heidelberg : Winter, 1989.

[10] Pascale Bourgain. « Les verbes en rapport avec le concept d'auteur ». In : *Auctor et auctoritas. Invention et conformisme dans l'écriture médiévale. Actes du colloque tenu à l'Université de Versailles-Saint-Quentin-en-Yvelines (14-16 juin 1999)*. Sous la dir. de Michel Zimmermann. Paris : Ecole des Chartes, 2001, p. 361-374.

[11] Nicolaus Bubnov. *Gerberti Opera Mathematica (972-1003)*. Berlin : R. Friedländer & Sohn, 1899.

[12] John Burrows. « 'Delta' : A Measure of Stylistic Difference and a Guide to Likely Authorship ». In : *Literary and Linguistic Computing* 17.3 (2002), p. 267-287.

[13] Bernard Cerquiglini. *Éloge de la variante. Histoire critique de la philologie*. Paris : Editions du Seuil, 1989.

[14] Nicole Charbonnel et Jean-Eric Iung, éd. *Gerbert l'Européen. Actes Du Colloque d'Aurillac, 4–7 Juin 1996*. Aurillac : Société des Lettres, Sciences et Arts "La Haute Auvergne", 1997.

[15] Carole E. Chaski. « Who Wrote It ? Steps Toward a Science of Authorship Identification ». In : *National Institute of Justice Journal* 233.233 (1997), p. 15-22.

[16] Hugh Craig. « Stylistic Analysis and Authorship Studies ». In : *A Companion to Digital Humanities*. Sous la dir. de Susan Schreibman, Ray Siemens et John Unsworth. Blackwell Publishing. Malden, Oxford & Victoria : Blackwell Publishing Oxford, UK, 2004, p. 273-288.

[17] Walter Daelemans. « Explanation in Computational Stylometry ». In : *Computational Linguistics and Intelligent Text Processing*. 14th International Conference, CICLing 2013 Samos, Greece, March 24-30, 2013 Proceedings, Part II. Sous la dir. d'Alexander Gelbukh. Berlin & Heidelberg : Springer, 2013, p. 451-462.

[18] Jeroen De Gussem. « Collaborative Authorship in Twelfth-Century Latin Literature : A Stylometric Approach to Gender, Synergy and Authority ». Thèse de doct. Gent & Antwerpen : Universiteit Gent & Universiteit Antwerpen, 2019.

[19] Jeroen De Gussem. « Larger than Life ? A Stylometric Analysis of the Multi-Authored Vita of Hildegard of Bingen ». In : *Interfaces : A Journal of Medieval European Literatures* 8 (2021), p. 125-159.

[20] Henrique F. de Arruda et al. « Paragraph-Based Representation of Texts : A Complex Networks Approach ». In : *Information Processing & Management* 56.3 (2019), p. 479-494.

[21] Christian Delcourt. « Stylometry ». In : *Revue belge de philologie et d'histoire* 80.3 (2002), p. 979-1002.

[22] Jeroen Deploige et Jeroen De Gussem. « Medieval Authorship and Canonicity in the Digital Age – an Introduction ». In : *Interfaces : A Journal of Medieval European Literatures* 8 (2021), p. 113-124.

[23] S. N. Dorogovtsev et J. F. F. Mendes. « Language as an Evolving Word Web ». In : *Proceedings of the Royal Society of London. Series B : Biological Sciences* 268.1485 (2001), p. 2603-2606.

[24] Joseph Drecker. « Hermannus Contractus Über Das Astrolab ». In : *Isis* 16.2 (1931), p. 200-219.

[25] Michael D.C. Drout. « "I Am Large, I Contain Multitudes" : The Medieval Author in Memetic Terms ». In : *Modes of Authorship in the Middle Ages*. Sous la dir. de Slavica Ranković. 22. Toronto : Pontifical Institute of Mediaeval Studies, 2012, p. 30-51.

[26] Maciej Eder. « A Bird's-Eye View of Early Modern Latin : Distant Reading, Network Analysis, and Style Variation ». In : *Early Modern Studies After the Digital Turn*. Sous la dir. de Laura Estill, Diane K. Jakacki et Michael Ullyot. Toronto : Iter Academic Press, 2016, p. 63-90.

[27] Maciej Eder. « Does Size Matter? Authorship Attribution, Small Samples, Big Problem ». In : *Digital Scholarship in the Humanities* 30.2 (2015), p. 167-182.

[28] Maciej Eder. « Visualization in Stylometry : Cluster Analysis Using Networks ». In : *Digital Scholarship in the Humanities* 32.1 (2017), p. 50-64.

[29] Ramon Ferrer i Ferer-i-Cancho et Richard V. Solé. « The Small World of Human Language ». In : *Proceedings of the Royal Society of London. Series B : Biological Sciences* 268.1482 (2001), p. 2261-2265.

[30] Richard S. Forsyth, David I. Holmes et Emily K. Tse. « Cicero, Sigonio, and Burrows : Investigating the Authenticity of the Consolatio ». In : *Literary and Linguistic Computing* 14.3 (1999), p. 375-400.

[31] Michel Foucault. « Qu'est-ce qu'un auteur ? » In : *Bulletin de la Société française de philosophie* 63/3 (1969), p. 73-104.

[32] Henryk Fros. « Les Vies de St-Adalbert - Vojtech, attribuées à Sylvestre II ». In : *"Gerberto : scienza, storia e mito" : atti del Gerberti Symposium (Bobbio 25-27 luglio 1983)*. Sous la dir. de Michele Tosi. Bobbio : Archivi storici bobiensi, 1985, p. 567-576.

[33] Gerbert d'Aurillac. *Correspondance*. Sous la dir. de Pierre Riché et Jean-Pierre Callu. 2 t. Paris : Les Belles Lettres, 1993.

[34] Woon Peng Goh, Kang-Kwong Luke et Siew Ann Cheong. « Functional Shortcuts in Language Co-Occurrence Networks ». In : *PLOS ONE* 13.9 (2018), e0203025.

