

REVUE DE
LINGUISTIQUE
FRANÇAISE
DIACHRONIQUE

7
2017

DIACHRONIQUES

LES ÉTATS ANCIENS
DES LANGUES À L'HEURE
DU NUMÉRIQUE

Richards – 979-10-231-2167-4



LES ÉTATS ANCIENS DES LANGUES À L'HEURE DU NUMÉRIQUE

JOËLLE DUCOS

Présentation

ROBERT MARTIN

À propos du *DMF* : réussites et pièges de la lexicographie électronique

SYLVIE BAZIN-TACHELLA & GILLES SOUVAY

De la gestion de la variation en moyen français à son élargissement aux états anciens du français : les développements du lemmatiseur LGeRM

XAVIER-LAURENT SALVADOR, FABRICE ISSAC & MARCO FASCILO

Herméneutique des similarités dans le *DFSM* : une expérience

ESTRELLA PÉREZ RODRÍGUEZ

Le *Lexicon Latinitatis Medii Aevi Regni Legionis* (VIII^e siècle-1230) : caractéristiques et quelques exemples (*ventrescas, iera, cumbo, plentum*)

ELISA GUADAGNINI

La lexicographie de l'italien médiéval et les corpus de l'OVI : un bilan provisoire et quelques nouvelles perspectives

ANA GÓMEZ RABAL

Le latin médiéval du *Glossarium Mediae Latinitatis Cataloniae* : un projet lexicographique dans un contexte européen

MICHÈLE GOYENS & CÉLINE SZECEL

Autorité du latin et transparence constructionnelle : le sort des néologismes médiévaux dans le domaine médical

CÉLINE GUILLOT, SERGE HEIDEN & ALEXEI LAVRENTIEV

Base de français médiéval : une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique

GÉRARD PETIT

Terminographie diachronique : le cas de la terminologie médiévale française

RAMON MASÍÀ

Numérisation et traitement de textes mathématiques grecs : méthodes, problèmes et résultats

EARL JEFFREY RICHARDS

À la recherche des communautés discursives au Moyen Âge : un regard numérique sur la connectivité dans la culture vernaculaire et le rôle des traductions dans l'évolution de la prose en moyen français



LES ÉTATS ANCIENS DES LANGUES
À L'HEURE DU NUMÉRIQUE

Les états anciens
des langues
à l'heure du numérique



Les PUPS, désormais SUP, sont un service général
de la faculté des Lettres de Sorbonne Université.

© Presses de l'université Paris-Sorbonne, 2018

© Sorbonne Université Presses, 2021

Diachroniques n° 7

ISBN papier : 979-10-231-0581-0

PDF complet – 979-10-231-2155-1

TIRÉS À PART EN PDF :

Ducos – 979-10-231-2156-8

Martin – 979-10-231-2157-5

Bazin-Tacchella & Souvay – 979-10-231-2158-2

Salvador, Issac & Fasciolo – 979-10-231-2159-9

Pérez Rodríguez – 979-10-231-2160-5

Guadagnini – 979-10-231-2161-2

Gómez Rabal – 979-10-231-2162-9

Goyens & Szeceł – 979-10-231-2163-6

Guillot, Heiden & Lavrentiev – 979-10-231-2164-3

Petit – 979-10-231-2165-0

Masià – 979-10-231-2166-7

Richards – 979-10-231-2167-4

Maquette initiale : Compo-Méca (64990 Mouguerre)

Réalisation : Emmanuel Marc Dubois/3d2s

SUP

Maison de la Recherche

Sorbonne Université

28, rue Serpente

75006 Paris

Tél. (33) 01 53 10 57 60

sup@sorbonne-universite.fr

sup.sorbonne-universite.fr

À la recherche des communautés discursives au Moyen Âge : un regard numérique sur la connectivité dans la culture vernaculaire et le rôle des traductions dans l'évolution de la prose en moyen français

Earl Jeffrey Richards
Bergische Universität Wuppertal

Depuis peu, l'étude numérique des textes est associée à l'idée qu'une nouvelle « technologie » de lecture, donnant lieu à une nouvelle visualisation de la production littéraire, s'est installée, non sans controverse¹. Une lecture « distante » (où la distance reste à définir) se réaliserait là à une plus grande échelle que celle de la critique traditionnelle. Dans toutes les études stylistiques menées depuis Leo Spitzer et Charles Bruneau, la détermination de l'« écart », qui produit des « effets de style sur fond de langue », reste une question centrale, qui présuppose une communauté discursive sans pour autant la démontrer². Avec la lecture à distance, il ne s'agit plus d'étudier cinq ou dix ouvrages de manière très focalisée – d'en réaliser une *microlecture* –, mais plutôt de découvrir grâce à une analyse algorithmique de la fréquence lexicale dans des corpus d'une centaine de textes l'existence d'affinités linguistiques entre les textes évalués : en d'autres termes, de procéder à une *macrolecture*. Non seulement ces deux modes de lecture doivent désormais être considérés comme complémentaires, mais leur articulation est réputée obligatoire. La problématique que nous aborderons ici concerne

1. Voir Franco Moretti, 2013.

2. L'analyse de Nicole Gueunier, soumise en 1969, reste toujours pertinente.

l'identification de certaines communautés discursives en France à la fin du Moyen Âge, dans la perspective de décrire comment ces communautés ont contribué à l'évolution de la prose, à travers des traductions du latin en français ou par l'élaboration d'œuvres originales.

Quelques éléments des méthodes appliquées méritent d'être précisés : l'analyse d'une éventuelle évolution de la prose médiévale telle qu'elle est ici proposée se fonde sur une base de données constituée de 150 textes, composés entre 1150 et 1450, tirés d'un corpus de textes en ancien et moyen français plus conséquent que mon collègue David Wrisley (New York University-Abu Dhabi) et moi-même avons numérisé, depuis quatre ans³. La plupart de ces 150 textes (qui comptent plus de 10 millions de mots) sont soit des traductions en prose sélectionnées parmi tous les genres possibles (œuvres littéraires, scientifiques, médicales, théologiques, juridiques, etc.), soit des œuvres originales en prose contemporaines aux traductions choisies. Les textes sont analysés à l'aide du logiciel StyloR : cette procédure produit dans un premier temps, après plusieurs rééquilibrages des critères statistiques, une architecture qui avance par tâtonnements, plusieurs partitionnements de données dont il faut calculer le « consensus » statistique en recourant à une seconde méthode, dite de *bootstrap consensus*. Cette opération génère par la suite un fichier Excel qui représente le « consensus » entre les échantillonnages, déjà exécutés, des partitionnements des données. Ce fichier Excel peut être à son tour visualisé dans un troisième graphique, une projection de réseau Gephi (version « beta 0.8.2 »), à partir de plusieurs algorithmes. Gephi produit d'abord un graphique brut sans algorithme qui ressemble à une jungle confuse de vecteurs. Il faut ensuite calculer la modularité des données, qui détermine le partitionnement des nœuds en communautés, présentant des arêtes intracommunautaires

3. Afin de constituer notre corpus, nous transformons les fichiers PDF en fichiers TXT UTF-8, au moyen d'un logiciel de reconnaissance optique de caractères (ABBYY Fine Reader 12). Après la conversion au format.doc(x), il est souvent nécessaire de corriger le texte généré. Ce corpus de 150 textes comprend des fichiers pour la plupart corrects ou corrigés.

« épaisses » et des arêtes intercommunautaires faibles. L'épaisseur des arêtes représente l'affinité entre deux œuvres en fonction de la fréquence des mots qu'elles partagent, et non par rapport aux thèmes qu'elles auraient en commun. Le partitionnement des données établit les accords lexicaux entre les textes, et non les variantes selon la méthode lachmannienne : il s'agit bien de réhabiliter une méthode d'analyse rejetée dès l'origine par les disciples du philologue allemand⁴.

Premier constat : à l'exception de la traduction anonyme de Gratien et des traductions d'Aristote effectuées par Nicolas Oresme, il est en général impossible de distinguer un texte traduit d'un ouvrage original à la seule aune de l'emploi des mots les plus fréquents. Autrement dit, la plupart des traductions ne sortent pas du lot, ce qui ne veut pas dire que les contextes de traduction n'aient pas exercé une influence importante sur l'évolution de la prose. Un des acquis les plus frappants de la projection Gephi est ainsi l'importance inattendue, dans le développement de la prose, d'une communauté qui comprend les traductions de Jean de Meun, la traduction anonyme du *Miroir des Dames*⁵ et des traductions de Jean de Vignay, et la traduction du *Policraticus* de Jean de Salisbury exécutée par Denis Foulechat. Deuxième constat : la diversité lexicale en général, sauf dans le cas d'Oresme, ne semble pas avoir de rapport avec le fait qu'un texte est une traduction ou un ouvrage original.

Revenons à notre sujet : en l'occurrence, la partition des 150 textes analysés a produit 9 communautés avec un score de modularité de 0,716 (un score supérieur à 0,3 est considéré comme significatif). Ce calcul effectué, reste une question aussi pratique qu'« artisanale » : le choix de l'algorithme permettant de représenter les communautés. Il faut souligner que les postulats sur lesquels repose toute visualisation sont par essence heuristiques. Le défi consiste dès lors à trouver un algorithme

4. Voir Paolo Trovato, 2014.

5. L'importance de cette traduction a été récemment rétablie par une équipe de chercheurs de Monash University (Melbourne). Leurs résultats préliminaires sont publiés dans *Virtue Ethics for Women*, 2011 ; se référer en particulier aux contributions de Rina Lahay, Constant Mews et Karen Green.

qui représente des distances calculées, tout en permettant de distinguer clairement les nœuds individuels sans fausser les écarts intra- et intercommunautaires. On n'a pas affaire ici à des communautés imaginaires, mais à des communautés statistiquement déterminées, dont le caractère empirique précis reste une question ouverte. En l'occurrence, une microlecture aidera à contextualiser les communautés discursives mises au jour par une macrolecture.

