REVUE DE
LINGUISTIQUE
FRANÇAISE
DIACHRONIQUE

2017

DIACHRONIQUES

LES ÉTATS ANCIENS DES LANGUES À L'HEURE DU NUMÉRIQUE

PDF complet – 979-10-231-2155-1

PUPS

7

DIACHRONIQUES

REVUE DE LINGUISTIQUE FRANÇAISE DIACHRONIQUE

LES ÉTATS ANCIENS DES LANGUES À L'HEURE DU NUMÉRIQUE

ISBN 979-10-231-0581-0
9 791023 105810

12 €

http://pups.paris-sorbonne.fr

Maquette : www.stephanemercier.fr

LES ÉTATS ANCIENS DES LANGUES À L'HEURE DU NUMÉRIQUE

Les états anciens des langues à l'heure du numérique

PUPS

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 & Szecel – 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

Présentation

Joëlle Ducos
EA 4509 STIH
Université Paris-Sorbonne

Linguistique et informatique depuis plusieurs décennies font bon ménage : si les tableaux phonologiques ont correspondu à la binarité des premiers ordinateurs, les développements considérables du TAL, des lemmatisations, des annotations automatiques ont contribué à la fois à un renouvellement des outils d'analyse, avec une puissance considérable de traitement de la matière linguistique, mais aussi au développement de nouvelles approches, voire de nouvelles définitions ou de nouveaux concepts comme on peut le mesurer à la lecture des numéros 141 et 142 de l'*Information grammaticale* (2014) faisant le point sur la place de l'informatique et du numérique dans le traitement automatique de l'oral et de l'écrit. Ainsi Irène Tamba et Daniel Lugazzi tentent-ils de définir le *mot-numérique*, nouveau concept pour une réalité lexicale autre que le mot-forme ou le mot-lexème, un mot codé numériquement, qui ouvre des voies nouvelles à l'analyse linguistique et spécialement à celle du sens[1].

Nouvelles perspectives, révolution conceptuelle ? Il est certain que le numérique n'est pas du papier numérisé et consultable en ligne, en quelque sorte une pure évolution de support où le matériau papier disparaîtrait dans une image : les usages et les applications témoignent d'une transformation de fond, dans la méthode comme dans la réflexion. On pourrait penser qu'accoler l'adjectif *ancien* à *numérique* est paradoxal, voire antonymique. Or depuis que l'ordinateur est un outil pour

1. *Information grammaticale*, n° 142, 2014, p. 40-47.

les chercheurs, les nouvelles possibilités qu'il apportait ont tout de suite emporté leur adhésion, et spécialement celle des médiévistes, qu'ils soient historiens, linguistes ou littéraires. Citons la revue *Le médiéviste et l'ordinateur*, créée en 1979 par un groupe d'historiens et de chercheurs et publiée par l'IRHT, qui a contribué largement à la diffusion des perspectives liées à l'ordinateur, que ce soit dans la textométrie, comme appui à l'analyse des sources et des textes, ou par la création du site *Ménestrel*, à la fois plate-forme des ressources documentaires en ligne pour la médiévistique et lieu de publication pour les réflexions de chercheurs[2]. Les outils numériques ont été très vite utilisés pour les états anciens du français, ne serait-ce au début que par l'établissement de concordanciers, et par de multiples entreprises, soit de traitement de corpus, soit de dictionnaire. C'est ainsi que le corpus de Chrétien de Troyes, traité informatiquement par l'université d'Ottawa, a été le point d'origine du *Consortium international pour les corpus de français médiéval*, créé en 2004 par les universités d'Ottawa, du Pays de Galles, de Stuttgart, de Zürich, l'ENS-LSH, l'ATILF, et l'École nationale des chartes. Citons aussi la Base de français médiéval créée à l'initiative de Christiane Marchello-Nizia en 1989, et désormais projet phare du Laboratoire ICAR (UMR 5191 ENS LSH/CNRS). Pour les dictionnaires, c'est le *Dictionnaire de moyen français* (*DMF*) créé par Robert Martin qui apparaît comme la référence d'une entreprise conçue dès le départ sous un format informatique, transformant alors la lexicographie par l'idée d'un dictionnaire dont les versions successives soulignent les apports et les évolutions, en fonction d'un format qui n'est plus celui de la page éditée et qui est élaboré pour répondre aux contraintes et aux potentialités de la diffusion en ligne. Ces entreprises, qui reposent sur les textes en français médiéval, sont apparues alors que d'autres naissaient pour d'autres langues, dans d'autres pays, et reposaient sur les mêmes interrogations à propos de la place de l'informatique dans les disciplines de la

2. Ménestrel, en ligne : http://www.menestrel.fr/spip.php?rubrique397&lang=fr [consulté le 21 juin 2017].

médiévistique, qu'il s'agisse de la linguistique, de l'histoire, de la philologie ou la littérature. Ces travaux pionniers ont été à l'origine des réflexions actuelles et de la multiplication des projets numériques ou de linguistique outillée, que renforcent actuellement les financements nationaux et internationaux de la recherche. L'apport informatique y apparaît clairement comme une expansion considérable de potentialités de traitement, mais surtout comme un appel à la réflexion méthodologique et conceptuelle, ainsi qu'à la rigueur intellectuelle, qui fait progresser aussi dans la connaissance et dans l'analyse sémantique et lexicologique.

Il semblait donc nécessaire de procéder à une mise en perspective de l'apport du numérique à la connaissance des langues médiévales ou plus anciennes encore, qu'il s'agisse du latin, du français, de l'italien ou du grec, dans la mesure où le demi-siècle qui vient de s'écouler a permis la réalisation de plusieurs projets, anciens ou tout récents. Il s'agit de rendre compte d'expériences, de méthodes ou d'approches différentes, de tirer les leçons de ce qui relève désormais d'une histoire de la recherche informatique, de mesurer les résultats obtenus pour envisager l'étape suivante, à l'heure des *Big Data* et des *Linked Open Data* : représentation et modélisation des données qui permettent des interfaces et des requêtes multiples, gestion de ces données…

Le présent volume ne prétend pas répondre à tous les enjeux actuels issus aussi bien des nouvelles possibilités technologiques que de l'augmentation considérable des données disponibles. Il a pour ambition de mettre en évidence les convergences entre des projets séparés, sur des aires linguistiques différentes, menés par des chercheurs qui rendent compte de leurs méthodes, de leurs outils et de leurs avancées. Les applications sont multiples, et vont de la syntaxe à la stylistique en passant par les dictionnaires ou la sémantique. Loin de considérer en effet que le numérique est susceptible de détruire la diachronie et les études classiques, les expériences menées ces dernières années – et leurs résultats – prouvent combien il peut être le

moteur d'un renouvellement, à condition que la fin ne soit pas l'outil en lui-même, mais bien la recherche en linguistique : c'est ainsi que la nouvelle philologie a trouvé par l'édition numérique le moyen de mettre en œuvre de nouvelles formes d'édition et de rendre compte à grande échelle de la variance et de la variation. Qu'apporte le numérique ? Quelles en sont les limites, et quels nouveaux horizons ouvre-t-il ? Telles sont donc les questions sous-jacentes aux différentes contributions présentées ici. Elles témoignent toutes de l'importance des prémisses, de la nécessité de la mise au point d'une méthodologie rigoureuse, qui ne soit pas seulement technologique mais repose sur une connaissance véritable d'un état de langue. Elles révèlent aussi les difficultés, les nœuds éventuels qui peuvent se créer mais, surtout, elles témoignent de l'ouverture à de nouvelles dimensions, des ponts qui désormais sont praticables entre des domaines et des aires séparés alors que la technologie progresse, et permet d'aller bien au-delà de l'automaticité binaire des premières tentatives. Fouilles de textes, ontologies, lemmatiseurs, utilisation de standards internationaux, autant de perspectives qui se répondent et qui s'adaptent à la réalité mouvante et complexe des textes et des langues du Moyen Âge.

Ce numéro, issu du colloque final d'un programme ANR, CréaLScience (2010-2014)[3], qui a permis la conception du *Dictionnaire de français scientifique médiéval* (*DFSM*), se veut aussi une ouverture au-delà des frontières linguistiques et la mise en évidence d'une communauté de projets rendant compte de la diversité linguistique de la période médiévale. L'outil informatique permet ainsi de reconstruire la richesse médiévale, les réseaux, les relations et les échanges qui s'établissaient à cette époque par-delà l'obstacle de la langue. Cette communauté « numérique » qui se construit est assurément désormais le point de départ de nouvelles voies dans les territoires de la recherche médiévale.

3. En ligne : www.crealscience.fr.

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

Robert Martin
Académie des inscriptions et belles-lettres

La part du numérique ne cesse de croître – dans nos disciplines comme ailleurs. Le bénéfice est tel que l'on n'imagine pas un retour en arrière : un autre âge s'est ouvert, non pas épistémologique sans doute, les questions de fond restant inchangées, mais technique et méthodologique. Le stockage électronique des données, leur organisation, les accès instantanés que l'informatique autorise, les modèles interprétatifs et les représentations qu'elle suscite, la rigueur du contrôle qu'elle impose, tout cela est de si grande conséquence qu'il convient, avec le minimum de recul qui désormais s'instaure, d'en prendre une juste mesure. Si les avantages l'emportent, les pièges cependant ne sont pas inexistants. C'est vrai tout particulièrement en lexicographie. L'expérience du *Dictionnaire du moyen français* (*DMF*)[1] devrait en l'occurrence faciliter les entreprises similaires, qu'elles soient à leurs débuts ou déjà en cours. J'évoquerai tout d'abord ces avantages, avant d'insister sur les revers possibles, que l'expérience acquise devrait contribuer à déjouer.

Bénéfices de la lexicographie électronique

Dans l'histoire (encore récente) de la lexicographie électronique, trois grandes étapes dès à présent se dessinent : l'étape de l'informatique documentaire ; l'étape de l'informatisation des dictionnaires ; l'étape de l'élaboration

1. En ligne : www.atilf.fr/dmf.

lexicographique assistée par ordinateur. Chacune d'elles procure d'incontestables bénéfices.

La phase documentaire

Dans l'étape la plus ancienne, l'informatique se limite à la phase documentaire. Il s'agit alors :

– de rassembler, sous un format lemmatisé, un grand nombre de données ;
– de les affecter de références standardisées et immuables ;
– de les trier automatiquement selon des critères formels susceptibles d'en faciliter au mieux l'exploitation.

Fini le labeur fastidieux des fiches manuscrites, oubliées les références flottantes. Voyez le Godefroy : un même texte peut y être référencé sous des formes variables. Un exemple entre mille : *Le Racional des divins offices* de Guillaume Durand, adapté en français par Jean Golein, désormais accessible, du moins en partie, dans l'édition établie par Brucker et Demarolle. Ce texte figure dans Godefroy sous le nom d'auteur de « G. Durant » jusqu'à l'entrée *chapefol*, puis, à partir de l'entrée *collectaire*, sous l'étiquette bibliographique de « J. Goulain, *Ration.*, Richel. 437, p. ex., t. VII, 428b, *s.v. sincopiser* », ou « t. VII, 594a, *s.v. supererogation* » ; mais parfois aussi sous la forme « J. Goulain, *Trad. du Ration. de G. Durant*, B.N. 437, p. ex. GDC X, 652a, *s.v. segregation* ». Des références bibliographiques immuables et, grâce à l'informatique, commodes d'accès, évitent ces flottements.

Les opérations de tri, même élémentaires, aident à dominer la masse des informations. Ainsi, dans le *Trésor de la langue française* (*TLF*), dès 1966 (il y a près d'un demi-siècle !), le programme dit des « groupes binaires » a permis de donner une idée plus juste de la syntagmatique ; la démarche est simple : *maison* apparaît dans le corpus avec une certaine fréquence ; de même *campagne* ; on compare alors la probabilité de trouver *maison* et *campagne* côte à côte (comme dans *maison de campagne*) par le seul fait du hasard à la fréquence effective, en l'occurrence significativement supérieure ; le seul hasard ne

pouvant expliquer la fréquence de *maison de campagne*, il ne peut s'agir que d'un fait linguistiquement pertinent ; c'est là une donnée qui ne peut laisser le lexicographe indifférent. Les tris de cette espèce remontent aux débuts de l'informatique : ils ont marqué la discipline.

L'informatisation

La seconde phase de l'histoire de la lexicographie électronique est celle de l'informatisation des dictionnaires, d'abord par rétroconversion (comme pour le *Dictionnaire d'Oxford* ou pour le *TLF*), ensuite par balisage initial (comme pour le *DMF*). Les avantages en sont désormais si connus et si unanimement appréciés qu'il suffit de les rappeler brièvement sous quelques rubriques ; ils tiennent à la diversité des *accès* et à la commodité des *liens*.