[35] Virginie Greene, éd. *The Medieval Author in Medieval French Literature*. New York : Springer, 2006.

[36] Jack Grieve. « Quantitative Authorship Attribution : An Evaluation of Techniques ». In : *Literary and linguistic computing* 22.3 (2007), p. 251-270.

[37] J. Berenike Herrmann, Karina van Dalen-Oskam et Christof Schöch. « Revisiting Style, a Key Concept in Literary Studies ». In : *Journal of Literary Theory* 9.1 (2015), p. 25-52.

[38] David I. Holmes. « Authorship Attribution ». In : *Computers and the Humanities* 28.2 (1994), p. 87-106.

[39] David I. Holmes. « The Evolution of Stylometry in Humanities Scholarship ». In : *Literary and Linguistic Computing* 13.3 (1998), p. 111-117.

[40] David L. Hoover. « Quantitative Analysis and Literary Studies ». In : *A Companion to Digital Literary Studies*. Sous la dir. de Susan Schreibman et Ray Siemens. Malden, Oxford & Chichester : John Wiley & Sons, 2013, p. 517-533.

[41] Catherine Jacquemard. « Erectio, inclinatio / erectus, inclinatus : de Vitruve à Gerbert d'Aurillac (à propos de l'expression de la distance angulaire fin Xe - début XIe siècle) ». In : *Collection de l'Institut des Sciences et Techniques de l'Antiquité* 993.1 (2006), p. 157-162.

[42] Catherine Jacquemard, Olivier Desbordes et Alain Hairie. « Du quadrant vetustior à l'horologium viatorum d'Hermann de Reichenau : étude du manuscrit Vaticano, BAV Ott. lat. 1631, f. 16-17v ». In : *Kentron. Revue pluridisciplinaire du monde antique* 23 (23 2007), p. 79-124.

[43] Matthew L. Jockers et Ted Underwood. « Text-Mining The Humanities ». In : *A New Companion to Digital Humanities*. Sous la dir. de Susan Schreibman, Ray Siemens et John Unsworth. Malden, Oxford & Chichester : Wiley-Blackwell, 2016, p. 291-306.

[44] Patrick Juola. « Authorship Attribution ». In : *Foundations and Trends in Information Retrieval* 1.3 (2006), p. 233-334.

[45] Patrick Juola. « The Rowling Case : A Proposed Standard Analytic Protocol for Authorship Questions ». In : *Digital Scholarship in the Humanities* 30 (Suppl. 1 2015), p. i100-i113.

[46] Jakub Kabala. « Computational Authorship Attribution in Medieval Latin Corpora : The Case of the Monk of Lido (ca. 1101–08) and Gallus Anonymous (ca. 1113–17) ». In : *Language Resources and Evaluation* 54.1 (2020), p. 25-56.

[47] Mike KESTEMONT. « Function Words in Authorship Attribution. From Black Magic to Theory ? » In : *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*. Gothenburg, Sweden : Association for Computational Linguistics, 2014, p. 59-66.

[48] Mike KESTEMONT, S. MOENS et Jeroen DEPLOIGE. « Collaborative Authorship in the Twelfth Century : A Stylometric Study of Hildegard of Bingen and Guibert of Gembloux ». In : *Digital Scholarship in the Humanities* 30.2 (2015), p. 199-224.

[49] Mike KESTEMONT et al. « Authenticating the Writings of Julius Caesar ». In : *Expert Systems with Applications* 63 (2016), p. 86-96.

[50] Atle KITTANG. « Authors, Authorship, and Work : A Brief Theoritical Survey ». In : *Modes of Authorship in the Middle Ages*. Sous la dir. de Slavica RANKOVIĆ. 22. Toronto : Pontifical Institute of Mediaeval Studies, 2012, p. 17-29.

[51] Moshe KOPPEL, Jonathan SCHLER et Shlomo ARGAMON. « Computational Methods in Authorship Attribution ». In : *Journal of the American Society for Information Science and Technology* 60.1 (2009), p. 9-26.

[52] Moshe KOPPEL et Yaron WINTER. « Determining If Two Documents Are Written by the Same Author ». In : *Journal of the Association for Information Science and Technology* 65.1 (2014), p. 178-187.

[53] Paul KUNITZSCH. « Les Relations Scientifiques Entre Occident et Monde Arabe ». In : *Gerbert l'Européen. Actes Du Colloque d'Aurillac, 4–7 Juin 1996*. Sous la dir. de Nicole CHARBONNEL et Jean-Eric IUNG. Aurillac : Société des Lettres, Sciences et Arts "La Haute Auvergne", 1997, p. 193-203.

[54] Erik KWAKKEL et Stephen PARTRIDGE. *Author, Reader, Book : Medieval Authorship in Theory and Practice*. Toronto : University of Toronto press, 2011.

[55] Eveline LECLERCQ et Mike KESTEMONT. « Advances in Distant Diplomatics : A Stylometric Approach to Medieval Charters ». In : *Interfaces : A Journal of Medieval European Literatures* 8 (2021), p. 214-244.

[56] Uta LINDGREN. « Gerbert et les arts libéraux ». In : *Gerberto d'Aurillac da abate di Bobbio a papa dell'anno 1000 : atti del congresso internazionale, Bobbio, Auditorium di S. Chiara, 28-30 settembre 2000...* Sous la dir. de Flavio G. NUVOLONE. Bobbio : Associazione culturale Amici di Archivum Bobiense, 2001, p. 107-125.

[57] Uta LINDGREN. « Ptolémée chez Gerbert d'Aurillac ». In : *"Gerberto : scienza, storia e mito" : atti del Gerberti Symposium (Bobbio 25-27 luglio 1983)*. Sous la dir. de Michele TOSI. Bobbio : Archivi storici bobiensi, 1985, p. 619-644.

[58] Harold LOVE. *Attributing Authorship : An Introduction*. Cambridge : Cambridge University Press, 2002. 284 p.