Les deux grands avantages que présente StyloR tiennent d'abord à ce qu'il facilite la comparaison des textes édités selon des principes différents, et ensuite qu'il permet les comparaisons même avec les fichiers dits « sales », c'est-à-dire les fichiers mal corrigés. Cette dernière caractéristique pourrait bien faire horreur aux philologues – et à moi également, qui ai consacré trois années au déchiffrement d'une leçon difficile chez Christine de Pizan: on a lu pendant quarante ans la description de la fortification de la Cité des Dames comme « bastides don[é]es et vraies », alors qu'il s'agit très simplement de « bastides, douves et braves [= *palissades*] », une configuration de fortification urbaine traditionnelle au Moyen Âge. Néanmoins il faut admettre que la précision d'une microlecture guidée par la philologie diffère fortement de celle offerte par une macrolecture suivant la statistique, laquelle, sans être particulièrement gênée par un corpus « contaminé », envisage une marge d'erreur de 15 %⁶. Les résultats suggestifs portant sur la connectivité entre les textes analysés doivent servir à susciter de nouvelles questions. Le fait qu'un partitionnement détecte une communauté ne signifie pas qu'il s'agisse automatiquement d'une communauté discursive.

Commençons avec la question générale des traductions du latin en ancien et en moyen français, en premier lieu parce que nous connaissons très souvent le nom du traducteur et du dédicataire, c'est-à-dire que nous avons des indices d'une éventuelle communauté. Dans l'introduction de l'étude, aussi monumentale que magnifique, des traductions médiévales qu'il

6. Voir Maceij Eder, 2013.

a dirigée, Claudio Galderisi⁷ distingue cinq « grands moments » dans ce processus de *translatio* : le moment anglo-normand du ^{xiii} siècle, le moment des romans antiques, le moment « Philippe le Bel », le moment « Charles V » et le moment des premiers humanistes français (t. 2/1, p. 54). On peut rebaptiser le moment « Philippe le Bel » en lui accolant le nom de Jeanne de Bourgogne, la femme de Philippe VI, qui a commandité entre 1330 et 1350 plusieurs traductions de Jean de Vignay.

En remarquant que « souvent les balbutiements d'une langue naissante se confondent dans ces œuvres fragmentaires [*Serments de Strasbourg, Eulalie*] avec des rémanences morphologiques de la langue-mère », Galderisi souligne le rôle de l'interface latin/vernaculaire comme substrat ou adstrat dans l'éclosion du français écrit. Autrement dit : pouvait-on ou non constater une continuité dans le développement de la prose ? Et quel rôle jouait la coexistence latin/vernaculaire dans ce processus ?

Tentons de considérer de manière plus concrète l'interface latin/vernaculaire, puisqu'elle acquiert au cours des siècles de plus en plus d'importance dans l'évolution des communautés discursives. Deux anecdotes rapportées par Christine de Pizan en illustrent la pertinence. Vers 1405, dans le prologue du *Livre de la Prod'homme de l'homme*, elle décrit une scène apparemment quotidienne à la cour de son époque. Elle remarque l'aisance avec laquelle s'exprime en latin le frère cadet du roi, Louis d'Orléans : « Et je vous ouioe descrire tant bien et notablement, allegant a propos auctoritez saintes, tant a preuves vraies, comme legiste de la prodommie du noble et vertueux homme » (Vatican Reg lat. 1238, fol. 2r^o). Observation précieuse : ce prince n'affecte pas de parler latin, il parle couramment cette langue. Louis semble avoir beaucoup profité de la culture latinophile et latinisante qui caractérisait la cour de son père Charles V, dont la maîtrise du latin fut moyenne, comme l'observe Christine de Pizan dans un chapitre fascinant (III, ch. 12) dédié aux intérêts intellectuels du

7. Voir Claudio Galderisi, 2011.

sage roi : « pour ce que peut-estre n'avoit le latin pour la force des termes soubtilz si en usage comme la lengue françoise, fist de theologie translater plusieurs livres⁸ ».

Ce milieu humaniste dans lequel un prince cite simultanément les autorités ecclésiastiques et juridiques reflète une situation diglossique distinctive, caractérisée par une interface complémentaire entre le latin et le français, dans laquelle le latin n'est plus un substrat mais un adstrat. À la différence de la plupart des traducteurs analysés ici, Nicole Oresme, Évrart de Trémaugon, Jean Courtecuisse, Jean de Montreuil et Jean Gerson écrivaient en français et en latin, et leurs écrits en français – œuvres originales et traductions – se regroupent au sein d'une communauté discursive qui se distingue des textes produits par la majorité des écrivains de la cour, généralement monolingues. Cette situation ne fut pas la règle à la cour royale, même si le latin fut sa langue officielle ; dans la dédicace à Philippe le Bel qui ouvre sa traduction de Boèce, Jean de Meun dépeint un roi nettement moins compétent en latin que Louis d'Orléans : « Ja soit ce que tu entendes bien le latin, mais toutevois est de moult plus legiers a entendre le françois que le latin⁹. »

On regarde souvent la « couche latine » dans la prose médiévale comme un vestige inerte et infécond, à l'image du Pilier des Nautes, enterré sous le maître-autel de Notre-Dame, redécouvert des siècles plus tard et objet d'une admiration qui n'y distingue pas le syncrétisme entre christianisme et traditions païennes – exemple même d'une interface religieuse parallèle à l'interface latin/vernaculaire impliquée dans toute traduction. Reprenons l'ancienne question de l'importance de la latinité sous-jacente dans les *Serments de Strasbourg*. On a, par exemple, toujours évoqué les formules latines présentes en creux dans les serments, dont en 1935 Alfred Ewert a justement proposé une traduction latine : il interprète le segment « pro Deo amur et pro christian poblo et nostro commun saluament »

8. Christine de Pizan, 1936-1940, v. 2, p. 43. Voir également : Serge Lusignan, 1987 et Thelma Fenster, 1998, p. 91-107.

9. Se reporter à Venceslas Louis Dedeck-Héry, 1952, p. 165.

par « *ad Dei voluntatem et a populi christiani* », mais sa reconstruction est fautive¹⁰. Si on traduit littéralement « por Deo amur » en latin comme « *pro Dei amore* », et qu'on insère cette formule dans le moteur de recherche de la *Patrologia latina* et de la *Library of Latin Texts* de Brepols, on constate que la phrase française adapte directement la formule latine « *pro Dei amore et pro illorum salute* », qui invoque l'amour de Dieu et le salut, attestée pour la première fois chez Césaire, évêque d'Arles au VI^e siècle¹¹. Cette « piste Brepols », comme on désigne cette application du numérique aux études consacrées à Christine de Pizan que j'ai développée en collaboration avec Liliane Dulac, révèle souvent une latinité à peine sous-jacente au vernaculaire. Elle pose le problème épineux de savoir s'il s'agit là de survivances inconscientes ou de termes (ou emprunts) érudits conscients. Continuons avec les *Serments de Strasbourg* : « *nostro commun saluament* » fait allusion à l'expression « *pro communi salute* », apparemment d'origine cicéronienne, mais attestée chez Augustin, Césaire d'Arles et Léo le Grand (ces deux derniers sont contemporains de l'auteur des *Serments*), et même plus tard chez Thomas d'Aquin ; mais, à la différence de l'auteur des serments allemands, qui emploie le vocable *gehaltnissi* (lequel se traduit comme *salus*), l'auteur des serments français a bel et bien distingué entre *salus* et *salvamentum* : *salus* désigne le salut au sens religieux¹², tandis que *salvamentum*, terme utilisé exclusivement en moyen latin et souvent attesté dans la *Patrologia latina*, indique non seulement le salut du Christ mais surtout, comme le montrent les exemples cités dans

10. Alfred Ewert, 1935, p. 16-35.

11. Caesarius Arelatensis, *Sermones Caesarii uel ex aliis fontibus hausti* (CPL 1008), SL 104, sermo : 189, cap. 4, l. 16 : « *Quod si essent aliqui, qui pro Dei amore et pro illorum salute* » ; *Epistularium Guiberti* : *Epistulae Guiberti*, epist. 52, l. 441 : « *Nec uero hec ita dico, ut dissuadeam uel retraham quoslibet, qui opus omni laude et benedictione dignum et toti ecclesie gratissimum facerent, qui pro Dei amore et salute animarum suarum* » ; Guillelmus Alvernus, *Sermones de communi sanctorum et de occasionibus*, sermo : 61 (*De uno confessore*), p. 217, l. 3 : « *Bonus latro significat illos qui a dextris pendent, id est qui pro Dei amore et animarum salute laborant* ».

12. Voir, par exemple, *verbum salutis*, Act. 13, 26 ; Rom. 10, 1 ; 13, 11 ; mais une concurrence entre *salus* et *salutatio* se dessine très tôt, qui explique peut-être pourquoi l'auteur des serments choisit *salvamentum* afin d'être plus précis.

le Du Cange, une « protection » dans le sens politique¹³. La différence entre *salus* et *salvamentum* est un exemple classique de « l'écart » stylistique. Pour le dire autrement, grâce à cette « piste Brepols », on peut constater précisément la survivance des formules légales qui semble accuser, mais faiblement, l'existence d'une communauté discursive documentée par des bribes aussi éloquents que fragmentaires, communauté dont Paul Zumthor a déjà envisagé l'existence¹⁴.