Les *accès* sont en effet très divers :

– ils sont déliés de la linéarité (on peut afficher toutes les occurrences de *campagne*, même en dehors de l'article *campagne*) ;
– ils peuvent engager une lemmatisation (dans le *DMF*, l'accès dit « Mot ou forme » propose l'ensemble des articles auxquels une forme peut théoriquement appartenir : ainsi en demandant *flageole*, on obtient le lemme *flageole*, substantif qui existe, mais aussi *flageoler*, dont *flageole* est une forme fléchie ; il va sans dire que la pertinence des propositions faites par le lemmatiseur ne saurait être absolue ; cependant les propositions correctes sont à présent nettement supérieures à 95 %, et proches de 99 % si l'on tient compte des réponses plurielles parmi lesquelles figure la réponse correcte) ;
– ils peuvent s'opérer par fragments (dans le *DMF*, sous l'intitulé « Filtre » : on peut afficher par fragments initiaux, p. ex. les mots qui commencent par *in-* ; par fragments terminaux, p. ex. tous les mots qui se terminent par *-tendre* : *attendre*, *contendre, contrattendre, détendre, distendre, entendre, forestendre, mesentendre, parentendre, partendre, pourtendre, prétendre, protendre, retendre*… ; ou par

fragments internes, p. ex. tous les mots qui contiennent *-fil-* : *affiler*, *défiler, effiler*, *effiloir, enfiler, forfiler, profiler, pourfiler*, *refiler*…) ;

– les accès peuvent aussi se réaliser par types d'informations, p. ex. par l'étymon, par la syntagmatique ou par le repérage sous une balise donnée, p. ex. *maison* ou *campagne* dans les définitions.

Dans le *DMF*, les *liens hypertextuels* se diversifient en trois types de liens :

– des *liens internes*, comme le lien des références abrégées avec la Bibliographie, ou bien le lien d'une forme quelconque dans une citation avec le ou les articles qui en traitent ;
– des *liens avec les Bases* qui ont servi à construire le Dictionnaire, pour le *DMF* avec la Base des « Lexiques préalables » et avec les « Bases textuelles » ;
– des *liens avec d'autres ouvrages* ; ainsi le *DMF* permet, article par article, d'ouvrir le *TLF*, le Godefroy ou le *FEW* (en 2015 l'*AND*, le *DECT* et en partie le *DEAF*).

L'élaboration par voie électronique

La troisième phase est celle de l'élaboration du Dictionnaire par voie électronique, celle d'une lexicographie assistée par ordinateur. Grâce à Gilles Souvay, le *DMF* dispose de divers outils qui assistent le rédacteur tout au long de sa démarche.

Ainsi le *DMF* s'élabore selon une *grammaire lexicographique* qui garantit l'homogénéité de l'écriture. Cette grammaire est un système qui, au fil de la rédaction, spécifie le type d'information qu'il convient de fournir ; elle affiche au fur et à mesure les balises à remplir et les choix à faire parmi les balises possibles à l'endroit où l'on est arrivé. Ainsi la première balise, obligatoire, est celle de la vedette : le curseur indique où il convient de l'écrire ; la forme s'enregistre automatiquement en gras ; vient ensuite un premier choix : on peut se borner à un simple renvoi (qui doit être précisé à l'endroit où le curseur s'est placé) ou bien on choisit de poursuivre, auquel cas, un exposant est possible (en cas d'homonymie), puis c'est le « code grammatical » qu'il

convient de préciser ; le système l'écrit en bas de casse romain ; vient ensuite la balise « Dictionnaires » ; le système affiche les dictionnaires possibles ; le choix de l'un d'entre eux entraîne l'affichage du mode de référence que le *DMF* exige, par exemple le lemme pour le GD (Godefroy) ; le curseur demande à nouveau que le lemme soit indiqué ; il l'écrit automatiquement en bas de casse italique ; et ainsi de proche en proche pour toutes les autres balises, jusqu'à la signature de l'article.

Le *DMF* est soumis, article par article, à un *contrôle lexicographique*. Plusieurs correcteurs limitent les erreurs : d'abord un correcteur bibliographique (il vérifie que les références bibliographiques sont bien identifiées et correctement appliquées) ; puis un autre correcteur effectue un « bilan sur les lemmes » (il vérifie que les articles prétendument corrigés par le rédacteur correspondent bien à des lemmes existants ; il signale les formes réputées être des occurrences du lemme qui paraissent suspectes, p. ex. celles qui ne commencent pas par la même lettre ; *queuillir* est-il bien une forme de *cueillir* ? *tresnoble* une forme de *noble* ?…). Enfin un correcteur d'ensemble signale toutes sortes d'incohérences : des erreurs de balisage (p. ex. des valeurs incorrectes : sous la balise DOM[aine], on ne peut utiliser que des domaines reconnus, de même sous la balise CODE GR[ammatical] ; certaines balises ont pu rester incomplètement remplies : le système le décèle sans faille) ; des erreurs matérielles : typographiques (espaces multiples ; absence d'espace après un signe de ponctuation double comme le point-virgule ; absence d'espace après la balise OCC[urrence] ; guillemets, parenthèses ou crochets qu'on a oublié de refermer ou qui ne sont pas ouverts…) ; ordre alphabétique défectueux des entrées ; ordre chronologique défectueux des exemples dans un même paragraphe ; numérotation incohérente… ; des erreurs diverses : mauvaise forme de l'étymon (le système contient une liste complète, avec leurs références, des étymons du *FEW*) ; mauvaise référence de l'étymon (qui par exemple se trouve dans un autre volume que celui que le rédacteur a erronément indiqué) ; mauvaise graphie du lemme de Tobler-

Lommatzsch (là aussi le système contient toutes les graphies de la nomenclature) ; erreurs de renvoi (comme le renvoi à un lemme qui n'existe pas), etc.

Le *DMF* se prête par ailleurs à la *révision lexicographique*. Elle porte sur tel ou tel type d'information (sous telle ou telle balise). Ainsi, un programme dit « des mots cachés » se révèle particulièrement efficace. Tous les dictionnaires contiennent des « mots cachés », c'est-à-dire des mots qui figurent dans des exemples mais qui, par mégarde, ne sont pas entrés dans la nomenclature. Comment les détecter automatiquement ? La révision consiste à soumettre le corpus complet des exemples (le TEXTE sous la balise EXemple) à une lemmatisation systématique : quand la lemmatisation échoue, la probabilité est forte qu'il s'agisse d'un « mot caché » ; la procédure permet d'ajouter (dans la version de 2015) plusieurs centaines de mots oubliés (*abondable* « qui abonde », *acrisie* « cécité », gr. *akrisia*, *adizeler* « mettre par groupes de dix gerbes » ; notamment des mots qui ont survécu, comme *réprimander*, *réservoir*, *rudoyer*…). Naturellement, l'échec de la lemmatisation peut être due à d'autres causes : le lemmatiseur ne reconnaît pas toujours les noms propres ou les mots étrangers, souvent des formes latines non balisées ; il piétine aussi en cas de faute de frappe, mais c'est là une bénédiction : toutes sortes de coquilles sont par la même occasion réparées (*accoderoit* a toute chance d'être *accorderoit* ; *avoità* est à corriger en *avoit à* ; *bataillle* est assurément *bataille*…). Le programme des « mots cachés » est une des facettes de la plateforme LGeRM que Sylvie Bazin et Gilles Souvay nous présentent ici-même ; je n'en dirai pas plus pour ne pas déflorer leur sujet.

Ajoutons que le *DMF* se met en page automatiquement ; de ce fait même, les articles sont immédiatement imprimables sous le format Word (la publication du *DMF* sur papier serait facilement réalisable, avec toutes les exigences de l'imprimé traditionnel : colonnes, retraits, italiques, changements de corps…).

Parmi les bénéfices les plus marquants de la lexicographie électronique, l'un tient à l'extraordinaire facilité du remodelage.

Le rédacteur, construisant son article à l'écran, est à même de le corriger, de le réorganiser et de l'augmenter indéfiniment. De surcroît, les limitations éditoriales disparaissent ; la hantise du nombre maximal de signes, indissociable de la forme imprimée, n'a plus de raison d'être. Certes un certain équilibre doit s'instaurer entre les articles ; mais c'est là une exigence liée à l'économie interne, et non plus d'ordre budgétaire.

Tous ces outils nouveaux et toutes ces possibilités évolutives représentent en lexicographie une avancée inestimable. Naturellement les erreurs les plus graves ne sont pas reconnues par l'automate : l'incompétence du lexicographe ne sera jamais palliée par aucun ordinateur.

Pièges de la lexicographie électronique

La lexicographie électronique n'est toutefois pas exempte de pièges qui lui sont propres. Je voudrais maintenant les évoquer de manière plus détaillée. Ces pièges peuvent se regrouper sous trois enseignes : l'*instabilité*, une croissante *complexité* de plus en plus difficile à dominer, l'*immatérialité* et le risque de l'inexistence qu'elle entraîne.

Le piège de l'instabilité

Pour un dictionnaire électronique, l'un des principaux dangers est de se présenter comme une base de données en constante évolution. Dès lors que l'ouvrage se modifie de jour en jour, voire d'heure en heure, de manière imprévisible, il n'est plus possible de s'y référer si ce n'est en indiquant à chaque fois la date et l'heure de la consultation, ce qui est tout à fait dissuasif. L'ouvrage risque alors de se perdre dans le gigantesque fouillis des données anonymes de la « toile », où l'on puise sans le dire dans le déni de la source ; et peu à peu le dictionnaire s'apparente à un inexistant. Alors qu'une publication sur papier donne lieu au dépôt légal et que s'agissant de ce support l'on se réfère toujours à une édition précise, le risque se crée sur la « toile » qu'aucune date éditoriale ne s'attache plus à l'objet.

Le remède cependant est relativement simple. Il faut veiller à garder disponibles des versions successives correctement identifiables et, surtout, ne pas livrer une production incessamment remodelée. Les étapes doivent être dénommées sans ambiguïté, et demeurer stables comme le sont les éditions d'un livre. Les erreurs sont corrigées de version en version, et non pas au jour le jour. Le *DMF* s'inscrit ainsi dans une perspective de lexicographie évolutive : l'avantage est considérable ; s'il est très éloigné de la perfection, il est au moins indéfiniment perfectible ; les versions se succèdent et assurément s'améliorent. Le *DMF* 1 (en 2002, sur Internet en 2003) n'était qu'un cumul, sous des lemmes communs, de Lexiques préalables ; sa nomenclature ne dépassait pas les 25 000 entrées. Le *DMF* 2 a porté la Nomenclature à plus de 60 000 entrées ; le *DMF* 2009 a ajouté divers Lexiques et synthétisé près de la moitié de la Nomenclature ; le *DMF* 2010 s'est accru de nouveaux Lexiques et a synthétisé le reste ; le *DMF* 2012 a repris les vocables qui n'apparaissaient que dans un seul Lexique préalable et a fait une place à l'immense documentation sur papier dont le *DMF* dispose ; le *DMF* 2015 a réparé des bévues diverses, augmenté de beaucoup les exemples et surtout a multiplié les liens hypertextuels.

En somme, l'ouvrage électronique s'améliore peu à peu et ses versions antérieures restent disponibles, avec leurs insuffisances et leurs erreurs, comme c'est le cas pour les éditions successives d'un livre. Dès lors le renvoi à l'ouvrage se fait sans difficulté, dans les conditions habituelles de l'imprimé.

On objectera à juste titre que l'on renonce tout de même à l'avantage considérable de la mise à jour immédiate. Dans le *DMF* 2012, nous avons imaginé un subterfuge qui combine la nécessité absolue d'éditions successives et la possibilité si souhaitable d'une constante progression : c'est la technique de l'« Annotation ». Les corrections et ajouts sont consignés sous forme d'« Annotation » au bas des articles concernés. L'article lui-même reste ainsi inchangé, mais on indique au fur et à mesure les modifications qui y seront apportées dans la version suivante (p. ex. sous FLOT 2 : « V. aussi FLOC 4, à regrouper ici »). La version

DMF 2012 comporte de surcroît une rubrique nouvelle, intitulée « Compléments au *DMF* 2012 » : elle permet d'enregistrer des données plus consistantes qui devraient trouver leur place dans les versions futures. Ainsi les « Compléments » ajoutent des mots nouvellement rédigés, des mots à rédiger pour lesquels on fournit dès maintenant un exemple, des ajouts divers à des articles existants, des versions provisoires de lexiques en cours (p. ex. un Lexique des Préfixes et des Suffixes, ou encore un Lexique des mots grammaticaux).

Ce dispositif est maintenu dans le *DMF* 2015 (et le sera assurément au-delà), à cette nuance près que l'on privilégie l'« Annotation » plutôt que le « Complément ». Les corrections ponctuelles, les indications pour un plan remodelé, les exemples supplémentaires : tout cela peut aller en « Annotation ». On ne réservera aux « Compléments » que les articles nouveaux et d'éventuels Lexiques inédits.

Le piège de l'instabilité est ainsi contourné. La formule retenue concilie ce qui au départ a pu paraître inconciliable.