[59] Vanessa Queiroz Marinho, Graeme Hirst et Diego Raphael Amancio. « Authorship Attribution via Network Motifs Identification ». In : *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*. 2016 5th Brazilian Conference on Intelligent Systems (BRACIS). 2016, p. 355-360.

[60] Ian Marriott. « The Authorship of the Historia Augusta : Two Computer Studies ». In : *The Journal of Roman Studies* 69 (1979), p. 65-77.

[61] Armando Martins et al. « Historia Augusta Authorship : An Approach Based on Measurements of Complex Networks ». In : *Applied Network Science* 6.1 (2021), p. 50.

[62] Ali Mehri, Amir H. Darooneh et Ashrafalsadat Shariati. « The Complex Networks Approach for Authorship Attribution of Books ». In : *Physica A : Statistical Mechanics and its Applications* 391.7 (2012), p. 2429-2437.

[63] Jacques-Paul Migne, éd. *Patrologiae Cursus Completus. Series Latina*. 217 t. Paris : Frères Garnier, 1841-1855.

[64] Josep Millàs Vallicrosa. *Assaig d'Història de Les Idees Fisiques i Matemàtiques a La Catalunya Medieval*. Barcelone : Institucio Patxot, 1931.

[65] A. J. Minnis. *Medieval Theory of Authorship : Scholastic Literary Attitudes in the Later Middle Ages*. 2nd ed. with a new preface by the author. Philadelphia : University of Pennsylvania Press, 2010.

[66] Frederick Mosteller et David L. Wallace. « Inference in an Authorship Problem : A Comparative Study of Discrimination Methods Applied to the Authorship of the Disputed Federalist Papers ». In : *Journal of the American Statistical Association* 58.302 (1963), p. 275-309.

[67] Marco Mostert. « Gerbert d'Aurillac, Abbon de Fleury et la culture de l'An Mil : étude comparative de leurs oeuvres et de leur influence ». In : *Gerberto d'Aurillac da abate di Bobbio a papa dell'anno 1000 : atti del congresso internazionale, Bobbio, Auditorium di S. Chiara, 28-30 settembre 2000...* Sous la dir. de Flavio G. Nuvolone. Bobbio : Associazione culturale Amici di Archivum Bobiense, 2001, p. 397-431.

[68] Marco Mostert. « Les Traditions Manuscrites Des Œuvres de Gerbert ». In : *Gerbert l'Européen. Actes Du Colloque d'Aurillac, 4–7 Juin 1996*. Sous la dir. de Nicole Charbonnel et Jean-Eric Iung. Aurillac : Société des Lettres, Sciences et Arts "La Haute Auvergne", 1997, p. 307-324.

[69] Tempestt Neal et al. « Surveying Stylometry Techniques and Applications ». In : *ACM Computing Surveys* 50.6 (2017), p. 1-36.

[70] Flavio G. Nuvolone, éd. *Gerberto d'Aurillac - Silvestro II, Linee per Una Sintesi : Atti Del Convegno Internazionale, Bobbio, Auditorium Di S. Chiara, 11 Settembre 2004, Sotto La Presidenza Del Prof. Pierre Racine...* Bobbio : Associazione culturale Amici di Archivum bobiense, 2005.

[71] Flavio G. Nuvolone, éd. *Gerberto d'Aurillac da abate di Bobbio a papa dell'anno 1000 : atti del congresso internazionale, Bobbio, Auditorium di S. Chiara, 28-30 settembre 2000...* Bobbio : Associazione culturale Amici di Archivum Bobiense, 2001.

[72] Flavio G. Nuvolone. « Il *Sermo pastoralis* Pseudoambrosiano e il *Sermo Giberti philosophi papae urbis Romae qui cognominatus est Silvester de informatione Episcoporum*. Riflessioni ». In : *"Gerberto : scienza, storia e mito" : atti del Gerberti Symposium (Bobbio 25-27 luglio 1983)*. Sous la dir. de Michele Tosi. Bobbio : Archivi storici bobiensi, 1985, p. 379-565.

[73] Alexandre Olleris. *Oeuvres de Gerbert, pape sous le nom de Sylvestre II*. Clermont-Ferrand & Paris : F. Thibaud & Ch. Dumoulin, 1867.

[74] Marek Otisk. « Gerbert of Aurillac (Pope Sylvester II) as a Clockmaker ». In : *Teorie vědy/Theory of Science* 42.1 (2020), p. 25-49.

[75] Emmanuel Poulle. « Gerbert Homme de Science ». In : *Gerberto d'Aurillac - Silvestro II, Linee per Una Sintesi : Atti Del Convegno Internazionale, Bobbio, Auditorium Di S. Chiara, 11 Settembre 2004, Sotto La Presidenza Del Prof. Pierre Racine...* Sous la dir. de Flavio G. Nuvolone. Bobbio : Associazione culturale Amici di Archivum bobiense, 2005, p. 95-123.

[76] Emmanuel Poulle. « L'Astronomie de Gerbert ». In : *"Gerberto : scienza, storia e mito" : atti del Gerberti Symposium (Bobbio 25-27 luglio 1983)*. Sous la dir. de Michele Tosi. Bobbio : Archivi storici bobiensi, 1985, p. 597-617.

[77] Laura V. C. Quispe, Jorge A. V. Tohalino et Diego R. Amancio. « Using Virtual Edges to Improve the Discriminability of Co-Occurrence Text Networks ». In : *Physica A : Statistical Mechanics and its Applications* 562 (2021).

[78] Slavica Ranković, éd. *Modes of Authorship in the Middle Ages*. 22. Toronto : Pontifical Institute of Mediaeval Studies, 2012.

[79] Pierre Riché. *Gerbert d'Aurillac : Le Pape de l'an Mil*. Paris : Fayard, 1987.

[80] Pierre Riché. *Les Grandeurs de l'an Mille*. Paris : Bartillat, 1999.

[81] Richer de Saint-Rémy. *Histoire de France (888-995)*. Sous la dir. de Robert Latouche. 2 t. Paris : H. Champion, 1930.