Revenons-en maintenant à l'analyse stylométrique. On voit ici la projection Gephi déjà évoquée, qui distingue 9 communautés discursives dans les oeuvres médiévales considérées. Certains regroupements sont anticipés, mais plusieurs éléments moins attendus sont révélés. De manière générale la projection montre une progression chronologique, à quelques exceptions près, qui mériteraient un commentaire plus détaillé.

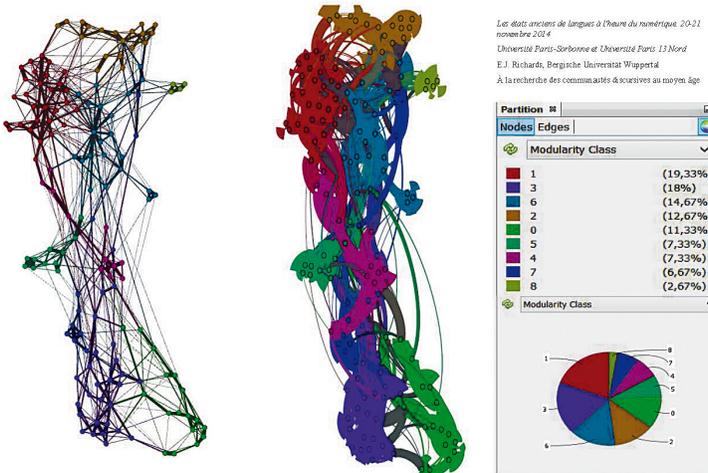


Fig. 1. Les 9 communautés discursives (projection Gephi)

13. « SALVAMENTUM: tutela, immunitas, protectio [...] præstatio a tenentibus facta dominis, pro tutela ac protectione personarum ac rerum suarum », dans Du Cange et al., 1883-1887, t. 7, col. 289c, en ligne : <http://ducange.enc.sorbonne.fr/SALVAMENTUM1>.

14. Paul Zumthor, 1959, p. 211-233 ; voir aussi Konrad Ewald, 1964, p. 35-55.

Le groupe qui apparaît en vert (n^o – **fig. 1**) tout en bas du réseau représente les œuvres anglo-normandes et franco-italiennes, souvent dites franco-vénitiennes, pour marquer l’origine de ces textes issus du Nord de l’Italie – les écrivains franco-italiens provinrent aussi du Nord : c’est le cas de Brunetto Latini (qui n’apparaît d’ailleurs pas dans cette communauté), Nicola da Verona, Rusticiano da Pisa, etc. La surprise, ici, tient au regroupement même des deux dialectes géographiquement si éloignés l’un de l’autre. Ils sont habituellement étudiés séparément, bien qu’on ait toujours commenté la présence des picardismes dans le franco-italien¹⁵. Le fait que les deux dialectes appartiennent à la même communauté discursive induit la possibilité que les Normands du royaume des Deux Siciles aient été les principaux responsables de la diffusion du français en Italie. (L’affinité linguistique que les Italiens du Nord écrivant en franco-italien entretiennent avec l’anglo-normand peut s’expliquer aussi par les affinités politiques, d’ailleurs très compliquées, existant entre les Angevins du royaume anglo-normand et les Guelfes de l’Italie septentrionale.)

15. Je cite François Avril : « Les traces picardes [...] relevées dans les copies italiennes d’œuvres françaises s’expliquent, me semble-t-il, par la place prépondérante occupée par la librairie picarde (ce terme englobant toute la production des provinces septentrionales de la France) dans la diffusion des textes français depuis au moins la première moitié du XIII^e siècle. [...] Une autre explication de ces picardismes pourrait être aussi l’origine picarde de certains copistes travaillant en Italie » (extrait d’une correspondance privée datée du 17 mars 1978, publiée dans Earl Jeffrey Richards, 1981, p. 15).

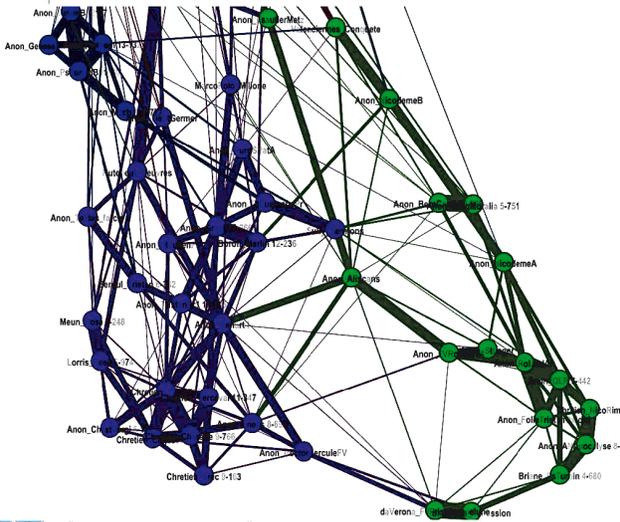


Fig. 2. Communautés n°0 et 3

Le groupe de couleur violette situé au bas de la figure (n° 3 – fig. 2) associe d’abord des textes traditionnellement associés à la littérature courtoise: l’*Eneas*, les romans de Chrétien de Troyes, le *Tristan* de Béroul, *La Châtelaine de Vergi*, le *Tristan en prose* et la *Queste del saint Graal* – mais aussi le *Roman de la Rose*, les œuvres de Rutebeuf, le *Roman de Renart*, l’*Ovide moralisé*, deux œuvres franco-italiennes (l’anonyme *Roman d’Hector et Hercule* du *xiv^e* siècle, et le *Milione* de Marco Polo), et finalement la traduction des *Sermons* de Maurice Sully. Il est remarquable et également inattendu de constater que les arêtes intracommunautaires qui relient les deux romans chevaleresques et les romans de Chrétien de Troyes sont plutôt fines. Cette communauté suscite de nombreuses interrogations touchant à notre compréhension de la littérature courtoise. On songerait à un élément d’oralité qui serait commun à tous ces textes, mais c’est une hypothèse qui mérite plus d’attention, compte tenu du fait surtout que les chansons de geste ne montrent pas le moindre rapport avec les romans courtois, un lien qu’on aurait pu s’attendre à trouver si on accepte la thèse proposée par Erich Köhler au sujet de l’existence d’une « épopée courtoise »

(*höfische Epik*). Malheureusement l'affinité entre ces ouvrages, définie selon la fréquence des mots qui les composent, ne dit rien quant à la culture rhétorique profonde ni de la *Chanson de Roland*, qui selon Ernst Robert Curtius aurait subi une influence prononcée de Virgile, ni de Chrétien de Troyes, culture mise en évidence par Danièle James-Raoul dans son étude *Chrétien de Troyes. La griffe d'un style*¹⁶.

Si l'on considère à présent la figure suivante, le groupe apparaissant en bleu marine (n° 7 – **fig. 3**) qui « serpente » dans la projection rassemble dix ouvrages dont la parenté est difficile à expliquer : la « langue du serpent » comprend la traduction des *Macchabées*, puis sa « tête » quatre œuvres très proches, *Li Fet des Romains* (1213/14), la traduction de la Genèse datée du XIII^e siècle, deux versions du *Pseudo-Turpin* ; et sa « queue » est constituée d'une œuvre en vers, *La Mappemonde* de Pierre de Beauvais (1184-1218), puis de quatre traductions en prose : le *Livre des moralitez* (1275), une traduction de la *Moralium dogma philosophorum* de Guillaume de Conches, *L'information des princes* (1282), une traduction par Henri de Gauchy du *De regimine principum* de Gilles de Rome, une traduction anonyme, datée de 1314, de la *Chirurgie* de Henri de Mondeville, chirurgien de Philippe le Bel ; et enfin *Les Livres du roy Modus et de la reine Ratio*, attribué à Henri de Ferrières (1354-1376), dont le premier, le *Livre des deduis*, est un traité de chasse et le second, le *Songe de pestilence*, une critique allégorique des vices et malheurs du temps. D'une certaine manière, il est aussi réjouissant de constater qu'une arête relativement épaisse lie cet ouvrage au *Livre de la Chasse* de Gaston Phébus (1387) qui se trouve très proche, mais dans une autre communauté.

16. Étude parue chez Champion (2007).

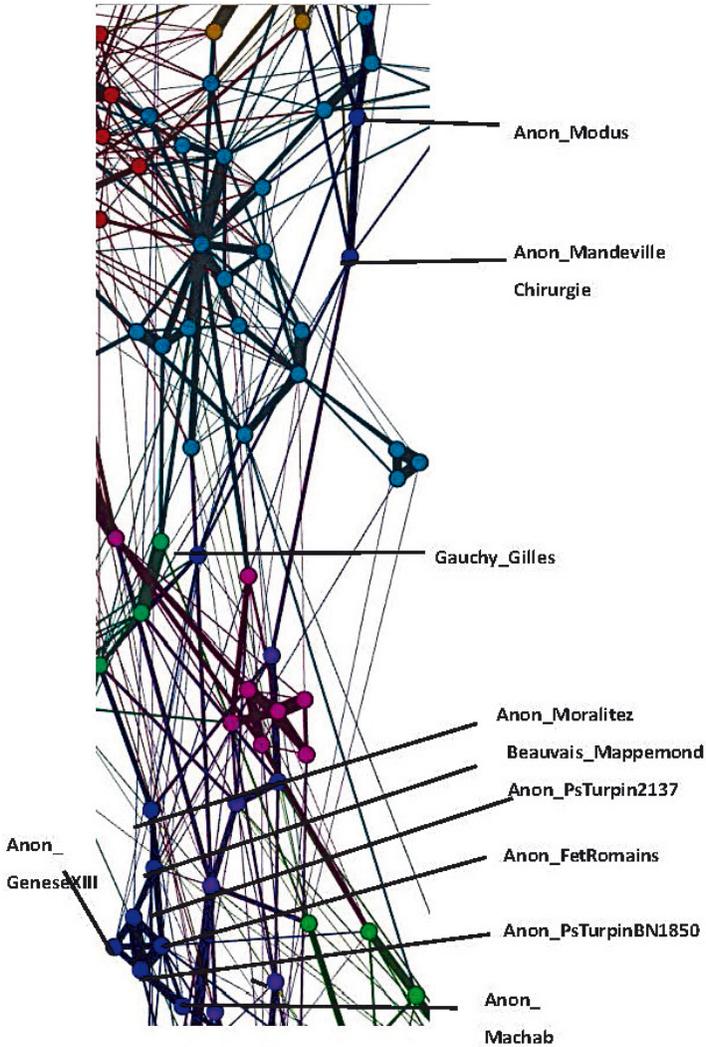


Fig. 3. Communauté n°7

Au milieu du graphique suivant apparaissent deux communautés bien distinctes. En magenta (n° 4 – fig. 4), en haut de la projection, se regroupent les ouvrages de Villehardouin, de Robert de Clari, de Philippe de Novare, le roman anonyme du Templier de Tir et, voilà la surprise présentée par ce groupe, les *Chroniques* de Froissart. Est-il possible que Froissart ait