Le piège de la complexité

Un autre piège est celui d'une croissante complexité, qui risque peu à peu de n'être plus dominable. Les logiciels du *DMF* sont régulièrement modifiés, améliorant les accès, ajoutant des fonctionnalités nouvelles, corrigeant les insuffisances manifestes. Comment s'en plaindre ? Mais à la limite, un seul informaticien détient encore la clé de l'édifice ; et si, pour une raison ou une autre, il se retirait de l'entreprise, la suite deviendrait fort incertaine. Le « montage » d'une version nouvelle prend des allures acrobatiques. Les erreurs conjuguées des rédacteurs et la superposition des programmes compliquent singulièrement la tâche. Ainsi une des faiblesses du *DMF* 2012 (corrigée autant que possible en 2015) tient aux « doublons » relativement nombreux que comporte cette version (comme souvent dans les dictionnaires qui portent sur des états de langues où les graphies ne sont pas stabilisées – les « doublons » sont en grand nombre dans le Godefroy) ; un même vocable peut

se présenter sous des lemmes légèrement distincts et qui font double emploi, p. ex. *affait*[1] n'est autre que *affect*[2] ; en 2015 *affait*[1] disparaîtra au seul profit de *affect*[2] nouvellement rédigé, et *affait*[2] apparaîtra en conséquence sous un lemme *affait* sans exposant. Reconnaissons que les modifications de cette espèce ont de quoi inquiéter…

Le piège de la complexité n'est pas facile à déjouer. Il est pourtant impératif d'y porter remède.

Dans le *DMF*, quelques progrès sont déjà réalisés, notamment dans les procédures de « montage ». Jusqu'ici, *grosso modo*, on combinait les fichiers source avec des fichiers d'ajouts (des fichiers d'articles nouveaux et d'articles revus) ; on aboutissait ainsi, de version en version, à une reconstruction entièrement nouvelle du *DMF*. La technique est désormais différente, beaucoup plus simple. La base sera exclusivement la version immédiatement précédente du *DMF* (en l'occurrence le *DMF* 2012), à charge pour les rédacteurs d'une part de signaler les articles qui sont à supprimer et d'autre part de fournir les articles qui les remplacent et les articles qui viennent s'y ajouter. L'extraction des articles balisés du *DMF* 2012 (permettant de les corriger) s'opère dès à présent sans difficulté majeure ; c'est un progrès technique tout à fait appréciable.

Mais le problème de la complexité est beaucoup plus général. Il y a tout à gagner, semble-t-il, quel que soit le projet en cause, le *DMF* ou un autre, à bien distinguer :

– d'une part les *données*, leur *balisage*, et les *programmes* de traitement ;
– et d'autre part, parmi les programmes, les *programmes de base* et les *programmes périphériques*.

À un moment donné, nous nous sommes demandé s'il ne fallait pas, pour sa sauvegarde, imprimer le *DMF* en deux ou trois exemplaires. Le plus utile est de distinguer nettement les données, les balises et les programmes. Les données (p. ex. en PDF) trouvent aisément place sur une simple clé USB. La

matière de l'ouvrage électronique devient ainsi très facile à sauvegarder.

Dans le *DMF*, les programmes de base gravitent autour des accès lemmatisés. Aussi, lors de la suppression d'articles, convient-il de supprimer simultanément les données correspondantes du lemmatiseur ; et, lors des ajouts, le lemmatiseur doit être configuré en conséquence.

Les programmes périphériques ne concernent que l'affichage. Leur évolution et leur accumulation ne mettent pas en cause les programmes de base. En voici deux exemples. L'affichage des familles de mots : en 2015, on a distingué l'affichage par « étymons » (sont réputés de la même famille les vocables que le *FEW* regroupe sous le même étymon) et l'affichage par « hyperétymons » (ainsi *creatio*, *creator* et *creatura* seront placés sous l'hyperétymon *creare*) ; c'est là une décision qui n'affecte que la surface. Autre exemple : les retouches purement graphiques de la nomenclature ; dans le *DMF* 2015, on a essayé de pallier une décision malheureuse qui a été retenue dans le *DMF* 1 et qui a consisté, non seulement à adopter les diacritiques des lemmes qui ont survécu en français moderne, mais à moderniser les lemmes de même famille ; les diacritiques conduisent alors à des inexistants : *anéantissement*, soit ; mais *anéantance* ? Cette forme n'a jamais existé ; la nomenclature du *DMF* prend de ce fait même un aspect étrange, qu'il convient de rectifier. On posera désormais l'équivalence stricte de *anéantance* et de *aneantance*, l'affichage étant toujours *aneantance* ; c'est là encore une décision de surface, qui ne touche pas les programmes de base.

Une chose est sûre : les plus grands efforts doivent aller à la simplification des procédures. C'est un point capital pour la survie et pour la transmission d'un projet.

Le piège de l'immatérialité

Le piège le plus redoutable est celui de l'immatérialité. Ses conséquences ne laissent pas d'être inquiétantes. Un dictionnaire électronique, pas plus qu'aucun autre ouvrage publié sous format électronique, n'a d'existence tangible ; là

où l'imprimé, voire le manuscrit (s'il n'est pas endommagé ou incendié) peut se conserver indéfiniment, l'objet électronique est voué à disparaître s'il n'est plus abrité par un site actif. Nous ne consultons plus guère aujourd'hui le *Dictionnaire* de La Curne ; il n'en reste pas moins parfaitement accessible (et parfois très utilement). Le temps viendra où le *DMF* sera complètement dépassé. À la différence du La Curne, le risque est alors grand qu'il tombe dans les oubliettes sans laisser la moindre trace.

Pourtant tout ouvrage scientifique, même mineur, devrait bénéficier d'une garantie minimale de survie. La voie d'avenir est celle du dépôt légal étendu aux ouvrages numériques. Il serait extrêmement souhaitable que les publications numériques bénéficient d'un dépôt légal comparable à celui des publications sur papier, la production scientifique passant de plus en plus par cette voie. Seul le dépôt légal garantirait leur pérennité. Le site internet de la BnF, hélas, indique qu'« à ce jour, il n'y a pas de dépôt *à l'unité* des publications numériques en ligne ou téléchargeables, leur collecte passe par le site web qui les diffuse ». Cela signifie que le seul dépôt légal pour les ouvrages numériques est celui qu'opère le robot d'archivage Heritrix, qui explore automatiquement les sites en ligne. L'ennui est que le robot procède par échantillonnage ; de surcroît, dans le cas des dictionnaires, dont l'accès se fait lemme par lemme, la collecte est inopérante.

Pourtant il ne devrait pas être trop difficile de mettre en place un dépôt « à l'unité ». On se bornerait au contenu scientifique (à la matière imprimable) des ouvrages en cause, sans prendre nécessairement en compte les balisages, encore moins les logiciels d'interrogation. On s'en tiendrait à des éditions immuables, qui se suivraient à des intervalles suffisants pour en justifier la pérennité. On imposerait un format standardisé (actuellement le format PDF), dont l'évolution est universellement suivie et qui garantirait de ce fait même la sauvegarde des ouvrages dans la durée. Une correspondance récente (en date des 18 et 20 août 2014) avec Mme Hélène Jacobsen, qui dirige le département du Dépôt légal à la BnF, a au moins confirmé que la question est à l'étude. Espérons que la voie sera ouverte dans des délais rapprochés !

Bref, les progrès de l'informatisation sont immenses. Ils s'accompagnent certes de risques de plusieurs espèces, mais qui peuvent être contenus. Tout donne à penser que l'avenir de la lexicographie scientifique est désormais indissolublement lié à l'informatique.

Références bibliographiques

Duval, Frédéric, « *Dictionnaire du moyen français* (*DMF* 2) », *Romania*, n° 126, 2008, p. 530-539.

Gerner, Hiltrud et Souvay, Gilles, « Présentation de la seconde version du *DMF* », dans Iliescu, Maria, Siller-Runggaldier, Heidi M. et Danler, Paul (dir.), *Actes du XXV^e^ Congrès international de linguistique et philologie romanes* [2007], Tübingen, Niemeyer, 2007, p. 213-220.

Gorcy, Gérard, Martin, Robert et Maucourt, Jacques, « Le traitement des "groupes binaires" », *Cahiers de Lexicologie*, n° 17, 1970, p. 15-46.

Martin, Robert, « Pour un dictionnaire du moyen français », dans Wunderli, Peter (dir.), *Du Mot au Texte. Actes du III^e^ Colloque international sur le moyen français* [1980], Tübingen, Narr, 1981, p. 13-24.

—, « Le *Dictionnaire du moyen français* (*DMF*) », *Comptes rendus des séances de l'année 1998*, Académie des inscriptions et belles-lettres, novembre-décembre 1998, p. 961-982.

—, « Note sur le *DMF* 2012 (*Dictionnaire du moyen français*, version de 2012) », *Romania*, n° 131, 2013, p. 173-178.

—, « Bref retour historique sur le *Dictionnaire du moyen français* », *Romania*, n° 133, 2015, p. 219-227.

Martin, Robert et Souvay, Gilles, « Le *Dictionnaire du moyen français*, *DMF* 2 (Note d'information) », *Comptes rendus des séances de l'année 2008*, Académie des inscriptions et belles-lettres, janvier-mars 2008, p. 49-57.

Schneider, Stefan, « *Dictionnaire du moyen français*, version 2012 », *Zeitschrift für romanische Philologie*, n° 129, 2013, p. 1232-1237.

Städtler, Thomas, « Die evolutive Lexikographie am Beispiel der Geschichte des *Dictionnaire du moyen français* », *Beiheft zur Zeitschrift für französische Sprache und Literatur*, n° 120, 2010, p. 1-13.

Trotter, David, « *Dictionnaire du moyen français* (1330-1500) : *DMF* 2 », *Journal of French Language Studies*, n° 20, 2010, p. 342-344.

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
ATILF/CNRS
Université de Lorraine

La variation est une donnée constitutive de la morphologie lexicale pour nombre de mots : ils peuvent recevoir des marques de genre, de nombre, de personne, de temps ou d'aspect selon leur catégorie et/ou leur emploi. Le dictionnaire de langue ne rend pas compte de cette diversité en français moderne et se contente d'entrées conventionnelles, comme l'infinitif pour le verbe, le singulier pour le nom ou le masculin singulier pour l'adjectif, qui regroupent, en les sous-entendant, les formes fléchies correspondantes. Il s'adresse à des utilisateurs avertis, qui maîtrisent les différents systèmes flexionnels dans une langue standardisée. Cependant, dès lors qu'il est question d'états anciens de la langue, à cette variation morphologique obéissant à des principes qui ont eux-mêmes évolué s'ajoutent d'importantes variations graphiques et diatopiques.

La langue médiévale en particulier ne se livre qu'à travers des témoignages écrits, par essence mouvants et variants, en raison d'une transmission manuscrite s'opérant au cœur d'un diasystème : les copistes sont partagés entre la fidélité au modèle et leurs propres habitudes linguistiques, au croisement des axes diatopique et diachronique, d'où des concurrences ou des variations, parfois dans le même document. Ce qui domine, ce sont des systèmes souples, moins contraints et non normés, ce qui ne veut pas dire aléatoires. Mais, pour un esprit moderne,

cela demande un sérieux effort d'ouverture aux possibles pour reconnaître ou regrouper les formes. Face à une telle variation, la consultation d'un dictionnaire construit selon les principes habituels s'avère bien peu pratique pour le spécialiste, et peu utile au néophyte.

Ainsi, sous quelle entrée trouver les formes *destroict*, *vis*, *ameroyent*, *acouremens*, *polra* ou *menra* que l'on peut rencontrer dans un texte médiéval ? La variation peut être graphique, ainsi *destroict / detroit* ou *ameroyent / ameroit* ; dans le premier cas, le copiste peut continuer de graphier des consonnes qui ne sont plus prononcées (/s/ intérieur ou /k/ avant /t/), ou choisir de les insérer pour rappeler l'étymologie, comme dans *obscur / oscur* ou *tens / temps* ; dans le second cas, la lettre *y* devient un équivalent de *i*, ainsi *loy*, *roy*, etc. Quelle est alors l'entrée choisie par le dictionnaire ? Celle qui se rapproche le plus du français moderne, ou celle qui est la plus fréquente dans la période considérée ? Même lorsqu'un dictionnaire comme le Godefroy donne la liste des formes rencontrées, il faut que l'utilisateur trouve l'entrée sous laquelle est mentionnée la forme qui l'intéresse. Il existe des renvois, mais ils ne sont pas systématiques. Dans le cas de la forme adjective masculine *vis* ou de la forme verbale *menra*, la difficulté est de nature morphologique : la forme *vis*, qui peut être cas sujet singulier ou cas régime pluriel en ancien français, qui est un pluriel lorsque la déclinaison disparaît, n'est pas une entrée du dictionnaire, il faut la chercher sous la forme du singulier *vif* ; le dictionnaire ne permettra pas de retrouver *menra*, forme usuelle en ancien et moyen français du futur simple du verbe *mener*, avec disparition de *e* dans la séquence *-ner-*, puisqu'il faut chercher sous un infinitif dont le lien avec la forme considérée est loin d'être évident. Parfois, l'entrée de référence peut paraître évidente, ainsi pour une forme telle que *vendra*, que l'on aurait tendance à rattacher au verbe *vendre*, selon la morphologie moderne, alors qu'il peut tout aussi bien s'agir du futur du verbe *venir*, construit dans l'ancienne langue sur la base faible du verbe. Des variantes

diatopiques se rencontrent également dans les textes rédigés en ancien et moyen français, selon la coloration dialectale des témoins, ainsi *bel/biel*, *chastel/castel* ou encore *chacier/cachier* (latin **captiare*).