[82] Joseph Rudman. « The State of Authorship Attribution Studies : Some Problems and Solutions ». In : *Computers and the Humanities* 31.4 (1998), p. 351-365.

[83] Alain Schärlig. *Un Portrait de Gerbert d'Aurillac : Inventeur d'un Abaque, Utilisateur Précoce Des Chiffres Arabes, et Pape de l'an Mil*. Lausanne : Presses polytechniques et universitaires romandes, 2012.

[84] Santiago Segarra, Mark Eisen et Alejandro Ribeiro. « Authorship Attribution Through Function Word Adjacency Networks ». In : *IEEE Transactions on Signal Processing* 63.20 (2015), p. 5464-5478.

[85] Santiago Segarra, Mark Eisen et Alejandro Ribeiro. « Authorship Attribution Using Function Words Adjacency Networks ». In : *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. Vancouver, Canada : IEEE, 2013, p. 5563-5567.

[86] Costantino Sigismondi. « Gerbert of Aurillac : Astronomy and Geometry in Tenth Century Europe ». In : *International Journal of Modern Physics : Conference Series* 23 (2013), p. 467-471. arXiv : 1201.6094.

[87] Efstathios Stamatatos. « A Survey of Modern Authorship Attribution Methods ». In : *Journal of the American Society for Information Science and Technology* 60.3 (2009), p. 538-556.

[88] Efstathios Stamatatos. « On the Robustness of Authorship Attribution Based on Character N-Gram Features ». In : *Journal of Law and Policy* 21.2 (2013), p. 421-439.

[89] Tomasz Stanisz, Jarosław Kwapień et Stanisław Drożdż. « Linguistic Data Mining with Complex Networks : A Stylometric-Oriented Approach ». In : *Information Sciences* 482 (2019), p. 301-320.

[90] Michele Tosi, éd. *"Gerberto : scienza, storia e mito" : atti del Gerberti Symposium (Bobbio 25-27 luglio 1983)*. Bobbio : Archivi storici bobiensi, 1985.

[91] André Van de Vyver. « Les Premières Traductions Latines (X^{e}-XI^{e} s.) de Traites Arabes Sur l'astrolabe ». In : *1er Congrès International de Géographie Historique*. Sous la dir. de Fritz Quicke. T. 2. 1931, p. 266-290.

[92] A. van de Vyver. « Les plus Anciennes Traductions Latines Médiévales (X^{e}-XI^{e} Siècles) de Traités d'astronomie et d'astrologie ». In : *Osiris* 1 (1936), p. 658-691.

[93] Hans Van Halteren et al. « New Machine Learning Methods Demonstrate the Existence of a Human Stylome ». In : *Journal of Quantitative Linguistics* 12.1 (2005), p. 65-77.

[94] George Huntston Williams. « The Golden Priesthood and the Leaden State. A Note on the Influence of a Work Sometimes Ascribed to St. Ambrose : The Sermo de Dignitate Sacerdotali ». In : *Harvard Theological Review* 50.1 (1957), p. 37-64.

[95] Michel ZIMMERMANN, éd. *Auctor et auctoritas. Invention et conformisme dans l'écriture médiévale. Actes du colloque tenu à l'Université de Versailles-Saint-Quentin-en-Yvelines (14-16 juin 1999)*. Mémoires et documents de l'Ecole des Chartes 59. Paris : Ecole des Chartes, 2001.

[96] Michel ZINK. «Auteur et autorité au Moyen Âge». In : *De l'autorité. Colloque annuel du Collège de France*. Paris : Odile Jacob, 2008, p. 143-158.

[97] Marco ZUCCATO. «Arabic Singing Girls, the Pope, and the Astrolabe : Arabic Science in Tenth-Century Latin Europe». In : *Viator* 45.1 (2014), p. 99-120.

[98] Marco ZUCCATO. «Gerbert of Aurillac and a Tenth-Century Jewish Channel for the Transmission of Arabic Science to the West». In : *Speculum* 80.3 (2005), p. 742-763.

[99] Paul ZUMTHOR. *La Lettre et La Voix. De La "Littérature" Médiévale*. Paris : Editions du Seuil, 1987.

Annexe A

Code R

Le code R utilisé pour les différentes étapes de mes analyses est reproduit ci-après. Il se compose de quatre fichiers, dont le nom est précédé par un numéro qui représente l'étape de la méthodologie mise en œuvre : dans l'ordre, le pré-traitement, l'analyse de réseaux, la classification.

Le fichier `01_pretraitementTexte.R` reprend le code relatif à un fichier texte (le code pour traiter un fichier PDF n'a pas été utilisé dans le cadre de ce mémoire). Le fichier `02_matriceFeatures.R` transforme les textes pré-traités en réseaux dont les propriétés sont calculées. Enfin, les fichiers `03_ACP.R` et `03_classification.R` implémentent les tâches de classification non-supervisée (ACP) pour le premier et supervisée (KNN et SVM) pour le second.