Par exemple, dans la communauté n° 1 (fig. 5), qui rassemble plusieurs traductions exécutées à la cour de Charles VI et de nombreuses œuvres de Christine de Pizan, se trouvent aussi les *Grandes Chroniques de France*, dont on conserve plus de 700 manuscrits. Il en va de même pour le *Roman de la Rose*, dont on conserve plus de 300 manuscrits : le recours à la stylométrie ne permet pas de montrer l'influence importante de cet ouvrage, ou plutôt la stylométrie semble indiquer que l'influence de la *Rose* fut vraisemblablement plutôt thématique que linguistique, distinction jusqu'ici en grand partie ignorée dans les études littéraires.

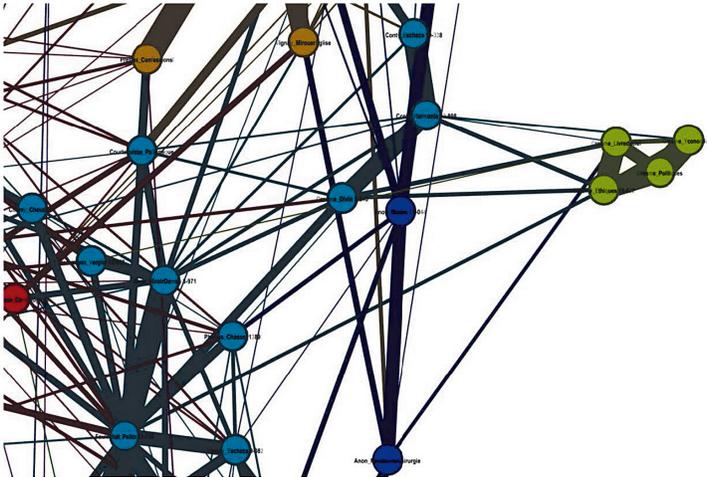


Fig. 6. Communautés n° 6 et 8

Dans le cas de la communauté n° 6 (fig. 6), représentée en bleu horizon, il s'agit de la période « Philippe le Bel et Jeanne de Bourgogne », jusqu'à la cour de Charles V. En réalisant cette projection, j'ai par erreur inclus dans les 150 textes du corpus considéré trois textes à la thématique identique, puisque consacrés à saint Louis. Dans la projection on les retrouve ensemble, de manière séparée des autres. Mais j'avais de fait numérisé les *Miracles de saint Louis* de Guillaume de Saint Pathus à deux reprises successives, l'une en les recensant

comme ouvrage anonyme, une faute d'inattention que j'ai découverte en contrôlant la projection Gephi – exercice aussi nécessaire que salutaire. Cependant, la projection montre l'identité entre le fichier « anonymisé » de manière erronée et le fichier « authentique », qui, tous les deux, sont rapprochés de la biographie de saint Louis par Joinville. On peut dater ces deux œuvres authentiques de manière très précise : celle de Guillaume de Saint Pathus, commanditée par Blanche de France, de 1297, et celle de Joinville, présentée à Louis le Hutin, de 1309. En même temps, deux autres ouvrages semblent être d'une influence primordiale au sein de cette communauté : *Les Évangiles des domées* (« Les évangiles des dimanches »), datés de 1235 mais qui, apparemment, n'ont aucun lien avec la Bible du XIII^e siècle, et *Le Livre de Sydrac le philosophe*, composé après 1268. Or, ces quatre œuvres, toutes « en marge » de la communauté, comme deux traductions de Jean de Meun, plusieurs traductions de Jean de Vignay et la traduction anonyme du *Miroir des Dames*, simplement datée du début du XIV^e siècle, sont néanmoins liées par des arêtes épaisses au *Policratique* de Denis Foulechat (1372). Sa position centrale n'indique pas son influence sur les autres textes, mais plutôt le fait que son écriture représente un « consensus » linguistique, voire une « norme stylistique » pour cette communauté. La question centrale ici n'est pas l'importance de l'influence des œuvres antérieures de Foulechat dans l'éclosion du style du discours à la cour royale pendant presque quatre-vingts ans, c'est le simple fait que l'analyse stylométrique identifie un style particulier déterminant le développement de la prose française.

La fécondité du style de cette communauté se révèle dans ses rapports avec trois autres communautés, qui semblent en émerger : les communautés n°1 (voir **fig. 5**), modélisée en brun-rouge, n°2 (**fig. 7**) en beige foncé et n°8 (voir **fig. 6**) en chartreuse ; toutes trois plus ou moins contemporaines et toutes trois associées à la cour royale de Charles V et Charles VI. Certains auteurs sont groupés dans plus d'une communauté : cet « entrecroisement », ou cette « coopération » pourrait représenter

l'indice d'une certaine continuité dans la production littéraire autour de la cour royale à partir de 1300. Dans la communauté n°1, on trouve des textes en prose plus ou moins contemporains, des œuvres en vers et en prose d'Eustache Deschamps, les écrits de Philippe de Mézières, d'Honoré Bouvet et de Christine de Pizan, et les traductions de Laurent de Premierfait – mais aussi, ce qui surprend, *Renart le Contrefait* (1328-1342) et deux romans chevaleresques en prose: le *Conte de la Charrette* (version divergente de la Vulgate) daté du XIII^e siècle et le *Perceforest* de la première moitié du XIV^e, une constellation qui devrait faire l'objet de recherches futures. L'autre anomalie, ici, mais qui aide un peu à expliquer les caractéristiques de cette communauté, est la présence de la version du *Roman de la Rose* de Clément Marot, publiée en 1527. Dans les projections préliminaires, ce texte, par erreur étiqueté comme « Lorris_Rose », s'est toujours néanmoins retrouvé dans cette communauté, un résultat surprenant, qui a exigé une nouvelle inspection du fichier. Il s'est avéré qu'il était l'un des premiers textes numérisés, quatre années plus tôt. Par la suite, l'édition de Lecoy a été numérisée, mais l'étiquette de ce fichier n'a malencontreusement pas été corrigée. Après sa correction, suivie d'une nouvelle projection, la présence de cette réécriture de la *Rose* dans la communauté considérée ici semble répondre à sa propre logique, une logique compréhensible. Autrement dit, et c'est là une conclusion provisoire, les caractéristiques propres à cette communauté semblent signaler une tendance « vulgarisante » quoiqu'érudite; mais non pas aussi latinisante que celle qui définit la communauté n°2 voisine, dominée par les œuvres françaises de Jean Gerson, mais qui comprend aussi les sermons de Jean Courtecuisse, le *Grand Coutumier de France* de Jacques d'Ableiges, ancien secrétaire de Jean de Berry, des œuvres composées pour la cour de Berry (*Mélusine* de Jean d'Arras) et pour la cour de Bourgogne (les deux versions de *l'Erec en prose*). La plus grande surprise qui naît face à cette communauté tient à la présence de Marguerite Porete. Au cours de plusieurs projections préliminaires, son *Miroir des simples ames* s'est toujours retrouvé à proximité des œuvres de Gerson. Face à ce résultat j'ai ajouté au corpus, en guise de

test, la *Montaigne de Contemplation* de Gerson (détachée de ses œuvres, analysées en un seul fichier), puisque la piété féminine est le thème central de cet ouvrage, et un sujet sur lequel Gerson prend un parti nettement opposé à celui défendu par Marguerite Porete. Malgré leurs positions contraires, la dernière projection montre que l'affinité entre ces deux œuvres, au regard de la fréquence des mots qui les composent, est encore plus accusée que celle présente entre les autres œuvres de Gerson et le *Miroir des simples ames*. La critique a toujours remarqué l'érudition dont faisait preuve Marguerite Porete: en voici une représentation très claire.

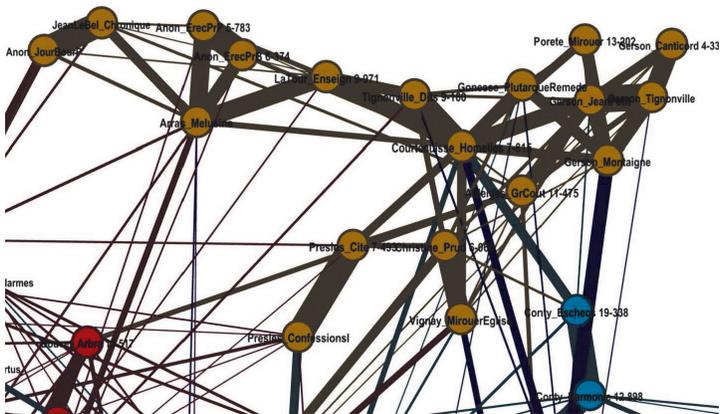


Fig. 7. Communauté n° 2

La projection Gephi révèle aussi d'étonnants moments de discontinuité: les traductions de la Bible et les plus anciennes mises en prose des romans chevaleresques ne semblent pas avoir exercé une influence sur le développement de la prose. D'autre part, on observe une éclosion véritable de la prose dès la fin du XIII^e siècle, surtout sous l'impulsion de la cour royale, et c'est à partir de ce moment que l'on observe une continuité et un véritable élan dans l'évolution de la prose française.