Le *Dictionnaire du moyen français*, dès ses débuts, a été confronté à la difficulté de la variation, car le rédacteur face aux multiples graphies possibles d'un même terme doit lui aussi choisir une entrée. On peut décider de moderniser l'entrée, lorsque le terme s'est maintenu en français moderne – c'est l'option qui a été suivie dans le choix du lemme, ainsi *destroict* est rangé sous le lemme DÉTROIT.

Dans sa première phase, qui a débuté dès 1984, le *Dictionnaire du moyen français* s'est élaboré à partir de la synthèse de matériaux fournis par des lexiques préalablement publiés individuellement, et par la suite complétés par d'autres ressources documentaires, ce qui a conduit en 1998 à la publication d'un premier fascicule regroupant les articles A-AH, appelé *DMF*o. Mais le projet, qui paraissait alors démesuré, a été réorienté, en privilégiant une construction par étapes, à partir non pas des lettres de l'alphabet mais des lexiques déjà rédigés, structurés en base de données balisées sous XML ; la lemmatisation des vedettes a été nécessaire pour construire la base de données. Un outil a alors été développé pour répondre au problème représenté par la variation : il s'agit du lemmatiseur LGeRM, dont il sera plus spécifiquement question dans cette contribution. Le lemmatiseur 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. Il permet d'interroger à partir de la forme rencontrée dans un document : si l'on reprend les exemples précédents, l'utilisateur saisit les formes *vis*, *ameroyent*, *menra* et le lemmatiseur propose un ou plusieurs lemmes possibles.

Prenons la forme *embache* rencontrée dans un texte ; la page d'accueil du *DMF* propose un formulaire d'interrogation :

Fig. 1. Formulaire d'interrogation

La forme est tapée dans la zone de saisie, et la proposition d'affichage des dictionnaires a été sélectionnée. Voici la réponse obtenue :

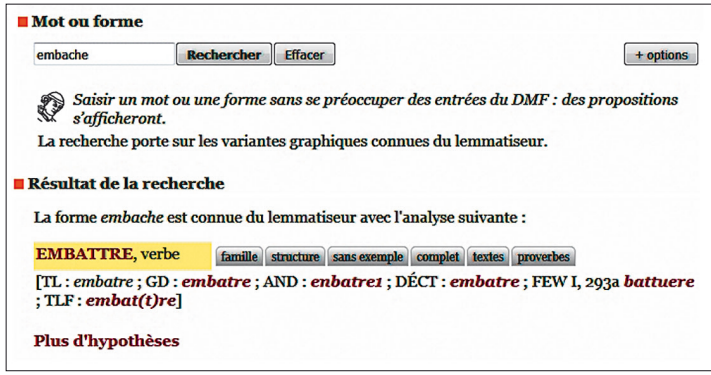


Fig. 2. Résultat de la recherche

On voit apparaître une proposition de lemme, EMBATTRE, et la liste de tous les dictionnaires où le lemme apparaît ; la couleur utilisée – sauf pour le Tobler-Lommatzsch – montre qu'il s'agit de liens hypertextuels et que l'on peut accéder à une version électronique de ces dictionnaires. Mais on aurait tout aussi bien pu sélectionner d'autres fonctionnalités, et obtenir ainsi une réponse plus complète :

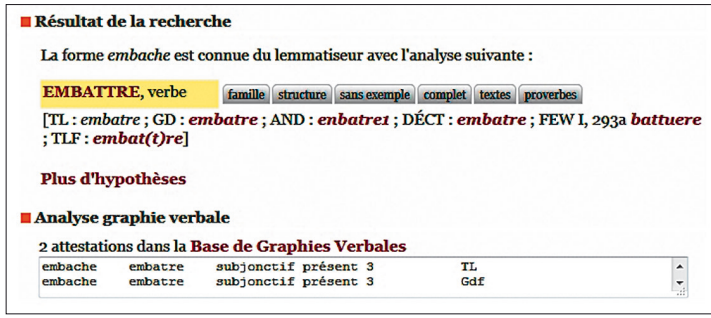


Fig. 3. Résultat de la recherche avec analyse de la graphie verbale

La Base de graphies verbales (BGV) a été constituée à partir de la révision et de la saisie électronique du fonds des formes flexionnelles établi dans les années 1960 par Robert Martin, constitué de fiches manuscrites (entre 16 000 et 20 000) analysant des formes verbales de l'ancien français au XVI^e^ siècle ; cette base, entreprise dans le cadre d'un accord de collaboration scientifique établi entre l'INaLF et le LFA, a reçu l'appui financier de l'université d'Ottawa. La saisie et la lemmatisation avaient été effectuées sous la responsabilité de Pierre Kunstmann, le lemme retenu était celui du Tobler-Lommatzsch, à défaut dans l'ordre celui du Godefroy, du Godefroy complément et celui de Huguet.

Le formulaire de recherche propose également d'indiquer les attestations dans les bases. Voici par exemple la réponse donnée pour CHASSER :

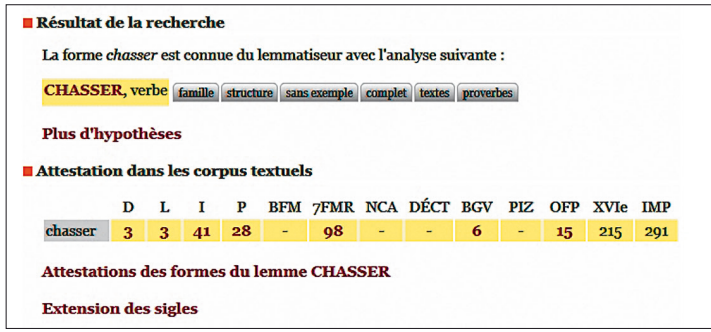


Fig. 4. Résultat de la recherche avec attestations dans les bases

En cliquant sur « Attestations des formes du lemme CHASSER », on peut accéder à un tableau rassemblant toutes les formes qui sont lemmatisées sous CHASSER, rangées par ordre alphabétique, avec mention du nombre d'occurrences par corpus textuel[1]. Voici le début du tableau, qui comporte au total 156 formes répertoriées :

Famille | Structure | Sans exemple | Complet | Formes | Exemples | Lexiques | Textes | Sources | Impression | Aide

CHASSER — Voir diachronie dans FRANTEXT

	D	L	I	P	BFM	7FMR	NCA	DÉCT	BGV	PIZ	OFP	XVIe	IMP	
cacha	-	-	2	2	1	4	4	-	5	-	1	26	44	CACHER (2)
cachans	1	1	1	-	-	1	-	-	-	-	1	5	1	CACHER (2)
cache	9	6	8	14	1	15	8	-	11	-	58	125	301	CACHER (3)
cachez	1	1	1	2	1	13	-	-	3	-	-	90	163	CACHER (2)
cachie	1	1	1	2	-	2	1	-	-	-	-	-	-	CHASSER (2)
cachier	7	6	16	15	7	16	77	-	9	-	24	-	-	CACHER (3)
cachierent	2	2	5	-	3	5	5	-	-	-	6	-	-	CACHER (2)
cachiers	-	-	-	-	-	-	-	-	-	-	1	-	-	
cachies	-	-	-	-	-	-	-	-	1	-	-	-	-	CASSER (2)
cachiet	3	3	4	1	-	4	-	-	-	-	12	-	-	
cachèrent	-	-	-	-	-	-	-	-	1	-	-	1	-	
cacier	-	-	-	-	1	-	34	1	7	-	-	-	-	
caciet	1	1	1	-	-	1	-	-	-	-	-	-	-	
cacièrent	1	1	-	-	-	-	-	-	-	-	-	-	-	
casser	15	11	24	14	1	27	63	16	6	2	1	36	40	CASSER (2)

Fig. 5. Attestations des formes de CHASSER (extrait)

1. Si les sigles utilisés sont familiers des rédacteurs du *DMF*, ils sont opaques pour un utilisateur occasionnel. À partir du lien « Extension des sigles », l'utilisateur pourra accéder à la liste des sigles développés de tous les corpus textuels, avec des précisions sur le nombre de textes, de mots ou de formes selon le cas, leur source et un lien vers le projet quand il s'agit d'un projet extérieur à l'ATILF. Certains liens ne sont plus actifs, c'est le cas du NCA. Voici la liste des sigles développés :
   - D — *Dictionnaire du moyen français* (*DMF*)
   - L — Lexiques du *DMF*
   - I — Intégraux (base de textes à saisie intégrale du corpus *DMF*)
   - P — Partiels (base de textes à saisie partielle du corpus *DMF*)
   - BFM — Copie locale de la *Base de français médiéval* (2006) de l'ENS Lyon
   - 7FMR — Corpus *DMF* réactualisé (base accessible depuis le menu « Recherche dans les textes »)
   - NCA — Copie locale du *Nouveau corpus d'Amsterdam*
   - *DÉCT* — Corpus de textes du *Dictionnaire électronique de Chrétien de Troyes*
   - BGV — Base de graphies verbales
   - PIZ — Liste des mots présents dans l'édition électronique de Christine de Pizan (ms. British Library, Harley 4431, extraction mai 2011)
   - OFP — Liste des mots présents dans *The Online Froissart Project* (extraction juin 2012)
   - XVIe — Sous-corpus de textes du XVI^e^ siècle dans Frantext
   - IMP — Mots présents dans le corpus Impact

Il est possible de cliquer sur les chiffres colorés pour accéder aux exemples eux-mêmes. Voici par exemple les sept occurrences de la forme *cachier* dans le *Dictionnaire du moyen français* :

[1]	(…) et lui remonstra comment il estoit trompé et que le roy Amydas avoit ung petit filz, lequel avoit donné à entendre qu'il estoit mort et l'avoit fait *cachier* et nourrir en loingtain payz, affin qu'il n'en fut nouvelle, pour mieulx marier sa fille et trouver homme qui le secourust à son besoing et remist en sa seigneurie (BUEIL, II, 1461-1466, 252)
[2]	Item, deux marteaux de fer appellez chacheurs, pour *cachier* mine, 2 s. 6 d. (…) Item, deux marteaux de moulin et ung de montaigne pour adouber le fourneau, 10 s. tournois. (Aff. Jacques Cœur M., 1453-1457, 268)
[3]	(…) tous chiaus qui le dit mestier leveront et qui se mesleront de taillier draps d'ore en avant en le dite ville, paieche, pour avoir tout chou que dit est, tout sitost qu'il leveront le dit mestier u qu'il conmencheront a taillier, as conpagnons qui a chou seront commis dou *cachier*, 50 s. tourn. (Drap. Valenc. E., 1369, 41)
[4]	Et *cachier* les pors qui eschiet au roy au pasnage de Bris quant il chiet de l'eau de Rade jusquez à Valloignes, eulx et ceulx de Damneville. Et auxi ilz sont subgetz, ceulx d'Orval, à fere le dit chassage, en tant que il en y a soubz l'onnour de Bris. (HECTOR DE CHARTRES, Cout. R.B., 1398-1402, 149)
[5]	Lors s'en vinrent tout cil de l'avant garde à chevauchant jusques sus les fossés de la cité de Rains, et là descendirent et fissent leurs gens descendre et entrer ens es fossés et *cachier* toutes hors ces bestes. (FROISS., Chron. R., IX, c. 1375-1400, 253)
[6]	« Nous ne poons faire milleur exploit, mais que nos pourveances soient toutes venues, que de aler che chemin que nostre ennemi font, et tant *cachier* que nous les trouvons, et eux combatre. » (FROISS., Chron. R., XI, c. 1375-1400, 271)
[7]	… li jones rois Edouwars d'Engleterre et la roine s'en vinrent a Evruich euls tenir et lor estat, et *cachier* as cerfs, as dains et as cheviruels. (FROISS., Chron. D., p. 1400, 208)

Sur la droite du tableau des attestations dans les bases apparaissent des onglets indiquant les ambiguïtés lexicales que le lemmatiseur répertorie sans les résoudre ; il s'agit des cas où la forme peut renvoyer à deux lemmes différents, ou davantage encore. Leur nombre est indiqué entre parenthèses, et il suffit de cliquer sur l'onglet pour voir apparaître les autres lemmes. Ainsi la forme *cachie* correspond-elle à deux lemmes possibles : CHASSER et CHASSIE, alors que la forme *cachier* est susceptible quant à elle de correspondre à trois lemmes : CACHER, CASSER et CHASSER. C'est là le résultat d'un traitement automatique : cela ne signifie pas

que les lemmes indiqués sont toujours pertinents pour la forme considérée en contexte. Par exemple, plus bas dans le tableau, on découvre que la forme *chace* correspond à cinq lemmes : CHASSE 1, CHASSE 2, CHÂSSE, CHASSER et CHAUD. Si les quatre premières propositions paraissent assez évidentes, la dernière est plus curieuse, et demanderait certainement confirmation. Seule l'étude contextuelle permettra de lever ces ambiguïtés, comme le montre l'examen des occurrences de *cachier* dans le *DMF* : les contextes (4) à (7) orientent clairement vers le lemme CHASSER, puisqu'il est question de l'activité de la chasse ou de pourchasser des animaux ou des ennemis ; le contexte (1) renvoie au lemme CACHER ; le contexte (2), avec le terme de *mine*, au lemme CASSER ; seul le cas (3) peut poser problème – en tout cas l'exemple n'a pas été retenu par le rédacteur – mais il s'agit vraisemblablement également de CASSER, avec le sens particulier de « écraser, assouplir une peau » dans le domaine de la peausserie.