A.1 01_pretraitementTexte.R

```
rm(list=ls())
library(stringr)

# Fonctions pour appeler script de tokenisation et
   TreeTagger

monTreeTagger <- function(input,output){
        # chemin vers l'executable TreTagger lui-m?me
        exeTreeTagger = "C:/TreeTagger/bin/tree-tagger.exe"
        # options à passer à la commande
        options = "-lemma -token"
        parametres = "C:/TreeTagger/lib/mediolatin-6.par"

        ligneAExecuter = paste(exeTreeTagger,options,
```

```
            parametres,input,output,sep=" ")

        system(ligneAExecuter)
}

tokenisationPerl <- function(input){

        scriptPerl = "perl tokenizelat-6.pl"

        ligneAExecuter = paste(scriptPerl,input,sep=" ")

        system(ligneAExecuter)
}

# Pour chaque fichier, retirer le maximum d'éléments qui ne
    sont pas le fait de l'auteur
# Réécrire le fichier nettoyé préfixé de clean_

listeTextes <- list.files(path = "./textes/", pattern='*.
   txt')

for (fichier in listeTextes) {
        texte <- readLines(paste("./textes/", fichier, sep
           =""))
        texte1 <- gsub("<.*?>", "", texte) # retire les
           balises HTML
        texte2 <- gsub("[0-9]", "", texte1) # retire les
           chiffres arabes
        texte3 <- gsub("\\(.*?\\)", "", texte2) # retire
           les parenthèses et leur contenu
        texte4 <- gsub("\\[.*?\\]", "", texte3) # retire
           les crochets et leur contenu
        texte5 <- str_squish(texte4) # retire les espaces
           blancs
        writeLines(texte5, paste("./textes/tmp/clean_",
           fichier, sep=""))
}

# Exécuter le script de tokenisation
```

```
listeTextesClean <- list.files(path = "./textes/tmp/",
   pattern='^clean_')

for (fichier in listeTextesClean) {
        setwd("./textes/tmp")
        tokenisationPerl(fichier)
        print(fichier)
        setwd("../..")
}

# Exécuter TreeTagger

listeTextesToken <- list.files(path = "./textes/tmp/",
   pattern='^token_')

for (fichier in listeTextesToken) {
        setwd("./textes/tmp")
        monTreeTagger(fichier,paste("lemma_",fichier,sep
           =""))
        print(fichier)
        setwd("../..")
}
```

A.2 02_matriceFeatures.R

```
rm(list=ls())

library(stringr)
library(igraph)
library(tidyverse)
library(tibble)
library(ggplot2)

# Créer une matrice avec tous les auteurs (avec colonnes
   titre, n° paquet et les différentes features)

K = 500 # nbre de tokens par paquet
```

```
listeTextes <- list.files(path = "./dataLemmatisees",
   pattern = "txt$")

matriceFeatures = tibble(auteur = NA, titre = NA, paquet =
   NA, nbSommets = NA, diametre = NA, rayon = NA,
   degreSortant = NA, distributionDegre = NA, degrePondere
   = NA, plusPetitCheminMoyen = NA, intermediarite = NA,
   clustCoef = NA, assortCoef = NA, nombreMotifs = NA,
   nombreCliques = NA, taillePlusGrandeClique = NA, .rows =
    0)

auteurNom = NA
titreNom = NA

# Peupler la matrice avec les données en lisant les textes

for (fichier in listeTextes) {

        # pour chaque fichier (texte lemma)

        texte = readLines(paste("./dataLemmatisees/",
           fichier, sep = ""))
        texte = str_split_fixed(texte, "[\t ]", n = 4) #
           lire le texte sous format de matrice
        texte = as.data.frame(texte) # en faire un data
           frame

        # Trier le data frame selon les POS que l'on veut
           conserver (modifiable)

        # suppression des noms propres (pas lemmatisés) et
           des lemmes inconnus
        texte = filter(texte, V2!="NAM"&V3!="<unknown>")

        # suppression de la ponctuation
        texte = filter(texte, V2!="PON"&V2!="SENT"&V2!="NUM
           ")

        # suppression des function words
```

```
texte = filter(texte, V2!="CON"&V2!="INT"&V2!="PRO
   ")

# ne garder que les function words
#texte = filter(texte, V2=="CON"|V2=="INT"|V2=="PRO
   ")

# Créer les paquets

MO = dim(texte)[1] # nbre de tokens du texte
print(MO)

if (MO >= K) { # évite une out of bound error pour
   les plus petits textes si on prend un K élevé

        nbPaquets = floor(MO/K) # prendre l'entier
           en dessous

        # nom de l'auteur
        auteurNom = str_replace(fichier,".*[0-9]{3}
           _([A-z]+)_.*","\\1")

        # titre
        titreNom = str_replace(fichier,".*[0-9]{3}_
           [A-z]+_([0-z]+)\\.txt$","\\1")

        # Crée un clone de la matrice de features
           de longueur équivalente au nombre de
           mots du paquet

        ajout = tibble(auteur = auteurNom, titre =
           titreNom, paquet = NA, nbSommets = NA,
           diametre = NA, rayon = NA, degreSortant
           = NA, distributionDegre = NA,
           degrePondere = NA, plusPetitCheminMoyen
           = NA, intermediarite = NA, clustCoef =
           NA, assortCoef = NA, nombreMotifs = NA,
           nombreCliques = NA,
           taillePlusGrandeClique = NA, .rows =
```

```
   nbPaquets)

# Pour chaque paquet

for (i in 1:nbPaquets){
        ajout$paquet[i] = i # garde le nr
           du paquet

        # POS et lemmes de paquet

        indDeb = (i-1)*K+1
        indFin = i*K
        motsDuPaquet = texte[indDeb:indFin,
            3] # 1 = mot et 3 = lemme
        posDuPaquet = texte[indDeb:indFin,
           2]

        # Création des réseaux + sauvegarde
            des propriétés

        listeAretes = cbind(motsDuPaquet
           [1:(K-1)],motsDuPaquet[2:K]) #
           crée la matrice d'adjacence
        grPaquetsMots = graph_from_edgelist
           (el = listeAretes,directed =
           TRUE) # utilise la matrice pour
           créer un graphe dirigé

        ajout$nbSommets[i] = length(V(
           grPaquetsMots))

        ajout$diametre[i] = diameter(
           grPaquetsMots)

        ajout$rayon[i] = radius(
           grPaquetsMots)

        ajout$degreSortant[i] = mean(degree
           (grPaquetsMots, mode = "out"))
```

```
            ajout$distributionDegre[i] = mean(
               degree_distribution(
               grPaquetsMots))

            ajout$degrePondere[i] = mean(
               strength(grPaquetsMots))

            ajout$plusPetitCheminMoyen[i] =
               mean_distance(grPaquetsMots)

            ajout$intermediarite[i] = mean(
               betweenness(grPaquetsMots,
               normalized = TRUE))

            ajout$clustCoef[i] = transitivity(
               grPaquetsMots)

            ajout$assortCoef[i] =
               assortativity_degree(
               grPaquetsMots)

            ajout$nombreMotifs[i] =
               count_motifs(grPaquetsMots) #
               motifs de 3 par défaut

            ajout$nombreCliques[i] =
               count_max_cliques(grPaquetsMots)

            ajout$taillePlusGrandeClique[i] =
               clique_num(grPaquetsMots)
}

# La matrice "ajout" est collée à la
   matrice de features complète

matriceFeatures = rbind(matriceFeatures,
   ajout)
print(fichier)
```

```
        }
}

# Visualiser le nombre de paquets par auteur dans la
   matrice de features

ggplot(matriceFeatures) +
geom_bar(aes(x = auteur, y = after_stat(count))) +
labs(x = "Auteur", y = "Nombre de paquets")
ggsave(path = "figures", filename = paste(K,"
   _nbrePaquetsMatrice.png", sep=""))

# Visualiser un graphe donné (en l'occurrence le dernier
   créé)

net.graphopt <- layout_with_graphopt(grPaquetsMots, charge
   = 0.009, mass = 50, spring.length = E(grPaquetsMots)
   $weight)
plot(grPaquetsMots, vertex.size=1, edge.arrow.size=0.3,
   layout=net.graphopt)

# Sauvegarde la matrice pour future utilisation (en changer
    le nom suivant les POS retenus)

save(matriceFeatures, file = paste("./RData/
   matriceFeatures_", K ,".RData", sep = ""))
```