Références bibliographiques

- AINSWORTH, Peter, « The Online Froissart ». En ligne : <http://www.hrionline.ac.uk/onlinefroissart/apparatus.jsp?type=intros&intro=f.intros.PFA-Froissart>
- CHRISTINE DE PIZAN, *Le Livre des fais et bonnes meurs du sage roy Charles V*, éd. Suzanne SOLENTE, Paris, Champion, 1936-1940.
- DU CANGE *et al.*, *Glossarium mediae et infimae latinitatis*, éd. augm., Niort, L. Favre, 1883-1887. En ligne : <http://ducange.enc.sorbonne.fr/>.
- DEDECK-HÉRY, Venceslas Louis, « Boethius' *De Consolatione* by Jean de Meun », *Mediaeval Studies*, n° 14, 1952, p. 165-275.
- EDER, Maceij, « Mind Your Corpus: Systematic Errors In Authorship Attribution », *Literary and Linguistic Computing*, n° 28, 2013/4, p. 603-614.
- EWALD, Konrad, « Formelhafte Wendungen in den Straßburger Eiden », *Vox Romanica*, n° 23, 1964, p. 35-55.
- EWERT, Alfred, « The Strasburg Oaths », *Transactions of the Philological Society*, 1935, p. 16-35.
- FENSTER, Thelma, « *Perdre son latin*: Christine de Pizan and Vernacular Humanism », dans DESMOND, Marilyn (dir.), *Christine de Pizan and the Categories of Difference*, Minneapolis, University of Minnesota Press, 1998, p. 91-107.
- GALDERISI, Claudio (dir.), *Translations médiévales. Cinq siècles de traductions en français au Moyen Âge (XI^e et XV^e siècles)*, 2 vol., Turnhout, Brepols, 2011.
- GREEN, Karen, « From *Le Miroir des Dames* to *Le Livre des trois vertus* », dans GREEN, Karen et MEWS, Constant (dir.), *Virtue Ethics for Women*, New York, Springer, 2011.
- GREEN, Karen et MEWS, Constant (dir.), *Virtue Ethics for Women*, New York, Springer, 2011.
- GUEUNIER, Nicole, « La pertinence de la notion d'écart en stylistique », *Langue française*, n° 3, 1969/3, p. 34-45.

- LAHAY, Rina, « A Mirror of Queenship: The *Speculum dominarum* and the Demands of Justice », dans GREEN, Karen et MEWS, Constant (dir.), *Virtue Ethics for Women*, New York, Springer, 2011.
- LUSIGNAN, Serge, *Parler vulgairement. Les intellectuels et la langue française aux XIII^e et XIV^e siècles* [2^e éd], Paris, Vrin, 1987.
- MEWS, Constant, « The *Speculum dominarum*/Miroir des dames and Transformations of the Literature of Instruction for Women in the Early Fourteenth Century », dans GREEN, Karen et MEWS, Constant (dir.), *Virtue Ethics for Women*, New York, Springer, 2011.
- MORETTI, Franco, *Distant Reading*, London, Verso, 2013.
- RICHARDS, Earl Jeffrey, *Dante and the « Roman de la Rose »: An Investigation into the Vernacular Narrative Context of the « Commedia »*, Tübingen, Niemeyer, 1981.
- TROVATO, Paolo, *Everything You Always Wanted to Know about Lachmann's Method. A Non-Standard Handbook of Genealogical Textual Criticism in the Age of Post-Structuralism, Cladistics, and Copy-Text*, Padova, Liverariauniversitaria.it, 2014.
- ZUMTHOR, Paul, « Une formule gallo-romane du VIII^e siècle », *Zeitschrift für romanische Philologie*, n° 75, 1959/3-4, p. 211-233.

Résumés / Abstracts

Sylvie BAZIN-TACHELLA et Gilles SOUVAY,
De la gestion de la variation en moyen français à
son élargissement aux états anciens du français :
le développement du lemmatiseur LGeRM

Résumé

La langue médiévale ne se livre qu'à travers des témoignages écrits, essentiellement mouvants et variants. Le *Dictionnaire du moyen français*, dès ses débuts, a été confronté à cette difficulté. La lemmatisation des vedettes a été nécessaire pour construire la base de données et un outil, le lemmatiseur LGeRM (acronyme de « Lemmes, Graphies et Règles Morphologiques »), a permis de faire du DMF un dictionnaire véritablement électronique, à la fois dans sa conception et dans sa consultation, deux aspects différents mais liés. C'est lui qui permet d'interroger à partir de la forme rencontrée dans un document. Lors de la recherche d'une entrée dans le dictionnaire, l'analyseur isole un mot – hors contexte – et fournit des hypothèses de lemmes. Il utilise pour cela un lexique et des règles de flexion et de variation graphique. Le lexique est constitué des graphies connues avec leur analyse (graphie, lemme, étiquette). Conçu au départ pour le dictionnaire, le lemmatiseur a pu être intégré dans de nouveaux environnements. Grâce à la lemmatisation d'un texte source encodé en XML/TEI, il est possible de l'interroger par forme, ou par lemme, ou en suivant le texte en continu, ce qui est d'une aide considérable pour mener à bien la préparation d'une édition et la construction d'un glossaire. LGeRM a connu d'autres types de développements, en s'adaptant à la morphologie et aux variations spécifiques d'autres états de langue que celui pour lequel il avait été conçu, ce qui a abouti à la construction de deux lexiques distincts : un lexique LGeRM médiéval, optimisé pour la période 1300-1500 et un lexique LGeRM ^{xvi}^e-^{xvii}^e pour 1550-1700, désormais utilisés par le moteur de recherche de FRANTEXT pour

la recherche par lemme. En accès libre sur demande, LGeRM est devenu un outil d'interrogation des textes anciens, en moyen français (cible du *DMF*) et en amont et en aval de la période (ancien français et français des *xvi^e* et *xvii^e* siècles), complémentaire des outils d'étiquetage morphosyntaxique.

Abstract

Medieval language reveals itself only through diverse and unsettled written accounts. Right from the beginning, the creators of the *Dictionnaire du moyen français (DMF)* have tried to overcome this challenge. The lemmatization of the entries was necessary in order to construct the dictionary's database. The team have also used a lemmatizing tool, LGeRM (*Lemmes Graphies et Règles Morphologiques*), to create an electronic dictionary in both its conception and consultation. When an user researches an entry from the dictionary, the analyzer takes a word out of context and provides hypothesis of lemmas. In order to do this, the analyzer utilizes a lexicon and various rules of inflection and spelling variations. The lexicon is made of known written forms with their analysis (spelling, lemma, tag). The lemmatizer was firstly designed for the dictionary, but is now fit for further use. Thanks to the lemmatization of source texts encoded in XML/TEI, LGeRM can analyze an original text per forms, lemma or even pages which is of significant assistance when preparing a text edition or constructing a glossary. LGeRM has undergone other types of developments, being adapted to the morphology and specific variations of other states of language. Therefore, we now have two distincts LGeRM lexicons; one for the medieval period (1300-1500), and another one for the early-modern period (1550-1700). Both are being used by the FRANTEXT search engine for the research by lemma. LGeRM can thus be used to work on Middle French (the target of the DMF), but also on Old French as well as French of the 16th and 17th Centuries. To finish, this query tool is on open access and complementary to Morphosyntactic taggers.

Ana GÓMEZ RABAL, *Le latin médiéval du Glossarium Mediae Latinitatis Cataloniae: un projet lexicographique dans un contexte européen*

Résumé

Le *Glossarium Mediae Latinitatis Cataloniae* (GMLC), dictionnaire du latin médiéval des territoires correspondant au domaine linguistique du catalan entre le IX^e et le XII^e siècle, est réalisé grâce à la collaboration de la section de lexicographie latine du département d'Études médiévales de l'Institut Milà y Fontanals du CSIC (Consejo superior de investigaciones científicas, à Barcelone) avec le département de Lettres latines de l'université de Barcelone. Les responsables de l'élaboration et de la publication de ce glossaire ont comme objectif scientifique de fournir aux philologues, aux historiens et aux juristes, ainsi qu'à toute personne intéressée par le Moyen Âge, un outil qui rende compréhensible la documentation notariale et les textes littéraires, juridiques et scientifiques latins produits dans les lieux et à l'époque cités, textes qui sont le témoignage écrit non seulement de la langue latine médiévale, mais aussi de la langue romane naissante et dont la lecture est, très souvent, compliquée même pour ceux qui ont une certaine habitude de travailler sur des textes en latin.