Ces premiers éléments font donc apparaître une gestion assez fine de la variation formelle et des liens entre formes et lemmes, grâce à un programme développé dès 2001, au moment où il a été décidé de faire du *DMF* un dictionnaire véritablement électronique : l'informatique, qui devait servir à rassembler et exploiter les ressources lexicales, a également été sollicitée dans la structuration du dictionnaire lui-même, grâce au lemmatiseur LGeRM. Il nous faut nous arrêter sur l'histoire de son développement, et sur les différents éléments qui le constituent.

LGeRM est l'acronyme de « Lemmes, Graphies et Règles Morphologiques ». L'outil est né en 1986 dans le cadre du travail de DEA mené par Gilles Souvay, fruit d'une collaboration entre informaticiens du Centre de recherche en informatique de Nancy[2] et linguistes de l'Unité de recherche sur le français ancien (URFA) de l'université de Nancy II[3]. Il s'agissait de concevoir un système expert (à base de règles) ayant pour but de réduire la variation graphique de textes médiévaux, dans le cadre de travaux préparatoires au *DMF*. L'étude portait sur la

2. Aujourd'hui LORIA.
3. Aujourd'hui université de Lorraine. Mais l'URFA n'existe plus.

variation, hors variation verbale. Les premiers essais de mise en ligne du *DMF*, au début des années 2000, ont montré la difficulté que représentait le fait de trouver un mot dans le dictionnaire ; d'où l'idée de reprendre et de compléter les travaux de 1986. Le lemmatiseur a été présenté pour la première fois en 2004 au Congrès international de linguistique et philologie romane à Aberystwyth, au Pays de Galles (Souvay, 2004). Un article lui a été consacré en 2010 dans un numéro de la revue *TAL* consacré au traitement automatique des langues anciennes (Souvay et Pierrel, 2010).

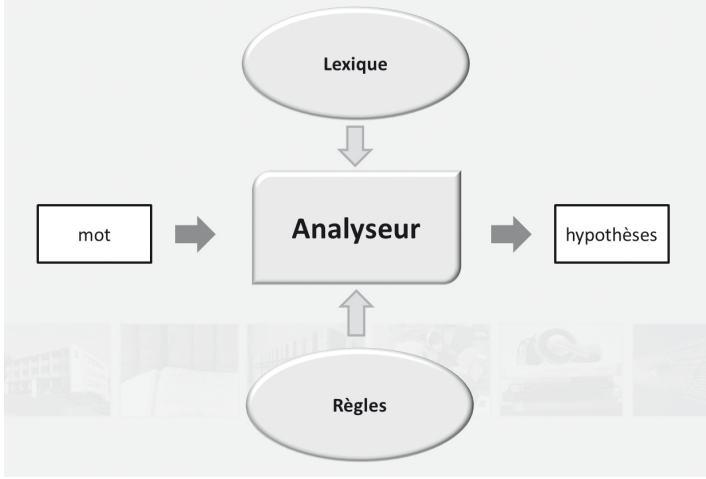


Fig. 6. Architecture du système

Lors de la recherche d'une entrée dans le dictionnaire, l'analyseur saisit 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.

L'analyseur

L'algorithme mis en œuvre est le suivant : si la graphie saisie est répertoriée dans le lexique, le lemmatiseur propose l'analyse. Si tel n'est pas le cas, il applique les règles sur la forme de départ, ce qui produit de nouvelles formes. On réitère

le processus sur les formes produites, le but étant de trouver une forme connue. Ainsi, pour le mot *maulvitiéz*, une première règle transforme le *z* final en *s* pour donner *maulvitiés* ; une seconde règle fait tomber le *l* pour donner *mauvitiés* ; enfin une troisième transforme le premier *i* en *e* pour donner *mauvetiés*, qui est une forme connue du lemme MAUVAISETÉ. Mais le système doit gérer le nombre de formes produites et l'arrêt de la production de formes. Il est inutile de produire trop de formes, car chaque règle permet une transformation de la forme de départ et l'application de plusieurs règles à la suite éloigne considérablement de la forme initiale, d'autant que les règles sont indépendantes les unes des autres. L'application d'un nombre trop élevé de règles sur un même mot génère une cascade de transformations qui conduisent à une forme généralement aberrante. Par exemple pour le mot *prouhomme*, analysé fin 2014, avant correction du lemmatiseur : une première règle réduit le double *m* en *prouhome* ; une seconde règle supprime le *h* et donne *prouome* ; enfin, une troisième transforme le *u* en *n* pour donner *pronome*, qui est une forme connue du lemme PRONOM. Chaque règle prise individuellement est logique, mais l'application des trois règles à la suite déforme complètement la forme de départ. Une étude des résultats a montré qu'à partir de trois règles appliquées, le taux de reconnaissance de la forme passait en dessous de 50 % et qu'il fallait alors vérifier les propositions faites par LGeRM.

Le lexique

Le lexique rassemble des graphies connues avec leur analyse. Il se présente sous la forme d'une liste de triplets (graphie, lemme, étiquette). Les étiquettes sont les grandes catégories grammaticales du *DMF*. Par exemple, pour la forme *amer*, il existe deux lemmes possibles : AIMER et AMER – ce qui donne les deux triplets suivants :

amer, AIMER, verbe
amer, AMER, adj.

Le lexique initial a été constitué à partir des exemples du *DMF*. Comme les articles du *DMF* sont balisés en XML, il était

aisé de constituer une liste initiale. Le lexique a par la suite été enrichi manuellement ou semi-automatiquement à partir de textes ou de corpus traités dans le cadre de collaborations formelles ou informelles, en lien avec les projets *DMF* et Frantext. L'enrichissement est toujours en cours, grâce à chaque nouveau texte traité par le lemmatiseur. En novembre 2014, celui-ci comportait un peu moins de 900 000 entrées.

Les règles

Dans LGeRM, la formulation « règle morphologique » est un raccourci qui englobe non seulement les règles morphologiques, mais aussi des règles de variation graphique. La structure générale d'une règle est la suivante : si des conditions sont remplies, alors on effectue une action.

si conditions *alors* action *finsi*

Les conditions peuvent porter sur la position d'un graphème dans le mot : en finale, en initiale ; sur le contexte du graphème : précédé de, suivi de… Elles peuvent concerner le lemme lui-même (verbal ou non, sur le suffixe du lemme…), ou encore la réussite ou non de l'application d'une règle. Le système permet de tester des hypothèses, mais on n'ajoutera la forme produite dans le flux des formes engendrées que si elle est pertinente. Il faut toujours avoir en tête que l'application d'un trop grand nombre de règles risque de déformer démesurément les mots.

Les règles sont regroupées par familles : règles de flexion (verbale ou nominale), règles de transformation de la flexion, règles d'archaïsations, règles de transcription des phonèmes… Il existe des règles classiques pour ramener une forme verbale à l'infinitif du lemme, une forme adjective à la forme du masculin singulier, etc.

si (en finale) alors DRONT → DRE finsi
pondront → (*pondre*, PONDRE, verbe)
si (en finale) alors DRONT → DRA finsi
pondront→ (*pondra*, PONDRE, verbe)
si (en finale) alors IVE → IF finsi
vive→ (*vif*, VIF, adjectif)

Mais l'infinitif du lemme (ou la forme de l'adjectif au masculin singulier) ne se trouvent pas forcément dans la base de connaissances. LGeRM teste alors une autre personne, un autre mode, un autre temps, un autre genre, un autre nombre… de la forme rencontrée :

si (en finale) alors DRÉS → DRA finsi
poundrés → (*poundra*, PONDRE, verbe)
si (en finale) alors IVE → IF finsi
vives → (*vive*, VIF, adjectif)

Il existe des règles pour la variation graphique et morphologique de la flexion du lemme. C'est le cas par exemple du *e* dit svarabhaktique pour les futurs et conditionnels présents.

si (en finale) et (précédé de [D,T,V]) alors ERAI → RAI finsi
ponderai → (*pondrai*, PONDRE, verbe)
si (en finale) alors NRA → NERA finsi
menra → (*menera*, MENER, verbe)
si (en finale) alors ES → EFS finsi
nes → (*nefs*, NEF, subst. fém.)

La base de connaissances contient également des règles de modernisation ou d'archaïsation des formes :

Y → I
fayre → (*faire*, FAIRE, verbe)

Elle contient des règles d'équivalence graphique :

C→ SS
mesfacent → (*mesfassent*, MÉFAIRE, verbe)

Ainsi que des règles d'agglutination avec l'adverbe, le pronom ou un élément formant :

si (précédé de TRES) alors TRES tombe finsi
tresadvisé → (*très,* TRÈS, adv.) + (*advisé,* AVISÉ, adj.)

La base de connaissances contient aussi des variations régionales, utiles par exemple pour le traitement d'un mot d'origine lorraine :

si (en finale) alors EI → É finsi
abandonei → (*abandoné*, ABANDONNER, verbe)

Au total, le système comporte environ 6 500 règles. Aux 200 règles initiales de 1986, se sont ajoutées les règles de la

flexion verbale, construites à partir des exemples du *DMF*, des corpus textuels (Frantext, Christine de Pizan, Froissart…) et de la BGV (Base de graphies verbales). Environ quatre cinquièmes des règles portent sur la flexion des verbes.

Les règles sont utilisées pour pallier les lacunes du lexique. Il n'est pas possible de produire un lexique exhaustif : la combinatoire est trop élevée, plus particulièrement dans le cas des verbes et pour les mots à plusieurs syllabes[4].

En 2007, dans le cadre d'un projet franco-britannique mené en partenariat avec les équipes travaillant sur l'édition électronique de Christine de Pizan, *The Making of the Queen's Manuscript* (Édimbourg) et de Froissart, *The Online Froissart* (Sheffield/Liverpool), nous avons intégré le lemmatiseur dans un environnement permettant d'aider à construire le glossaire d'un texte. LGeRM permet de lemmatiser un texte source encodé en XML/TEI. Le résultat de la lemmatisation peut être consulté de trois manières : en passant sur les mots du texte en continu, en interrogeant par forme, ou par lemme. L'outil permet de détecter des erreurs de transcription ou d'océrisation[5], des mots absents du *DMF*… ; il permet de réaliser une édition électronique à orientation lexicographique. Il est possible d'intervenir sur le résultat de l'analyse, de choisir parmi les hypothèses proposées

4. Ainsi le lemme CONNAISSANCE correspond-il à 44 formes attestées dans le corpus diachronique de Frantext (octobre 2014) : cognescence, cognissance, cognissanche, cognoeissance, cognoiscences, cognoiçance, conaisanche, congnioissance, congnoessance, connissanche, conoissances, cougnoissance… Pour trouver toutes les attestations de ce lemme dans Frantext, il faudrait utiliser l'expression régulière suivante :
[c|k|q][o|oi|e|eo|ei][n|nn|gn|ngn][oi|ai|i|i|oi|e|eo|e][s|ss|sc|sç|ç|c][i]?[en|an|ā|ē][s|ss|sc|sç][e][sz].

5. « La technique d'OCR (*Optical Character Recognition*) permet de situer et de reconnaître les chaînes de caractères dans une image, et donc de faire la conversion des mots qui peuvent ensuite être utilisés pour faire une recherche plein texte. Cette conversion est assurée automatiquement par un logiciel et fait l'économie de la retranscription manuelle, beaucoup plus chère. Les mots et chaînes de caractères stockés dans un fichier texte peuvent être réutilisés pour une nouvelle mise en page, exploités dans une base de données, etc. Le principe est la reconnaissance des différentes zones de la page et des caractères contenus dans les zones textuelles. Les caractères sont identifiés à partir de formes mémorisées par le logiciel et de termes déjà connus car présents dans le dictionnaire utilisé par l'outil. »
En ligne : http://www.bnf.fr/fr/professionnels/numerisation_boite_outils/a.num_conversion_mode_texte.html [consulté le 21 juin 2017].

par le lemmatiseur, de gloser certains termes… Le travail final peut être exporté sous forme d'index lemmatisé, de texte XML/TEI, voire de schéma d'article.

ABREUVER, verbe	3 attestations 2 formes	***abuvrees*** 1 att. 5va ***abuvrés*** 2 att. 5va 5vb
ACCÈS, subst.	13 attestations 2 formes	***acceps*** 1 att. 9rb ***accés*** 12 att. 9rb 9rb 9rb 9rb 9rb 14va 14vb 14vb 14vb 15ra 20ra 20ra
ACCROÎTRE, verbe	1 attestation 1 forme	***accroist*** 1 att. 26va
ACHE, subst.	10 attestations 2 formes	***ace*** 1 att. 11rb ***ache*** 9 att. 9va 9va 10ra 13vb 14vb 15vb 20rb 20vb 24vb

Fig. 7. Extrait d'un index lemmatisé

En 2010, LGeRM a été adapté à la langue du XVII^e^ siècle dans le cadre du projet européen Impact[6]. Ce projet avait pour but de fournir des outils pour l'océrisation[7] et l'interrogation des fonds anciens des bibliothèques nationales dans chacune des langues des partenaires (allemand, anglais, bulgare, espagnol, français, néerlandais, polonais, slovaque). Dans ce projet l'ATILF, en partenariat avec la BnF, était chargée de produire un lexique pour le français. Le travail a consisté à archaïser un lexique moderne MORPHALOU (entrées du *TLF* et leur flexion) et à le projeter sur un corpus textuel. Le corpus textuel était composé d'une centaine de textes issus de Frantext et de seize textes spécialement numérisés dans le cadre du projet pour constituer la « vérité-terrain » de l'expérimentation, en conservant leur graphie d'origine[8].