A.3 03_ACP.R

```
rm(list=ls())

library(FactoMineR)
library(factoextra)

load(file = "RData/matriceFeatures_500.RData") # charge la
   matrice de features désirée
colFeatures = 4:16 # indique sur quelles features
   construire l'ACP
```

```
# normalisation : l'ACP normalise pour nous, on obtient la
   même chose

# for (i in colFeatures){
        #  matriceFeatures[,i] = (matriceFeatures[,i] -
        #                           mean(unlist(
           matriceFeatures[,i]))) /
        #    (sd(unlist(matriceFeatures[,i])))
        #}

# Trier la matrice

matriceFeatures = subset(matriceFeatures, !(titre %in% c("
   DeUtAs2", "DeCoEtS5", "DeGeo"))) # retirer les textes
   douteux mais pas les textes étudiés

# Calcul des composantes principales

resACP = PCA(X = matriceFeatures[,colFeatures], graph =
   FALSE) # ne pas faire de graphique (ggplot préféré)
individus = resACP$ind # sauvegarder les individus pour la
   visualisation

# Mesurer la variance

barplot(resACP$eig[, 2], names.arg=1:nrow(resACP$eig),
xlab = "Composantes principales",
ylab = "Pourcentage de variance",
col ="steelblue")
lines(x = 1:nrow(resACP$eig), resACP$eig[, 2],
type="b", pch=19, col = "red")

# Comprendre ce qui contribue aux composantes principales

dimdesc(resACP, axes=c(1,2))

# Visualiser l'ACP
```

```
dataToPlot = tibble(auteur = matriceFeatures$auteur,
titre = matriceFeatures$titre,
coordAxe1 = individus$coord[,1],
coordAxe2 = individus$coord[,2])

# Extraire les données des orphelins puis les retirer
dataOrphelins = subset(dataToPlot, (titre %in% c("DeUtAs1",
    "SeDeInE")))
dataToPlot = subset(dataToPlot, !(titre %in% c("DeUtAs1", "
   SeDeInE")))

ggplot(dataToPlot) +
geom_point(aes(x = coordAxe1, y = coordAxe2, col = auteur))
    +
geom_point(data = dataOrphelins,
aes(x=coordAxe1, y = coordAxe2, col = titre), size = 3) +
labs(x="Composante 1", y = "Composante 2")
ggsave(path = "figures", filename = "ACP.png")

# Visualiser un sous-ensemble des données : les lettres de
   Gerbert et d'Abbon
dataLettres = subset(dataToPlot, (titre %in% c("Episto223",
    "EpScAnS")))

ggplot(dataLettres) +
geom_point(aes(x = coordAxe1, y = coordAxe2, col = auteur))
    +
labs(x="Composante 1", y = "Composante 2")
ggsave(path = "figures", filename = "ACP_Lettres.png")

dataHistoire = subset(dataToPlot, (titre %in% c("ExDeViR",
   "Histor4", "Chroni25")))

ggplot(dataHistoire) +
geom_point(aes(x = coordAxe1, y = coordAxe2, col = auteur))
    +
labs(x="Composante 1", y = "Composante 2")
ggsave(path = "figures", filename = "ACP_Histoire.png")
```