Les membres de l'équipe du GMLC travaillent en deux phases indissociables et complémentaires, qui évoluent vers un objectif ultime commun : la publication complète du glossaire. La première phase, la *rédaction*, consiste en la préparation, l'élaboration et la mise à jour des articles du glossaire lui-même. Pour la seconde phase, la *numérisation*, les textes utilisés comme matière première pour l'écriture des articles lexicographiques sont passés au scanner, reconnus et corrigés ; les textes corrigés forment un corpus à usage interne qui sert aussi bien pour la rédaction des articles lexicographiques que pour les recherches parallèles des membres du GMLC. Mais cette deuxième phase a désormais comme objectif le développement et l'expansion du *Corpus Documentale Latinum Cataloniae* (CODOLCAT), base de données lexicale de publication périodique (version 1,

en 2012 ; version 2, en 2013 ; version 3, en 2014 ; version 4, en 2015) qui permet l'accès, de façon libre et gratuite, au corpus textuel utilisé pour écrire le *GMLC* ; ce corpus textuel est traité, dépouillé et réédité lors de son introduction dans le CODOLCAT et, finalement, il est présenté sous forme de concordances.

La progression du travail amène l'équipe du *GMLC* à se confronter au défi de l'édition au format numérique du glossaire lui-même. Comme il en va pour les autres dictionnaires de latin médiéval – pour ceux qui sont en cours de publication autant que pour l'ancien Du Cange –, la publication numérique et en ligne s'impose. Le groupe s'est donc engagé, désormais, dans la préparation du balisage en langage XML des articles déjà rédigés. Le projet de publication en ligne des articles déjà publiés sur papier, et des articles futurs des autres lettres encore à rédiger, doit permettre une diffusion maximale de l'œuvre et rendre service aux chercheurs.

Abstract

The *Glossarium Mediae Latinitatis Cataloniae (GMLC)*, dictionary of Medieval Latin from the territories corresponding to the linguistic area of the Catalan from ninth to twelfth centuries, is realised through the collaboration between two institutions: the Department of Medieval Studies of Milá y Fontanals Institution (CSIC, Barcelona) and the Department of Latin Philology of the University of Barcelona. The developers of the glossary have the scientific purpose of providing philologists, historians and jurists, as well as anyone interested in the Middle Ages, a tool that makes understandable the Latin notarial documentation and the Latin literary, legal and scientific texts produced in the mentioned territories and centuries. All these acts and texts are the written testimony not only of the Medieval Latin language but also of the emerging Romance language, and whose comprehension is very often complicated even for those who have a certain habit of reading and working on texts in Latin.

The *GMLC* team divides and shares their functions between two lines of work, inseparable and complementary, which evolve

towards a common ultimate goal: the complete publication of the glossary. The first line is called *writing* and consists of the preparation, development and updating of glossary articles itself. In the second line of work, called *digitalisation*, the texts used as raw material for writing lexicographical items are passed to the scanner, recognized and corrected; the corrected texts form a corpus to internal utilisation, which is used both for writing lexicographical articles and for parallel searches for the members of the *GMLC*. But this second line of work now aimed at the development and expansion of the *Corpus Documentale Latinum Cataloniae* (CODOLCAT), lexical database of serial publication (version 1, 2012; version 2, 2013; version 3, 2014; version 4, 2015), which provides free access to the textual corpus used to write the *GMLC*, processed, marked, re-edited and presented in form of concordances.

As a result of the increase in the working lines described, the *GMLC* team now faces the challenge of publishing in digital format the glossary itself. Just as for the other teams of Medieval Latin dictionaries – those being published and the old Du Cange as well –, the digital and online publication is essential. So, the *GMLC* group is engaged now in the preparation of XML markup of the articles already drafted. The envisioning of the online digital publishing (of articles published in paper and of articles of letters to write) is strongly encouraged to give the work the maximum dissemination and usefulness.

Michèle GOYENS et Céline SZECEL, Autorité du latin et transparence constructionnelle: le sort des néologismes médiévaux dans le domaine médical

Résumé

Dans cette contribution, nous présentons le projet de recherche *Latin authority and constructional transparency at work: Neologisms in the French medical vocabulary of the Middle Ages and their fate*, subventionné par le Fonds de la recherche de la KU Leuven (OT/14/047). Ce projet étudie les raisons pour lesquelles certains néologismes créés dans le

domaine médical au cours du Moyen Âge existent toujours en français moderne, alors que d'autres ne se maintiennent pas. Notre hypothèse de travail est que des critères morphologiques, et plus particulièrement la transparence constructionnelle, jouent un rôle crucial pour la préservation de ce lexique. En d'autres mots, les termes présentant une relation formelle proche de l'élément latin dont ils sont issus se maintiendraient mieux que des créations françaises originales, c'est-à-dire des dérivés ou des composés réalisés à partir de bases morphologiques françaises. Concrètement, nous esquissons les objectifs du projet et ses hypothèses de travail, avant de présenter le corpus numérisé de textes médicaux du Moyen Âge, comprenant des traductions françaises de textes-sources latins ainsi que des textes directement composés en français. Nous expliquons ensuite les facteurs décisifs pour la survie de ces néologismes : ces critères peuvent être externes ou internes, aussi bien d'ordre général que d'ordre morphologique, ces derniers formant la grille d'analyse pour une base de données morphologique numérique de la terminologie médicale médiévale en français, qui sera mise à la disposition de la communauté scientifique. Nous présentons en dernier lieu le cadre théorique de la morphologie des constructions (Booij, 2010), qui permettra de dégager des corrélations au niveau des structures morphologiques relevées, et terminons par une série de perspectives.

Abstract

This article gives an overview of the research project *Latin authority and constructional transparency at work: Neologisms in the French medical vocabulary of the Middle Ages and their fate*, financed by the Research Fund of the KU Leuven (OT/14/047). This project aims at investigating why certain French neologisms that emerged in the field of medicine during the Middle Ages managed to survive, while others disappeared after some time. Our hypothesis is that morphological criteria, in particular constructional transparency, contribute in a crucial manner to lexical preservation. In other words, terms showing a close formal relation with the Latin equivalent from which they

were borrowed, could stand the test of time better than original French creations, i.e. derivations or compounds on the basis of genuinely French morphemes. In this contribution, we first present the objectives of the project and its working hypotheses, before describing the digitized corpus of medieval medical texts, containing both translations from Latin and texts directly written in French. We then set out the external and internal factors decisive for the survival of these neologisms. With respect to internal factors, a first set of criteria concerns more general linguistic characteristics; a second one, the morphological characteristics of each neologism. Those internal criteria form the guiding principles that will allow us to complete an online morphological database of medieval medical French vocabulary, which will be at the disposal of the scientific community. In a last section, we present the theoretical framework of Construction Morphology (Booij, 2010), which will allow us to extract correlations between morphological structures, before concluding our article with a series of prospects.

Elisa GUADAGNINI, La lexicographie de l'Italien médiéval et les corpus de l'OVI : un bilan provisoire et quelques nouvelles perspectives

Résumé

Ce travail décrit sommairement l'histoire de l'OVI (Opera del vocabolario italiano, CNR - Firenze) et de ses projets : depuis les années 1960, ce centre de recherche travaille à la rédaction d'un vocabulaire de l'ancien italien, le *TLIO* (*Tesoro della Lingua Italiana delle Origini*), et à la constitution d'une base de données textuelles. Le Corpus OVI est aujourd'hui librement consultable sur la toile (en ligne : <http://gattoweb.ovi.cnr.it>). Il recueille plus de 23 millions de mots, et représente une ressource incontournable pour toute étude consacrée à l'italien médiéval. Le *TLIO* compte plus de 30 000 articles : lui aussi publié sur internet (en ligne : <http://tlio.ovi.cnr.it/TLIO/>), il est le principal – et le plus ancien – projet italien de lexicographie électronique.

Abstract

This work outlines the history of OVI (Opera del Vocabolario Italiano, CNR - Firenze) and its projects: since the '60s, this research center is working on compiling a dictionary of old Italian, the *TLIO* (*Tesoro della Lingua Italiana delle Origini*), and on creating a textual database. The Corpus OVI is now freely available on the web (<http://gattoweb.oivi.cnr.it>). It collects more than 23 million words and is an indispensable resource for any study of medieval Italian. The *TLIO* has more than 30,000 items: also being published on the internet (<http://tlio.oivi.cnr.it/TLIO/>), it is the main – and the oldest – Italian project of electronic lexicography.

Céline GUILLOT, Serge HAIDEN et Alexis LAVRENTIEV, Base de français médiéval: une base de références de sources médiévales ouverte et libre au service de la communauté scientifique

Résumé

L'essor actuel de la linguistique diachronique a des répercussions importantes sur le développement de ressources numériques qui soient adaptées à la recherche en langue médiévale et accessibles à une très large communauté. L'enrichissement de ces ressources a en retour une influence très forte sur les objets et les méthodologies utilisés pour l'analyse des données ainsi constituées. C'est cette synergie complexe et les implications méthodologiques qui la sous-tendent que nous tenterons d'illustrer dans cet article, grâce à l'exemple du développement de la *Base de français médiéval*. Nous commencerons par donner un aperçu des possibilités offertes par ce corpus numérique et nous présenterons la double chaîne mise en place pour permettre les recherches : chaîne philologique pour la constitution et la préparation des données textuelles, chaîne analytique pour leur exploitation outillée. Nous montrerons de quelle façon ces deux chaînes s'articulent, et les principes qui fondent leur association en vue d'un développement intégré et communautaire: usage de standards internationaux pour

la représentation des données et pour l'architecture des outils d'analyse, licences *open-source* qui permettent la diffusion, l'enrichissement et la pérennisation des ressources textuelles/logicielles et qui garantissent la reproductibilité des analyses.

Abstract

Current developments in diachronic linguistics have an important impact on the production of digital resources that become more and more adapted to research on the medieval language and accessible to a large academic community. The enrichment of these resources has in turn a very strong influence on the objects and the methodologies used to analyse the data obtained in this process. It is this complex synergy and the methodological implications that underlie it that we will attempt to illustrate in this article through the example of the development of the *Base de Français Médiéval*. We will first give an overview of the possibilities offered by this online corpus and then present the double-fold data analysis workflow: a “philological chain” for the constitution and the preparation of the textual data, and the “analytical chain” for their exploitation powered by linguistic tools. We will show how these two chains interact and the principles that form the basis of their association for integrated and community development: international standards for data representation and for tools architecture, open source licenses that allow the distribution, enrichment and long-term preservation of textual and software resources and that ensure reproducibility of the results of analysis.