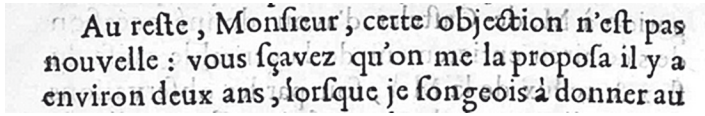

Au reſte, Monſieur, cette objection n'eſt pas
nouvelle : vous ſçavez qu'on me la propoſa il y a
environ deux ans, lorſque je ſongeois à donner au

Fig. 8. Extrait d'un texte du corpus de « vérité terrain »

6. *Improving Access to Text*, en ligne : http://www.impact-project.eu/ [consulté le 21 juin 2017].
7. Voir note 5.
8. Notamment les barres de nasalisation, *s* long et l'absence de distinction entre *i/j* et *u/v*.

Il a fallu adapter les règles de LGeRM à la morphologie et aux variations spécifiques de cet état de langue. Il s'agit en effet d'une période où l'on tend fortement à normaliser la graphie des mots, mais où celle-ci reste encore assez dépendante des choix des imprimeurs. On peut noter l'ajout de caractères étymologiques, *havons* pour *avons* du verbe AVOIR (latin *habere*) ou *poincter* pour POINTER, à partir du latin **puncta*. L'utilisation des diacritiques est en train de se mettre en place, mais ne correspond pas encore aux normes actuelles, comme dans *cinquiême*, ou manifeste un choix inverse de celui opéré par l'orthographe moderne pour *à*, forme du verbe AVOIR. L'adaptation de l'outil à ce projet et l'expérimentation sur des textes du XVII^e^ siècle lui ont permis de résoudre des formes telles que *hauoir*, *ſciẽce*, *neceßité*, *coꝰe*, et par là d'être utilisé également pour la période qui suit le moyen français afin de traiter les éditions anciennes.

Mais l'extension de l'utilisation de LGeRM s'est également opérée en amont de la période de référence du *DMF* (1330-1500). Néanmoins, comme les articles du *DMF* citent des dictionnaires tels que le Tobler-Lommatzsch et le Godefroy, qui portent sur l'ancien français, l'outil connait des graphies plus archaïques. Le *Dictionnaire électronique de Chrétien de Troyes* (*DECT*), qui a pris comme modèle le *DMF* et a été informatisé à l'ATILF, a également fourni des formes du XII^e^ siècle. La Base de graphies verbales complète le lexique avec des formes conjuguées pour la période de l'ancien français au français de la Renaissance, relevées dans des dictionnaires de référence et des éditions de textes. Enfin, le traitement des textes médiévaux au programme des agrégations de Lettres modernes, Lettres classiques et de Grammaire depuis 2010[9] a montré que l'outil était capable de traiter l'état de

9. Voici la liste des œuvres traitées par l'outil du *DMF* et mises en ligne sur le site du *DMF* : Charles d'Orléans, *Poésies* (2010-2011) ; Béroul, *Tristan* (2011-2012) ; Guillaume de Lorris, *Le Roman de la Rose* (2012-2013) ; *Le Couronnement de Louis* (2013-2014) ; *Le Roman d'Eneas* (2014-2015) ; Jean Renart, *Le Roman de la Rose ou de Guillaume de Dôle* (2015-2016) ; Christine de Pizan, *Le Livre du Duc des vrais amants* (2016-2017). On remarquera que les œuvres, en dehors des *Poésies* de Charles d'Orléans et du *Livre du Duc* de Christine de Pizan, n'appartiennent pas à la période de référence du *DMF* ; cependant, le lemmatiseur LGeRM et le *Dictionnaire* lui-même sont d'une grande utilité, même si le travail manuel demeure important après la phase de traitement automatique.

langue antérieur, même si la lemmatisation des textes d'ancien français offre un taux d'erreur plus important que pour le moyen français du fait de lacunes lexicales ou morphologiques, et de lemmes disparus ou inconnus du *DMF*. Ces expérimentations ont permis, non seulement de fournir aux candidats et préparateurs un texte lemmatisé et interrogeable avec précision, mais également d'améliorer les performances de l'outil.

Récemment, sous l'impulsion du projet Presto[10], qui a pour objectif d'étiqueter et de lemmatiser des textes de toutes périodes du français, nous avons été sollicités pour construire un lexique morphologique adapté au français du XVI^e^ siècle. Préalablement à ce travail, nous avons décidé de diffuser les ressources lexicales dont nous disposions sous licence *Creative Commons*. Deux lexiques sont désormais disponibles :

– Le lexique LGeRM médiéval est optimisé pour la période 1300-1500. Il comporte 880 192 entrées pour 66 976 lemmes. 142 687 graphies sont attestées dans Frantext, et 52 % des entrées sont attestées dans tous les corpus liés au *DMF*.
– Le lexique LGeRM XVI^e^-XVII^e^ est optimisé pour la période 1550-1700. Il comporte environ 3 millions d'entrées, dont seulement 116 161 formes sont attestées (3,9 %).

La différence dans les pourcentages d'attestation s'explique par le fait que des méthodes différentes ont été retenues dans la construction des lexiques. Le lexique médiéval a été construit par accumulation de formes, alors que le lexique XVI^e^-XVII^e^ l'a été par archaïsation d'un lexique moderne, ce qui a produit des formes théoriquement possibles, mais pas forcément attestées. Ces lexiques sont utilisés par le moteur de recherche de Frantext pour la recherche par lemme. Cette approche permet d'éviter de lemmatiser le corpus. Deux inconvénients sont néanmoins à signaler : la recherche produit du bruit (homographes), et le lexique possède des lacunes (formes absentes du lexique). Ainsi, la recherche du lemme AGNEAU dans les textes médiévaux grâce

10. Presto (*L'évolution du système prépositionnel du français : approche diachronique et quantitative*) est un projet ANR/DFG (2013-2016), coordonné par D. Vigier (Lyon 2). En ligne : http://www.hrionline.ac.uk/onlinefroissart [consulté le 21-06-2017].

à LGeRM permet de rassembler des exemples aux graphies très diverses :

Fig. 9a. Formulaire de recherche pour le mot AGNEAU

Fig. 9b. Flexion et variations médiévales d'AGNEAU

Pour Presto, l'adaptation au XVIe siècle a nécessité de prendre en compte ce qui se passe entre 1500 et 1550, qui sont les bornes de nos lexiques existants. Ces travaux en cours ont donné lieu à une publication (Diwersy, Falaise, Lay et Souvay, 2015).

On peut se demander si les lexiques LGeRM permettent vraiment d'interroger Frantext. Autrement dit, quel est le taux de couverture des lexiques ? Pour le savoir, nous avons observé leur contenu en les projetant sur les mots présents dans Frantext. Deux angles d'approche ont été envisagés : en termes de fréquence, et en termes de graphie. Dans la séquence « *les chas et les soris* », on dénombre 5 mots pour 4 graphies. Le mot *les* a une fréquence

de 2 pour 1 graphie. La comparaison par tranche chronologique de 50 années porte sur trois lexiques : médiéval, XVIe-XVIIe et moderne. Le premier graphique présente en abscisses le nombre de textes concernés. En termes de fréquence, on voit que chaque lexique couvre relativement bien la période pour laquelle il a été conçu. Le lexique XVIe-XVIIe reste assez performant pour le français moderne, en raison de sa construction par archaïsation à partir des formes modernes.

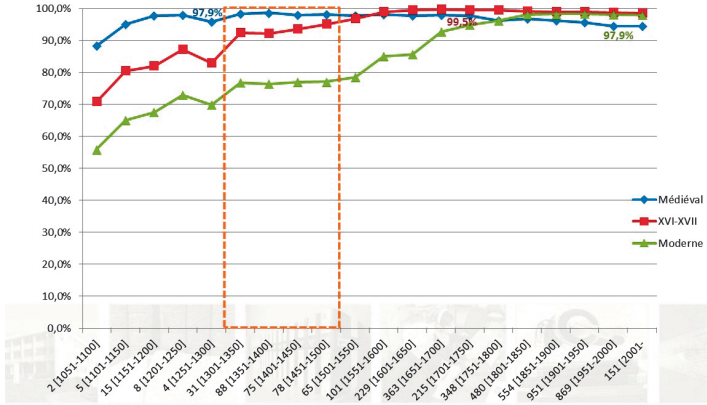


Fig. 10. Taux de couverture des lexiques : fréquences

En ce qui concerne les graphies, on remarque que le lexique médiéval et le lexique XVIe-XVIIe couvrent bien leur période. La construction automatique du lexique XVIe-XVIIe permet d'obtenir de meilleurs taux de couverture. S'agissant du lexique médiéval, on note un creux pour le repère 1251-1300. Cela est dû à un effet de corpus : seulement quatre textes, dont un texte en particulier, les *Actes de Ferry III, duc de Lorraine* qui représente 78 % des mots et possède un marquage dialectal lorrain fort, qui n'est pas encore pris en compte dans le lexique. Pour le français moderne, le taux de couverture est inférieur à 70 %, ce qui ne paraît pas très bon : il conviendrait donc de réaliser une étude plus poussée. La fréquence des mots étrangers et des noms propres est un début d'explication ; il manque aussi sans doute des lemmes.

Enfin, le graphique montre que les courbes du lexique médiéval et du lexique XVI^e^-XVII^e^ se croisent en 1550. Il serait sans doute intéressant d'étudier plus finement ce phénomène. Une première analyse a été effectuée sur les mots commençant par la lettre A ; les premiers résultats semblent indiquer que la nomenclature de lemmes joue un rôle plus important que la variation graphique. Les travaux liés au projet Presto devraient permettre de valider cette hypothèse.

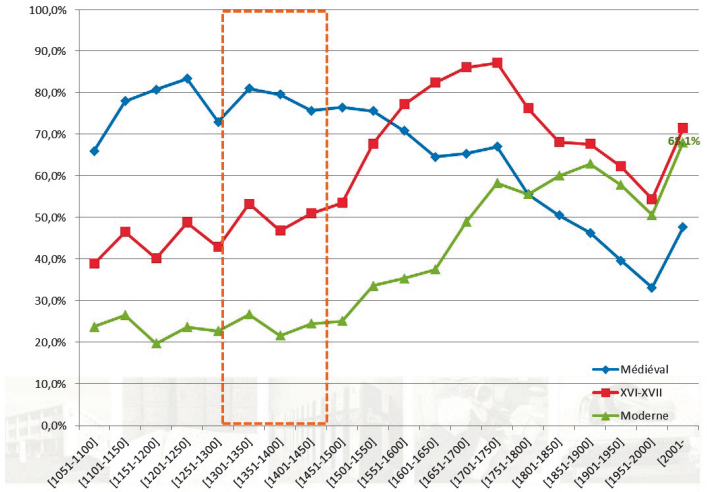


Fig. 11. Taux de couverture des lexiques : graphies

Si le bilan est nécessairement provisoire, avec des développements toujours en cours, LGeRM est tout de même à recommander en raison de sa capacité d'adaptation et de la souplesse d'utilisation dont il fait preuve en fonction des objectifs poursuivis. La lemmatisation automatique par LGeRM apporte déjà, dans ses résultats bruts, des éléments de vérification du texte transcrit de première importance. Ce qui n'est pas reconnu ou pose problème est souvent lié à des erreurs de transcription ou à des particularités de la copie. Il reste que pour pouvoir utiliser de façon certaine les matériaux obtenus par traitement automatique, il est nécessaire d'opérer non seulement une

longue et fastidieuse « désambiguïsation » dans le cas des propositions multiples de lemmes – la forme *appel* peut renvoyer au verbe APPELER ou au substantif APPEL, la forme *amer* à l'adjectif AMER ou au verbe AIMER –, mais également une vérification des lemmes choisis par le lemmatiseur pour pouvoir débusquer une forme mal interprétée – une forme *elle* qui n'est pas le pronom personnel sujet féminin, mais une graphie inhabituelle du lemme AILE, etc. Ce sont les étapes incontournables de vérification/validation des choix automatiques. Elles ont déjà été facilitées, sur le plan ergonomique, en faisant évoluer l'interface, et en partie systématisées, grâce à des procédures de tri ou de prise en compte du contexte linguistique. Depuis 2009, l'outil s'améliore au gré des projets divers qui l'utilisent, projets hébergés ou projets portés par l'équipe du *DMF*. Il n'est plus exclusivement lié au *Dictionnaire du moyen français*, bien qu'il reste au cœur de son fonctionnement et de ses développements. 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 correspondante (ancien français et français des XVI^e^ et XVII^e^ siècles), complémentaire des outils d'étiquetage morphosyntaxique.