A.4 03_classification.R

```
rm(list=ls())

library(class)
library(tidyverse)
library(e1071)
library(ggplot2)

load(file = "RData/matriceFeatures_500.RData") # charge la
   matrice de features désirée

######################
# adapter la matrice
######################

# normalisation
colFeatures = 4:16

#pour noLemma car NA
#matriceFeatures <- subset(matriceFeatures, select = -c(
   assortCoef))

# Normalisation 1
# for (i in colFeatures){
        #   matriceFeatures[,i] = (matriceFeatures[,i] -
           min(matriceFeatures[,i])) / (max(matriceFeatures
           [,i]) - min(matriceFeatures[,i]))
        # }

# Normalisation 2
for (i in colFeatures){
        matriceFeatures[,i] = (matriceFeatures[,i] -
        mean(unlist(matriceFeatures[,i]))) /
        (sd(unlist(matriceFeatures[,i])))
}

# Tri dans la matrice
```

```
matriceFeatures <- matriceFeatures %>% mutate_at("auteur",
   factor) # factor demandé par SVM
matriceFeaturesOrphelins = subset(matriceFeatures, (titre %
   in% c("DeUtAs1","SeDeInE"))) # sauvegarder les deux
   textes douteux étudiés
matriceFeatures = subset(matriceFeatures, !(titre %in% c("
   DeUtAs1","SeDeInE", "DeUtAs2", "DeCoEtS5", "DeGeo"))) #
   retirer les textes douteux

# Combien de paquets par auteur

ggplot(matriceFeatures) +
geom_bar(aes(x = auteur, y = after_stat(count))) +
labs(x = "Auteur", y = "Nombre de paquets")
ggsave(path = "figures", filename = "
   nbrePaquetsMatriceSansOrphelins.png")

ggplot(matriceFeaturesOrphelins) +
geom_bar(aes(x = titre, y = after_stat(count))) +
labs(x = "Titre", y = "Nombre de paquets")
ggsave(path = "figures", filename = "nbrePaquetsOrphelins.
   png")

##############################
# créer un training set
##############################

N = dim(matriceFeatures)[1]
colClass <- 1 # auteur
colARetirer <- 2:3 # titre et nbre paquets
set.seed(1)
indexTest <- sample(1:N, size = round(N/3), replace = FALSE
   ,prob = rep(1/N, N))

# Construire les deux sets avec indexTest
training.set <- matriceFeatures[-indexTest,-c(colClass,
   colARetirer)]
training.class<- as.factor(unlist(matriceFeatures[-
   indexTest,colClass]))
```

```
test.set <- matriceFeatures[indexTest,-c(colClass,
   colARetirer)]
test.class <- as.factor(unlist(matriceFeatures[indexTest,
   colClass]))

######
# KNN
######

# pour calculer le K idéal

accuracy = array(10)
for (k in 1:10){
        mydata.knn <- knn(training.set, test.set, training.
           class, k = k)
        cm<-table(as.factor(test.class), mydata.knn) #
           matrice de confusion
        accuracy[k] = (sum(cm) - sum(diag(cm)))/sum(cm)
}
plot(accuracy, t='l')

# modèle

mydata.knn <- knn(training.set, test.set, training.class, k
    = 2)

summary(mydata.knn)

cm<-table(as.factor(test.class), mydata.knn) # matrice de
   confusion
cm

# visualiser la cm
cmPond = cm / rowSums(cm) # créer data.frame pondérée
cmPond.df = as.data.frame(cmPond)
ggplot(cmPond.df,
aes(x=Var1,y=mydata.knn,fill=Freq)) + geom_tile() +
scale_fill_gradient(low="lightgreen",high="darkgreen",guide
   ="colorbar") +
```

```
labs(x = "Réalité", y = "Prédiction")
ggsave(path = "figures", filename = "cmTrainKNN.png")

# taux d'erreur
(sum(cm) - sum(diag(cm)))/sum(cm)

# Comparer avec Sermo

matriceFeaturesSermo = subset(matriceFeaturesOrphelins,
   titre  == "SeDeInE")
featuresSermo = matriceFeaturesSermo[,-c(colClass,
   colARetirer)]
predictionSermo.knn = knn(training.set, featuresSermo,
   training.class,prob = TRUE, k = 2)

# résultats Sermo
summary(predictionSermo.knn)
ggplot(data = as.data.frame(predictionSermo.knn)) +
geom_bar(mapping = aes(x=predictionSermo.knn,
y=..count../sum(..count..))) +
scale_x_discrete(drop = FALSE) +
labs(x = "Auteur", y = "Probabilité de paternité")
ggsave(path = "figures", filename = "cmSermoKNN_1000.png")

# Comparer avec Astro

matriceFeaturesAstro = subset(matriceFeaturesOrphelins,
   titre  == "DeUtAs1")
featuresAstro = matriceFeaturesAstro[,-c(colClass,
   colARetirer)]
predictionAstro.knn = knn(training.set, featuresAstro,
   training.class,prob = TRUE, k = 2)

# résultats Astro
summary(predictionAstro.knn)
ggplot(data = as.data.frame(predictionAstro.knn)) +
geom_bar(mapping = aes(x=predictionAstro.knn,
y=..count../sum(..count..))) +
scale_x_discrete(drop = FALSE) +
```

```
labs(x = "Auteur", y = "Probabilité de paternité")
ggsave(path = "figures", filename = "cmAstroKNN_1000.png")

######
# SVM
######

matriceFeatures.svm.model<- svm(auteur ~ ., data =
   matriceFeatures[-indexTest, -colARetirer])
matriceFeatures.svm.pred<-predict(matriceFeatures.svm.model
   ,test.set )

cm<-table(test.class, matriceFeatures.svm.pred)
cm

# visualiser la cm
cmPond = cm / rowSums(cm) # créer data.frame pondérée
cmPond.df = as.data.frame(cmPond)
ggplot(cmPond.df,
aes(x=test.class,y=matriceFeatures.svm.pred,fill=Freq)) +
   geom_tile() +
scale_fill_gradient(low="lightgreen",high="darkgreen",guide
   ="colorbar") +
labs(x = "Réalité", y = "Prédiction")
ggsave(path = "figures", filename = "cmTrainSVM.png")

# taux d'erreur
(sum(cm) - sum(diag(cm)))/sum(cm)

# Comparer avec Sermo

predictionSermo.svm = predict(matriceFeatures.svm.model,
   featuresSermo)
summary(predictionSermo.svm) # résultats Sermo

ggplot(data = as.data.frame(predictionSermo.svm)) +
geom_bar(mapping = aes(x=predictionSermo.svm,