Robert MARTIN, À propos du *DMF*

Résumé

Le *DMF* (*Dictionnaire du moyen français*) illustre les bénéfices que procure la lexicographie électronique; il fait prendre conscience aussi de tous les pièges qu'elle comporte: l'instabilité, une complexité informatique de plus en plus difficile à dominer, le risque de l'inexistence dans la durée.

Abstract

Das Mittelfranzösische Wörterbuch *DMF* veranschaulicht die grossen Vorteile der elektronischen Lexikografie; das Werk lässt aber auch verschiedene Schwierigkeiten wahrnehmen: die Unbeständigkeit, eine immer schwerlicher überwindbare informatische Komplexität und schliesslich auf die Dauer die Gefahr der Inexistenz.

Ramon MASIÀ, Numérisation et traitement de textes mathématiques grecs: méthodes, problèmes et résultats

Résumé

Le corpus des textes mathématiques grecs (CTMG) contient un peu plus de cent ouvrages qui ont survécu, totalement ou partiellement, depuis le IV^e siècle av. J.-C. C'est donc un corpus relativement restreint. Notre objectif est de le numériser, puis de le traiter avec les outils créés par la linguistique de corpus. D'une part, cet objectif est réalisable précisément parce que le corpus est de taille réduite, mais aussi parce qu'il ne contient presque pas d'ambiguïtés, le nombre d'occurrences du corpus restant faible et les différences de structure syntaxique peu abondantes. D'autre part, la mathématique grecque est rédigée dans une langue spécifique, que les mathématiciens eux-mêmes maîtrisaient très bien, puisque ce champ de savoir dépend entièrement du style dans lequel il a été écrit. Après avoir procédé à la numérisation des textes, nous avons lemmatisé une grande partie du corpus, puis avons procédé à une analyse comparative de différents textes et auteurs. Au cours de cette première étape, nous avons constaté qu'une telle approche quantitative dans le contexte de l'étude des CTMG était pertinente et nécessaire à la recherche consacrée aux mathématiques grecques.

Abstract

El corpus de los Textos Matemáticos Griegos (CTMG) contiene un poco más de 100 obras y abarca todas las que han sobrevivido, completa o parcialmente, desde el s. IV AC. Se trata, pues, de un

corpus relativement pequeño. Nos hemos planteado el objetivo de digitalizar dicho corpus, así como tratar el corpus digitalizado con las herramientas de la Lingüística de Corpus. Dicho objetivo, por un lado, es factible, precisamente por tratarse de un corpus pequeño, pero también porque presenta pocas ambigüedades, el número de ‘palabras diferentes’ (ocurrencias) del corpus es bajo y las estructuras sintácticas diferentes no són muy abundantes. Además, la Matemática Griega está escrita en un lenguaje muy específico, del cual los matemáticos eran conscientes, ya que en último término, y formalmente, la matemática griega depende completamente del estilo en que se escribió; la matemática griega puede identificarse con esta forma de escribirla. Después de la digitalización de textos, hemos lematizado gran parte del corpus y, posteriormente, hemos hecho análisis comparativos entre diversos textos y autores. En este primer estadio de este proceso de digitalización y análisis, hemos comprobado que este enfoque cuantitativo en el estudio del CTMG es pertinente y necesario para profundizar en la Matemática Griega.

Estrella PÉREZ RODRÍGUEZ, *Le Lexicon Latinitatis Medii Aevi regni Legionis* (VIII^e s.-1230)

Résumé

Le *Lexicon Latinitatis Medii Aevi Regni Legionis*, ou *LELMAL*, est un dictionnaire de latin actuellement élaboré en Espagne à partir d'un corpus formé par les textes écrits principalement en langue latine sur le territoire du Royaume des Asturies et de León entre le VIII^e siècle et 1230. L'objectif principal de cet article réunit deux aspects : en premier lieu, montrer la méthodologie de ce travail lexicographique et les caractéristiques externes fondamentales du dictionnaire ; en second lieu, exposer et commenter quelques exemples intéressants tirés du corpus léonais qui démontrent l'importance de l'étude lexicographique pour mieux connaître l'histoire de la langue d'un territoire. À titre d'exemples, on a choisi quatre romanismes : *uentresca*, à peine attesté en castillan avant le XVIII^e siècle ; *jera*, un mot relatif à la façon de mesurer les terres ; les adjectifs apparentés *combo* et

recombo, seulement attestés dans les sources asturiennes ; et, pour finir, la forme insolite *plentum*, inconnue en latin et résultat vraisemblablement d'une confusion du scribe médiéval (ce que nous appelons un « mot fantôme »).

Abstract

The *Lexicon Latinitatis Medii Aevi Legionis* or *LELMAL* is a Latin dictionary which is being created in Spain from the sources written mainly in Latin in the kingdom of Asturias and León between the 8th century and 1230. The twofold objective of this paper is, on the one hand, to explain the methodology of that lexicographical work and the main external features of the dictionary; on the other hand, to study some interesting examples from the sources of León which can show the important contribution of lexicographical studies to the knowledge of the history of the language of a territory. Five examples have been chosen, four vernacular words: *uentresca*, hardly found in Castilian before the 18th century; *jera*, a word in relation with land measurement, and the related adjectives *combo* and *recombo*, only used in the sources from Asturias; as well as the unique form *plentum*, a ghost-word, as it is called, because it does not exist in Latin and probably originated from a mistake of the medieval scribe.

Gérard PETIT, Terminographie diachronique: le cas de la terminologie médiévale française

Résumé

L'objectif de cet article est de prolonger la réflexion sur la description du lexique et des terminologies en diachronie, mais aussi de présenter un projet lexicographique novateur consacré au français technique et scientifique médiéval: il s'agit de CréalScience. Les présupposés attachés usuellement à la représentation du lexique postulent chez celui-ci une stabilisation des formes, des significations et des régimes syntaxiques. Si une approche en synchronie peut s'appuyer sur la permanence (même relative) des données, il n'en va pas

de même pour une description diachronique, surtout lorsque la synchronie T-1 envisagée – le Moyen Âge – constitue à elle seule une vaste diachronie. Dans cette étude nous montrerons que : (i) les réglages théoriques et méthodologiques préalables à la description sont fondamentalement tributaires de l'écart diachronique entre To et T-1; (ii) la procédure de description, demandant à être adaptée à chaque synchronie passée, ne peut permettre une modélisation de la démarche ou de ses paramètres, sauf sous forme de schémas déclinables; (iii) la notion d'état de langue constitue un objectif pour le chercheur. Elle est néanmoins facteur de risques pour la description qui veut éviter l'anachronisme.

Abstract

The objective of this contribution is to extend the reflection on the description of the lexicon and terminology diachronic, but also to present an innovative lexicographical project devoted to medieval scientific and technical French: CréalScience. Presuppositions usually attached to the lexical representation postulate in this stabilization of forms, meanings and syntactic systems. If an approach in synchrony can rely on permanently (even relative) data, the question arises for a diachronic description, particularly when considered synchrony T-1 – the Middle Ages – is in itself a vast diachronic. In this study we show that: (i) pre-theoretical and methodological adjustments to the description are fundamentally dependent on the diachronic difference between To and T-1; (ii) a description of procedure, asking to be adapted to each past synchrony can enable modeling of the process or its parameters, except as series of patterns; (iii) the concept of state language is an objective for the researcher. Nevertheless, it constitutes a degree of risk for the description aiming to avoid anachronism.

Earl Jeffrey RICHARDS, À la recherche des communautés discursives au Moyen Âge: un regard numérique sur la connectivité dans la

culture vernaculaire et le rôle des traductions dans l'évolution de la prose en moyen français

Résumé

Cette communication propose une analyse de l'évolution de la prose médiévale en français avec l'aide de quatre méthodes numériques : la « piste Brepols », la diversité lexicale calculée grâce à AntConc, la stylométrie du logiciel StyloR et la visualisation d'un réseau de communautés discursives grâce au logiciel Gephi.

Est montrée d'abord l'importance de la latinité sous-jacente dans les *Serments* de Strasbourg et la *Cantilène Sainte Eulalie*, en recourant au moteur de recherche de la *Patrologia latina* et de la *Library of Latin Texts* de Brepols, permettant de reconstruire plus précisément l'influence du latin comme substrat ou adstrat dans n'importe quel texte vernaculaire, ce qui implique l'existence d'une communauté discursive dès le IX^e siècle. La survivance des formules légales latines dans les *Serments* semble en effet montrer, mais faiblement, l'existence d'une communauté discursive documentée par des bribes aussi éloquentes que fragmentaires.

Il s'agit ensuite de savoir si les traductions commanditées dans des contextes historiques connus favorisent l'expansion du vocabulaire français. Une analyse de la diversité lexicale au moyen du logiciel concordancier AntConc, à la suite d'une conversion de traductions d'époques diverses en fichiers .txt, permet de calculer les *token/type*-ratio. Les résultats préliminaires suggèrent que la diversité lexicale présentée par les œuvres en prose est nettement plus élevée que celle des œuvres en vers, c'est-à-dire que l'expansion du vocabulaire dépend en premier lieu du choix de la prose par l'auteur. Un autre résultat important est constitué par la différence entre la diversité lexicale des traductions faites pour Philippe le Bel et celle des œuvres composées pour Charles V. Pour expliquer cette différence, les fichiers .txt de plusieurs centaines de textes ont été soumis à une analyse stylométrique StyloR. Ce logiciel combine plusieurs

fonctionnalités basées sur la fréquence des mots, et produit à la suite d'une analyse *bootstrap* un fichier Excel qui sert de base à la visualisation d'un réseau au moyen du logiciel Gephi. La communication se clôt par un commentaire sur cette mise en évidence de communautés discursives à travers trois siècles en France et une comparaison avec la littérature en prose composée en moyen anglais.