Références bibliographiques

Ressources électroniques

DMF = Dictionnaire du moyen français, version 2012, ATILF/CNRS - Université de Lorraine. En ligne : http://www.atilf.fr/dmf

Base textuelle Frantext, ATILF/CNRS - Université de Lorraine. En ligne : http://www.frantext.fr

LGeRM = Lemmes Graphies et Règles Morphologiques, ATILF/CNRS - Université de Lorraine. En ligne : http://www.atilf.fr/LGeRM/

BGV = Base de graphies verbales, ATILF/CNRS - Université de Lorraine ; LFA - Université d'Ottawa. En ligne : http://www.atilf.fr/bgv/

Le Réceptaire de Jean Pitart, projet coordonné par Sylvie Bazin-Tacchella. En ligne : http://www.atilf.fr/dmf/JeanPitart

Textes destinés à la préparation du concours de l'agrégation (Sylvie Bazin-Tacchella et Gilles Souvay)

Charles d'Orléans. En ligne : http://www.atilf.fr/dmf/CharlesOrleans

Christine de Pizan, *Le Livre du Duc des vrais amants* (2016-2017). En ligne : http.www.atilf.fr/dmf/pizan/VraisAmants

Le Couronnement de Louis (2013-2014). En ligne : http://www.atilf.fr/dmf/CouronnementLouis

Jean Renart, *Le Roman de la Rose ou de Guillaume de Dôle* (2015-2016). En ligne : http://www.atilf.fr/dmf/RomanRoseGuillaumeDole

Guillaume de Lorris, *Le Roman de la Rose* (2012-2013). En ligne : http://www.atilf.fr/dmf/RomanRoseStrubel

Le Roman d'Eneas (2014-2015). En ligne : http://www.atilf.fr/dmf/RomanEneas

Béroul, *Tristan* (2011-2012). En ligne : http://www.atilf.fr/dmf/Beroul/

Collaborations

Édition numérique des *Chroniques de Froissart*, University of Sheffield, University of Liverpool, Arts & Humanities Research Council. En ligne : http://www.hrionline.ac.uk/onlinefroissart

Christine de Pizan, *The Making of the Queen's Manuscript*. En ligne : http://www.pizan.lib.ed.ac.uk/

Communications et articles

Bazin-Tacchella, Sylvie, « Le "Réceptaire attribué à Jean Pitart" (XIVe siècle) : projet d'une édition et d'un glossaire électroniques », dans Ducos, Joëlle (dir.), *Sciences et langues au Moyen Âge. Wissenschaften und Sprachen im Mittelalter*, Heidelberg, Universitätsverlag Winter, 2012, p. 269-286.

Bazin-Tacchella, Sylvie et Souvay, Gilles, « Le *Dictionnaire du moyen français* : la version *DMF* 2010 », dans Casanova Herrero, Emili et Calvo Rigual, Cesáreo (dir.), *Actes del 26é Congrés de Lingüística i Filologia Romàniques (València, 6-11 de setembre de 2010)*, Berlin, De Gruyter, vol. VIII, 2013, p. 4452-4462.

Diwersy, Sascha, Falaise, Achille, Lay, Marie-Hélène et Souvay, Gilles, « Traitements pour l'analyse du français préclassique », *22e Conférence sur le Traitement Automatique des Langues naturelles*, Caen, 2015.

Gerner, Hiltrud, « Constitution et évolution des corpus textuels et lexicaux à l'ATILF. Interconnexion des ressources », dans Kunstmann, Pierre et Stein, Achim (dir.), *Le Nouveau Corpus d'Amsterdam. Actes de l'atelier de Lauterbad (23-26 février 2006)*, Stuttgart, Steiner, 2007, p. 101-109.

Martin, Robert, « Pour un dictionnaire du moyen français », dans Wunderli, Peter (dir.), *Du Mot au Texte. Actes du IIIe Colloque international sur le moyen français* [1980], Tübingen, Gunter Narr, 1982, p. 13-24.

Martin, Robert, Gerner, Hiltrud et Souvay, Gilles, « Présentation de la seconde version du *DMF* (*Dictionnaire du moyen français*) », dans Iliescu, Maria *et al.* (dir.), *Actes du XXVe Congrès international de*

linguistique et de philologie romanes (Innsbruck, 3-8 septembre 2007), Tübingen, Niemeyer, 2010, p. 213-220.

MARTIN, Robert et SOUVAY, Gilles, « Le *Dictionnaire du moyen français*, *DMF* 2 (Note d'information) », *Comptes rendus des séances de l'année 2008*, Académie des inscriptions et belles-lettres, janvier-mars 2008, p. 49-57.

PIERREL, Jean-Marie et BUCHI, Éva, « Research and Resource Enhancement in French Lexicography: the ATILF Laboratory's Computerised Resources », dans BRUTI, Silvia *et al.* (dir.), *Perspectives on Lexicography in Italy and Europe*, Newcastle upon Tyne, Cambridge Scholars Publishing, 2009, p. 79-117.

SOUVAY, Gilles, « LGeRM : un outil d'aide à la lemmatisation du moyen français », *Actes du XXIV^e Congrès international de linguistique et de philologie romane* (Aberystwyth, Pays de Galles, 1-6 août 2004), Tübingen, Niemeyer, 2007, t. I, p. 457-466.

SOUVAY, Gilles et PIERREL, Jean-Marie, « LGeRM : lemmatisation de mots en moyen français », *Traitement Automatique des Langues*, n° 50, 2010/2, p. 149-172.

SOUVAY, Gilles, « Des exemples des possibilités offertes par le *Dictionnaire du moyen français* », dans TROTTER, David (dir.), *Present and Future Research in Anglo-Norman: Aberystwyth Colloquium, Juillet 2011*, Aberystwyth, The Anglo Norman Online Hub, 2012, p. 163-171.

SOUVAY, Gilles et BAZIN-TACCHELLA, Sylvie, « Construction assistée de glossaires avec les outils du *DMF* », dans CASANOVA HERRERO, Emili et CALVO RIGUAL, Cesáreo (dir.), *Actes del 26é Congrés de Lingüística i Filologia Romàniques (València, 6-11 de setembre de 2010)*, Berlin, De Gruyter, t. VIII, 2013, p. 4682-4691.

TROTTER, David, « Configurer le ou les sens en moyen français : Problème sémantique et défi lexicographique », dans BLUMENTHAL, Peter (dir.), *Beiheft zur Zeitschrift für französische Sprache und Literatur*, Stuttgart, Steiner, 2010, p. 153-170.

Herméneutique des similarités dans le *DFSM* : une expérience

Xavier-Laurent Salvador, Fabrice Issac & Marco Fasciolo
Université Paris XIII

Contexte

Nous nous proposons de donner ici les éléments théoriques et la méthodologie permettant l'élaboration d'un outil automatique capable d'explorer un dictionnaire en langage naturel. Nous nous intéresserons, pour ce faire, à la constitution d'une ressource numérique qui enregistre autant les propriétés sémantiques inhérentes au lexique à un instant donné de son histoire qu'un positionnement relatif des unités lexicales par rapport à l'ensemble des mots de la langue. S'agissant d'explorer un dictionnaire en langage naturel, nous distinguons de fait deux degrés dans l'identification de relations :

1. un niveau fondamental, qui identifie l'existence d'une connexion entre les mots ;
2. la caractérisation de ce lien de connexion.

Traditionnellement, les outils d'accès aux entrées d'un dictionnaire utilisent (i) l'ordre alphabétique des vedettes, (ii) un moteur de recherche, plus ou moins sophistiqué. Le premier revient à une simple transposition de l'utilisation d'un dictionnaire physique, le second permet d'effectuer des recherches potentiellement très fines, mais nécessitant une bonne connaissance de la macrostructure. Par ailleurs aucune information contextuelle n'est fournie, et le résultat prend la forme d'un article isolé.

Afin de contextualiser l'ensemble des informations, et non pas seulement la vedette, on pourrait concevoir l'élaboration

d'une représentation qui ait comme base celle que nous faisons de notre propre lexique, le lexique mental. En plus de relier sémantiquement les concepts les uns aux autres, l'accès à l'information linguistique s'en trouve amélioré. Lors de la production d'un discours, le choix d'un mot ou d'une expression se fait à un rythme très soutenu, avec un taux d'erreur très faible[1]. Des expériences en psycholinguistique portant sur l'organisation du lexique mental (Aitchison, 2003) montrent que les relations entre les éléments du lexique sont de deux types, soit intrinsèques, soit associatives (Levelt, 1989).

Les relations intrinsèques, ou catégorielles, contiennent des informations linguistiques sur l'unité lexicale elle-même. On peut décomposer les relations intrinsèques en relations :

– sémantiques (elles concernent la synonymie, l'antonymie, l'hyperonymie, l'hyponymie, la méronymie) ;
– morphologiques ;
– phonologiques (pour les mots commençant ou se terminant par les mêmes phonèmes : p. ex. *travail* et *traverse*).

Les relations associatives, de leur côté, regroupent les unités dont la fréquence d'apparition dans un même contexte est importante. À *ouvrier* on associera par exemple *usine* ou *travail*. Ce type de relations dépend de la connaissance qu'a le sujet du monde qui l'entoure. Les relations entre les éléments mentaux forment ainsi un réseau dans lequel les nœuds ne sont pas les mots eux-mêmes, mais leurs sens particuliers. Dans cette représentation, une collocation n'est pas stockée comme une association de mots, mais comme une unité individuelle à part entière.

L'analyse par la grammaire traditionnelle des relations entre les lexies en termes d'antonymie ou de synonymie s'intéresse au second niveau. En revanche, l'analyse grammaticale proposée par la grammaire scolaire, voire – dans une moindre mesure – par

1. Le choix d'un mot se fait au rythme de 2 à 5 mots par seconde avec un taux d'erreur inférieur à 1 pour 2 000, alors qu'on estime le nombre de mots que connaît un locuteur moyen à 35 000. Cette vitesse est encore plus élevée en lecture (Levelt, 1989).

la stylistique, en termes de champs sémantiques ou notionnels, s'intéresse au premier niveau. C'est à ce titre que, dans le discours critique du texte littéraire, *naufrage* et *débarquement* peuvent être analysés comme appartenant tous deux au « vocabulaire de la mer[2] » indépendamment du fait que ces deux termes renvoient à des réalités connotées de manière pour le moins antinomique, puisque le premier évoque l'échec et le second, le succès. Toutefois, la « relation de pertinence » entretenue par le discours critique entre cette paire nominale et l'hyperonyme maritime est validée par la présence manifeste, dans une décomposition des unités sémantiques de chacun des termes, du sème /mer/.

Nous présupposons donc qu'il est pertinent d'établir deux niveaux d'analyse des relations entretenues par les mots entre eux : la connexion d'une part, et d'autre part la caractérisation du degré de similarité. Le premier niveau est binaire : la connexion existe ou n'existe pas. Le second niveau, en revanche, est scalaire : il va d'une similarité nulle, d'un certain degré de pertinence, jusqu'à une similarité maximale, qui coïncide avec la synonymie. Dans le domaine de la dictionnairique, la distribution du vocabulaire en « domaines de sens » renvoie à une classification des objets de la langue en fonction de leur apparition dans des contextes d'emplois que le corpus d'exemples rend manifeste. Ainsi, si le mot *gingembre* est classé dans le domaine de la botanique (c'est une plante) et de la pharmacopée (c'est un ingrédient), c'est en vertu du fait que son contexte d'emploi fréquent justifie la caractérisation de la polysémie du terme. Il s'agit donc là d'une analyse de discours opérée par le lexicographe, et non pas de la manifestation d'une propriété inhérente ni au mot de la langue, ni à l'objet. Ainsi, ce n'est pas parce que l'on consomme du gingembre qu'il est consommable, mais l'emploi du terme dans des livres de recettes et de médecine témoigne de ce que cet aspect « consommable » est bien une propriété discriminante de la plante, et que le mode, ou plutôt la destination, de cette consommation varie en fonction des effets. À ce stade de la description de la relation entretenue

2. Voir Georges Molinié (1995).

par les sens d'une unité polysémique, il semble que nous pouvons considérer qu'au niveau (1) il existe une connexion entre *gingembre* (bot.) et *gingembre* (pharm.), et que cette connexion peut être caractérisée par (2) un haut degré de similarité, sans pour autant parler, en l'occurrence, de synonymie.

Il existe en ancien français le terme « zédoaire », répertorié par le *Dictionnaire du français scientifique médiéval* (*DFSM*), qui présente les mêmes propriétés que le gingembre aussi bien du point de vue de la nature (c'est une racine), du goût (c'est aigre) et de la distribution des emplois. Il existe donc (1) une forte connexion entre *gingembre* (bot.) et *zédoaire* (bot.), et cette connexion peut être caractérisée par (2) un très haut degré de similarité : c'est une synonymie. En revanche, la distribution des relations de connexion entretenues par les unités croisées par domaine est moins forte. L'ensemble du corpus des relations entretenues par chacun des niveaux sémantiques de chaque unité polysémique, et des relations entretenues entre chaque unité est documenté, dans la ressource comme dans le dictionnaire, par le corpus des exemples qui illustrent le travail définitoire du lexicographe.