y=..count../sum(..count..))) +
scale_x_discrete(drop = FALSE) +
```

```
labs(x = "Auteur", y = "Probabilité de paternité")
ggsave(path = "figures", filename = "cmSermoSVM_500.png")

# Comparer avec Astro

predictionAstro.svm = predict(matriceFeatures.svm.model,
   featuresAstro)
summary(predictionAstro.svm) # résultats Astro

ggplot(data = as.data.frame(predictionAstro.svm)) +
geom_bar(mapping = aes(x=predictionAstro.svm,
y=..count../sum(..count..))) +
scale_x_discrete(drop = FALSE) +
labs(x = "Auteur", y = "Probabilité de paternité")
ggsave(path = "figures", filename = "cmAstroSVM_500.png")

#####################################
# Tests avec un même genre littéraire
#####################################

# Lettres

load(file = "RData/matriceFeatures_100.RData") # charge la
   matrice de features désirée
matriceFeaturesLettres = subset(matriceFeatures, (titre %in
   % c("Episto223", "EpScAnS")))
matriceFeaturesLettres <- matriceFeaturesLettres %>%
   mutate_at("auteur", factor) # factor demandé par SVM

N = dim(matriceFeaturesLettres)[1]
colClass <- 1 # auteur
colARetirer <- 2:3 # titre et nbre paquets
set.seed(1)
indexTest <- sample(1:N, size = round(N/3), replace = FALSE
   ,prob = rep(1/N, N))

# Construire les deux sets avec indexTest
training.set <- matriceFeaturesLettres[-indexTest,-c(
   colClass,colARetirer)]
```

```
training.class<- as.factor(unlist(matriceFeaturesLettres[-
   indexTest,colClass]))
test.set <- matriceFeaturesLettres[indexTest,-c(colClass,
   colARetirer)]
test.class <- as.factor(unlist(matriceFeaturesLettres[
   indexTest,colClass]))

# KNN

mydataLettres.knn <- knn(training.set, test.set, training.
   class, k = 2)

summary(mydataLettres.knn)

cm<-table(as.factor(test.class), mydataLettres.knn) #
   matrice de confusion
cm

# visualiser la cm
cmPond = cm / rowSums(cm) # créer data.frame pondérée
cmPond.df = as.data.frame(cmPond)
ggplot(cmPond.df,
aes(x=Var1,y=mydataLettres.knn,fill=Freq)) + geom_tile() +
geom_text(aes(label = Freq)) +
scale_fill_gradient(low="lightgreen",high="darkgreen",guide
   ="colorbar") +
labs(x = "Réalité", y = "Prédiction")
ggsave(path = "figures", filename = "cmLettres_KNN.png")

# taux d'erreur
(sum(cm) - sum(diag(cm)))/sum(cm)

# SVM

matriceFeaturesLettres.svm.model<- svm(auteur ~ ., data =
   matriceFeaturesLettres[-indexTest, -colARetirer])
matriceFeaturesLettres.svm.pred<-predict(
   matriceFeaturesLettres.svm.model,test.set )
```

```
cm<-table(test.class, matriceFeaturesLettres.svm.pred)
cm

# visualiser la cm
cmPond = cm / rowSums(cm) # créer data.frame pondérée
cmPond.df = as.data.frame(cmPond)
ggplot(cmPond.df,
aes(x=test.class,y=matriceFeaturesLettres.svm.pred,fill=
   Freq)) + geom_tile() +
geom_text(aes(label = Freq)) +
scale_fill_gradient(low="lightgreen",high="darkgreen",guide
   ="colorbar") +
labs(x = "Réalité", y = "Prédiction")
ggsave(path = "figures", filename = "cmLettres_SVM.png")

# taux d'erreur
(sum(cm) - sum(diag(cm)))/sum(cm)

# Histoire

load(file = "RData/matriceFeatures_500_FW.RData") # charge
   la matrice de features désirée
matriceFeaturesHistoire = subset(matriceFeatures, (titre %
   in% c("ExDeViR", "Histor4", "Chroni25")))
matriceFeaturesHistoire <- matriceFeaturesHistoire %>%
   mutate_at("auteur", factor) # factor demandé par SVM

N = dim(matriceFeaturesHistoire)[1]
colClass <- 1 # auteur
colARetirer <- 2:3 # titre et nbre paquets
set.seed(1)
indexTest <- sample(1:N, size = round(N/3), replace = FALSE
   ,prob = rep(1/N, N))

# Construire les deux sets avec indexTest
training.set <- matriceFeaturesHistoire[-indexTest,-c(
   colClass,colARetirer)]
training.class<- as.factor(unlist(matriceFeaturesHistoire[-
   indexTest,colClass]))
```

```
test.set <- matriceFeaturesHistoire[indexTest,-c(colClass,
   colARetirer)]
test.class <- as.factor(unlist(matriceFeaturesHistoire[
   indexTest,colClass]))

# KNN

mydataHistoire.knn <- knn(training.set, test.set, training.
   class, k = 2)

summary(mydataHistoire.knn)

cm<-table(as.factor(test.class), mydataHistoire.knn) #
   matrice de confusion
cm

# visualiser la cm
cmPond = cm / rowSums(cm) # créer data.frame pondérée
cmPond.df = as.data.frame(cmPond)
ggplot(cmPond.df,
aes(x=Var1,y=mydataHistoire.knn,fill=Freq)) + geom_tile() +
geom_text(aes(label = Freq)) +
scale_fill_gradient(low="lightgreen",high="darkgreen",guide
   ="colorbar") +
labs(x = "Réalité", y = "Prédiction")
ggsave(path = "figures", filename = "cmHistoire_500_FW_KNN.
   png")

# taux d'erreur
(sum(cm) - sum(diag(cm)))/sum(cm)

# SVM

matriceFeaturesHistoire.svm.model<- svm(auteur ~ ., data =
   matriceFeaturesHistoire[-indexTest, -colARetirer])
matriceFeaturesHistoire.svm.pred<-predict(
   matriceFeaturesHistoire.svm.model,test.set )

cm<-table(test.class, matriceFeaturesHistoire.svm.pred)
```

```
cm

# visualiser la cm
cmPond = cm / rowSums(cm) # créer data.frame pondérée
cmPond.df = as.data.frame(cmPond)
ggplot(cmPond.df,
aes(x=test.class,y=matriceFeaturesHistoire.svm.pred,fill=
   Freq)) + geom_tile() +
geom_text(aes(label = Freq)) +
scale_fill_gradient(low="lightgreen",high="darkgreen",guide
   ="colorbar") +
labs(x = "Réalité", y = "Prédiction")
ggsave(path = "figures", filename = "cmHistoire_500_FW_SVM.
   png")

# taux d'erreur
(sum(cm) - sum(diag(cm)))/sum(cm)
```