Abstract

In this contribution I present an analysis of the rise of prose in medieval French with the help of four digital methods: the “*piste Brepols*” (literally the “Brepols track”: a method which entails translating medieval French expressions into Latin and using this translation in the search engine at the online Brepols Library of Latin Texts), lexical diversity calculated on the on-line concordance program “AntConc” (<http://www.laurenceanthony.net/software/antconc/>), stylometry based on the software “Stylo Package for R”, and the visualization of a network of discursive communities at the internet platform “Gephi”.

It seems important to investigate the lexical and syntactic relationships among these highpoints in order to identify how French prose developed in the late medieval period, especially in order to assess the role of Latin as both substratum and adstratum in the development of both spoken and written French. In the first part of my communication I will briefly show the important of the Latin substratum in the *Strasburg Oaths* and *Eulalie*. Using the *piste Brepols*, the method permits a more precise reconstruction of Latin's influence as adstratum and substratum in many other vernacular texts, implying the existence of a Latin-vernacular interfaces in a discursive community as early as the 9th century. The survival of Latin legal formulae in the *Oaths* suggests, if perhaps only faintly, the existence of such a discursive community documented by scraps that are as eloquent as they are fragmentary.

The next question is ascertaining whether translations commissioned by the royal court in well-known historical

contexts were responsible for lexical expansion in French. To answer this question, I first present calculations of lexical diversity from representative works. I have used the platform AntConc to calculate the token/type ratio as a measure of lexical diversity. Preliminary results suggest that the prose works exhibit a higher lexical diversity than works written in verse: in other words, lexical expansion depended in the first instance on the choice of prose over verse. Another important result of this research was ascertaining the difference between lexical diversity in translations commissioned by Philip the Fair and those commissioned by Charles V. In order to explain these differences, I have performed a stylometric analysis of several hundred medieval French texts (as txt-files) using the StyloR platform. The software, combining several functionalities calculates the statistical differences between authors and produces an Excel-file which can be visualized as a network on the Gephi platform. The contribution ends with a brief commentary on the existence of different discursive communities over a period of three centuries in late medieval France and a comparison with a similar visualization of Middle English prose works.

Xavier-Laurent SALVADOR, Fabrice ISSAC et Marco FASCIOLO, *Herméneutique des similarités dans le DFSM: une expérience*

Résumé

L'avènement de l'informatique a engendré une double révolution pour la dictionnaire. Tout d'abord du point de vue des méthodologies, l'utilisation systématique de corpus numériques pour l'élaboration du *Trésor de la langue française (TLF)* en est un exemple, mais aussi, de manière moins massive cependant, en ce qui concerne les interfaces de consultation proposées aux utilisateurs.

Il existe de nombreux dictionnaires en ligne, de natures très diverses : dictionnaires, glossaires, spécialisés ou non, structurés ou non. Les outils et les ressources proposés ont tous la même forme : une base de données plus ou moins complexe associée à

une interface proposant un ou plusieurs outils de consultation ou de recherche. La grande majorité de ces applications se focalisent sur la mise à disposition de ressources linguistiques plus ou moins structurées. Le processus de constitution est totalement déconnecté du processus de consultation. Le principe – ou scénario – le plus fréquemment rencontré en terme d'interface est un calque, une transposition, plus ou moins réussi de l'utilisation des dictionnaires « papier ». Dans ce schéma l'utilisateur final est paradoxalement oublié et les possibilités offertes par l'ordinateur sous-exploitées, alors que parallèlement la masse d'informations proposée a considérablement augmenté.

Afin de pallier cette absence de *continuum*, nous avons développé un outil dictionnaire appelé Isilex, dont l'objectif est d'assister aussi bien les lexicographes dans l'élaboration du dictionnaire que les utilisateurs finaux pour le consulter. Notre présentation s'appuiera en grande partie sur le projet CréaLScience, dont l'objectif est de construire un dictionnaire du français scientifique médiéval. Nous présenterons les différents modules utilisés par l'ensemble des acteurs, les interfaces et les outils développés spécifiquement.

Abstract

The rise of academic computing has provoked a double revolution in lexical research. From the perspective of methodology, the systematic use of digital corpora in the creation of the *Trésor de la langue française (TLF)* is the first example of this revolution, and secondly as well, though in a less extensive manner, the kinds of interfaces available for readers consulting this on-line dictionary.

There are, of course, many on-line dictionaries, of highly different natures: dictionaries, glossaries, specialized or general. The tools and resources available all follow the same format: a more or less complex databank linked to a graphic user interface with one or many tools for consultation and research. The lion's share of these applications are focused on making more or less structured resources available for consultation.

The most frequently encountered principle or scenario as far as interfaces are concerned follows a transposed format, more or less successful, of hard-copy dictionaries. This format, however, paradoxically forgets the reader while at the same time under-exploiting the possibilities of a web-based environment which has vastly increased the amount of consultable data.

In order to remedy this rupture between hard-copy and on-line web-based dictionaries, we have developed a lexical tool called “Isilex” whose purpose is to help both lexicographers in expanding the dictionary as well as ordinary readers consulting it. Our presentation is based on the larger project CréaLSscience whose goal is to construct a dictionary of medieval scientific French. We present different modules used by both lexicographers and readers and the interfaces and tools specifically developed for them.

COMITÉ SCIENTIFIQUE

Hava BAT-ZEEV SHYLDKROT (Université de Tel Aviv)
Françoise BERLAN (Université Paris-Sorbonne)
Mireille HUCHON (Université Paris-Sorbonne)
Peter KOCH (Universität Tübingen)†
Anthony LODGE (Saint Andrews University)
Christiane MARCHELLO-NIZIA (École normale supérieure-LSH, Lyon)
Robert MARTIN (Université Paris-Sorbonne/Académie des inscriptions
et belles-lettres)
Georges MOLINIÉ (Université Paris-Sorbonne)†
Claude MULLER (Université Bordeaux Montaigne)
Laurence ROSIER (Université Libre de Bruxelles)
Gilles ROUSSINEAU (Université Paris-Sorbonne)
Claude THOMASSET (Université Paris-Sorbonne)

COMITÉ DE RÉDACTION

Claire BADIOU-MONFERRAN (Université de Lorraine)
Michel BANNIARD (Université Toulouse 2-Le Mirail)
Annie BERTIN (Université Paris Ouest Nanterre La Défense)
Claude BURIDANT (Université Strasbourg 2)
Maria COLOMBO-TIMELLI (Université Paris-Sorbonne)
Bernard COMBETTES (Université de Lorraine)
Frédéric DUVAL (École nationale des chartes)
Pierre-Yves DUFEU (Université Aix-Marseille 3)
Amalia RODRIGUEZ-SOMOLINOS (Universidad Complutense de Madrid)
Philippe SELOSSE (Université Lyon 2)
Christine SILVI (Université Paris-Sorbonne)
André THIBAUT (Université Paris-Sorbonne)

COMITÉ ÉDITORIAL

Olivier SOUTET (Université Paris-Sorbonne), Directeur de
la publication
Joëlle DUCOS (Université Paris-Sorbonne-EPHE), Trésorière
Stéphane MARCOTTE (Université Paris-Sorbonne), Secrétaire de rédaction
Thierry PONCHON (Université de Reims Champagne-Ardenne), Secrétaire
de rédaction
Antoine GAUTIER (Université Paris-Sorbonne), Diffusion de la revue

Table des matières

Présentation	
Joëlle Ducos	7
À propos du <i>DMF</i> :	
réussites et pièges de la lexicographie électronique	
Robert Martin	11
De la gestion de la variation en moyen français à son élargissement aux états anciens du français : les développements du lemmatiseur LGeRM	
Sylvie Bazin-Tacchella & Gilles Souvay	25
Herméneutique des similarités dans le <i>DFSM</i> : une expérience	
Xavier-Laurent Salvador, Fabrice Issac & Marco Fasciolo	49
Le <i>Lexicon Latinitatis Medii Aevi Regni Legionis</i> (VIII ^e siècle-1230) : caractéristiques et quelques exemples (<i>ventrescas, iera, cumbo, plentum</i>)	
Estrella Pérez Rodríguez	77
La lexicographie de l'italien médiéval et les corpus de l'OVI : un bilan provisoire et quelques nouvelles perspectives	
Elisa Guadagnini	101
Le latin médiéval du <i>Glossarium Mediae Latinitatis Cataloniae</i> : un projet lexicographique dans un contexte européen	
Ana Gómez Rabal	121
Autorité du latin et transparence constructionnelle : le sort des néologismes médiévaux dans le domaine médical	
Michèle Goyens & Céline Szeceł	141
Base de français médiéval : une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique	
Céline Guillot, Serge Heiden & Alexei Lavrentiev	167

Terminographie diachronique : le cas de la terminologie médiévale française Gérard Petit	185
Numérisation et traitement de textes mathématiques grecs : méthodes, problèmes et résultats Ramon Masià	213
À la recherche des communautés discursives au Moyen Âge : un regard numérique sur la connectivité dans la culture vernaculaire et le rôle des traductions dans l'évolution de la prose en moyen français Earl Jeffrey Richards	229
Résumés / Abstracts	249
Comité scientifique	267
Table des matières	269