La caractérisation des relations de similarité, ou *calcul de distance*, peut être schématisé de la façon suivante :

Mot 1 – (connexion, degré de similarité) – Mot 2

La similarité est une information pertinente pour la description d'un état de langue, car elle renseigne sur « ce que veut dire le lexicographe » et qui n'apparaît pas évident pour le lexicographe lui-même. Cela est d'autant plus vrai lorsque l'on travaille sur un état ancien de langue d'où les mots, mais aussi les choses ont disparu aujourd'hui et lorsque, comme dans notre cas, il y a deux niveaux de lexicographes. Nous rejoignons finalement la pensée du linguiste John Langshaw Austin qui écrivait (Austin, 1956) :

> *[…] our common stock of words embodies all the distinctions men have found worth drawing, and the connexions they have found worth marking, in the lifetimes of many generations: these surely are likely to be more numerous, more sound, since they have stood up to the long test of the survival of the fittest,*

> *and more subtle, at least in all ordinary and reasonably practical matters, than any that you or I are likely to think up in our armchairs of an afternoon – the most favoured alternative method.*

La méthode que l'auteur propose pour élucider les « distinctions conceptuelles que les hommes ont pu marquer au cours de nombreuses générations » est la suivante :

> *First we may use the dictionary – quite a concise one will do, but the use must be thorough. Two methods suggest themselves, both a little tedious, but repaying. One is to read the book through, listing all the words that seem relevant; this does not take as long as many suppose. The other is to start with a widish selection of obviously relevant terms, and to consult the dictionary under each: it will be found that, in the explanations of the various meanings of each, a surprising number of other terms occur, which are germane though of course not often synonymous. We then look up each of these, bringing in more for our bag from the « definitions » given in each case; and when we have continued for a little, it will generally be found that the family circle begins to close, until ultimately it is complete and we come only upon repetitions. This method has the advantage of grouping the terms into convenient clusters – but of course a good deal will depend upon the comprehensiveness of our initial selection.* (*Ibid.*)

Cette réflexion, que nous appliquons au domaine de la dictionnairique, permet d'établir des connexions entre les mots du lexique et des degrés de similarité entre eux.

Enjeux épistémologiques

Du côté philologique

Une partie de notre travail a pour point de départ la réalisation du *DFSM*. Cet ouvrage, consacré à la langue de spécialité médiévale, repose sur un corpus scientifique constitué par des traducteurs ou des vulgarisateurs qui travaillent presque comme des lexicographes au Moyen Âge, ce qui suppose de prendre en compte ces réflexions et pratiques de locuteurs.

La conception du dictionnaire permet ainsi une réflexion sur les processus de genèse et de néologie d'une terminologie.

Il doit aboutir à une meilleure connaissance du français médiéval dans ses usages spécialisés, mais aussi donner des outils d'étude et une méthodologie pour apprécier les modalités de création et en rendre compte dans un dictionnaire conçu comme une représentation d'évolution linguistique. Le développement technique au sein de l'équipe CréaLScience, de plus, complète les recherches épistémologiques, rendant au demeurant hommage à l'étymologie même du terme « informatique » – qui donne forme et corps aux pensées dont la nature est par essence informe.

Le *DFSM* présente donc une structure lexicographiquement stratifiée : d'un côté, il y a les rédacteurs médiévaux des définitions ; de l'autre côté, il y a l'équipe actuelle qui traduit ces définitions en français contemporain. Tout l'enjeu épistémologique du dictionnaire réside dans l'articulation de ces deux niveaux lexicographiques et dans la mise en forme des analyses du corpus multilingue qui en découle. Au premier niveau, parler de lexicographes pour les rédacteurs médiévaux peut étonner. Cet étonnement, cependant, est hors de propos. On trouve dans la littérature encyclopédique médiévale, comme dans les traductions de la Bible d'ailleurs, un ensemble de marqueurs caractéristiques de la glose encyclopédique : *c'est assavoir, si come dit le maistre, sicome est*,... Ces marqueurs embrayent sur une prise de parole originale des traducteurs afin d'introduire des paraphrases explicatives de certains mots, renvoyés en mention autonymique (Authier-Revuz, 1995 et 1998). Lorsque, dans un discours français, un terme français est autonyme et assorti d'une glose encyclopédique, n'est-on pas en droit de considérer que le traducteur a œuvré, en l'occurrence, en spécialiste de la langue ? Ne peut-on même considérer qu'il agit en lexicographe lorsqu'il renseigne son lecteur sur les emplois de *zurbe* ou sur les façons d'observer les apophtegmes ?

Dans le cadre de la définition des rapports qu'il entretient avec l'énonciation du texte traduit, le traducteur est indéniablement un sujet de l'énonciation. Il devient une forme d'interface de coïncidence entre le vouloir-dire du texte source et les horizons d'attente du texte traduit. Sa réflexion sur le lexique de la langue

source s'apparente fortement à celle menée par un lexicographe s'agissant de la recherche de la nature, du sens et des conditions syntaxiques d'avènement de ce dernier dans la langue cible.

Enjeu dictionnairique

Le traducteur intervient donc comme un lexicographe à part entière, chaque emploi qu'il fait de chaque unité du système étant le fruit d'une réflexion issue à la fois d'un enseignement universitaire et d'un souci d'enseignement scientifique ; et le texte traduit apparaît comme un recueil de prises de position lexicographiques à partir de relations méronymiques établies non pas entre deux langues, mais entre le latin, langue de communication savante, et le français en situation de diglossie. La traduction se conçoit *a fortiori* comme un discours rapporté, en témoigne le discours attributif qui l'introduit : « x (= l'auteur du texte original) dit que ». L'autorité de l'auteur / traducteur sur sa propre production est ainsi mise entre parenthèses, puisque le traducteur se présente comme citant et reprenant les mots du lexique de l'auteur premier, soit qu'il le suive au point de calquer le vocabulaire français sur le lexique latin – c'est le calque savant – soit au contraire qu'il le juge en inadéquation avec le lecteur français, et qu'il l'abandonne – c'est la glose paraphrastique. Ainsi, entre paroles rapportées et appropriation du discours d'un autre, la traduction scientifique engagée dans une réflexion métalinguistique sur le lexique construit un discours autonome. À l'intérieur de ce discours, nous remarquons des nœuds opaques qui posent le problème de la mention autonymique de quelques unités sémantiques, et c'est là sans doute que nous rejoignons de plein pied la problématique de la néologie et de son repérage. Un tel phénomène définit le paradoxe des unités lexicales placées en mention autonyme en contexte traductologique. Le traducteur travaille sans cesse à rendre son énoncé pertinent dans le cadre d'un enseignement fondé sur la transmission sémantique. Nous retrouvons dans ce phénomène la problématique du dictionnaire bilingue : construire un vocabulaire spécialisé (une définition), dont le sens est donné explicitement par l'introduction de xénismes en mention autonymique (l'entrée principale) et

dont le contexte se charge de saturer le signifié par le biais de mentions correctives.

Au second niveau lexicographique (celui de l'équipe actuelle), deux problématiques principales émergent.

La première d'entre elles concerne le rapport avec le premier niveau de lexicographes-traducteurs, à savoir l'identification de « ce qu'ils veulent dire ». Cette question, cruciale, ne peut pas être laissée à l'appréciation de chacun, et un ensemble de contraintes doivent peser sur la mise en forme du corpus des définitions. C'est un élément d'autant plus important que le nombre d'auteurs est élevé. L'équipe du *DFSM* doit maîtriser le métalangage de description et s'accorder sur un ensemble de termes propres à la description lexicographique. L'idée consiste à créer un lexique des termes autorisés dans le cadre de l'article de manière à ce que chacun de ces termes fonctionne comme un signal adressé à l'*uptake* du calculateur pour l'enregistrement d'un ensemble de traits descripteurs des relations sémantiques et des propriétés.

Il est également essentiel d'assurer l'autosuffisance du dictionnaire par le recoupement des données. Ainsi, le traitement de la néologie sera d'autant plus facilité qu'il existera un ensemble de lemmes orphelins, ensemble qu'il sera alors aisé de traiter.

La mise en œuvre de l'architecture métalexicographique s'accorde avec un travail sur l'article. Une définition doit intégrer des propriétés intrinsèques : « ce qu'est le x », et des propriétés extrinsèques : « à quoi sert x », par exemple dans le cas d'un instrument, ou encore « d'où vient x » dans le cas d'une maladie. Ainsi, la définition de *zodiaque* ne peut-elle pas reprendre les termes d'un dictionnaire contemporain[3] :

> *Zodiaque* : cercle situé sur le plan de l'écliptique et autour duquel évoluent le Soleil, la Lune et les planètes.

3. Larousse en ligne : http://www.larousse.fr/dictionnaires/francais/zodiaque/83170?q=zodiaque#82170 [consulté le 21 juin 2017].

Elle doit restituer le champ des savoirs relevant de la période couverte. En l'occurrence, dans la définition contemporaine, le terme *écliptique* est hétérogène et anachronique, alors que l'absence des renvois vers *astre* ou *maison*, qui sont les hyponymes directs de *zodiaque* dans l'astronomie médiévale (Boudet, 2006), est dommageable pour la compréhension du concept médiéval.

L'autre problématique, qui touche le second niveau lexicographique susmentionné, concerne la question des « nomenclatures », terme par lequel nous entendons « la liste des lemmes présents dans le dictionnaire ». Dans le cadre de la rédaction d'un dictionnaire d'histoire des sciences anciennes, la première difficulté que nous rencontrons réside dans la constitution de ladite nomenclature (Ducos, 2006). La forme du lemme n'est pas problématique lorsque la langue du dictionnaire est homogène avec la langue décrite, même lorsqu'il y a xénisme. Les choses peuvent devenir plus compliquées dans le cadre de la rédaction d'un dictionnaire bilingue (Kocourek, 1991). Mais elle sont plus problématiques encore s'agissant de la période médiévale, où la langue connaît trois degrés de variation : une variation dans le temps, variation diachronique ; une variation dialectale, variation diatopique ; une variation idiolectale, liée à l'intervention des copistes qui sont eux-mêmes des adaptateurs du texte original. À celles-là s'ajoute la variation graphique, qui peut en premier lieu laisser croire à l'existence de plusieurs lexèmes, là où il ne s'agit que d'une variante dialectale, mais également rendre difficile la création de règles claires pour le choix du terme à intégrer à la nomenclature ; ainsi, au même moment, en France, trouve-t-on *bourraiche, bourrache, borrache, borraige* pour le lemme « *bourrache* ».

Du côté de l'exploration dictionnairique

Dans le cadre de notre travail, une question se révèle cruciale : comment explorer un dictionnaire de langue naturelle ? L'interrogation peut être précisée de cette manière : comment relier les définitions d'un tel dictionnaire ? Il existe des solutions impraticables. Considérons les définitions suivantes :

Couteau :
Instrument tranchant servant à couper, composé d'une lame et d'un manche.

Chat :
Petit mammifère familier à poil doux, aux yeux oblongs et brillants, aux oreilles triangulaires et griffes rétractiles, qui est un animal de compagnie.

Sosie :
Personne qui a une parfaite ressemblance avec une autre.

Envieux :
Qui éprouve de l'envie.

Vendre :
Céder à quelqu'un en échange d'une somme d'argent.

Rapidement :
D'une manière rapide.

Une première hypothèse consisterait à les rapprocher sur la base de leur forme par rapport à la catégorie grammaticale de l'entrée. Dans la définition d'un nom comme *couteau*, par exemple, on peut distinguer un hyperonyme (un autre nom) et une portion de différences spécifiques, alors que dans la définition d'un adjectif comme *envieux* on peut distinguer un terme de relation en première position (en l'occurrence le relatif *qui*…) et le corps de la définition, qui contient le signifié.

Une deuxième hypothèse consisterait à rapprocher ces définitions sur la base de leur forme par rapport au type d'objet décrit. La catégorie des instruments, par exemple, sera définie par un hyperonyme suivi de la fonction, alors que les éléments des listes naturelles (comme les fleurs) recevront une description centrée sur leur forme. Ce type de solution, cependant, n'est pas envisageable. Tout d'abord, ce niveau de finesse est un obstacle pour l'exploitation automatique. Comparons les définitions de *sosie* (qui est un nom relationnel) et *couteau* (qui est un nom ponctuel). Le premier terme de ces définitions (à savoir, *instrument* et *personne*) est-il un vrai hyperonyme ? Certes, instrument est un hyperonyme de *couteau*, car l'assertion « un couteau est un instrument » est informative. Mais *personne* ne peut pas être considéré comme un hyperonyme de *sosie* : c'est,

dans la définition, un argument de « avoir une ressemblance parfaite » (le signifié de *sosie*), qui remplit la fonction d'un pronom. La définition de *sosie* est donc plus proche de celle de *envieux* que de celle de *couteau*. Il nous paraît très difficile de modéliser informatiquement une telle intuition.

Ensuite – ce qui est plus important – le rapport entre la forme de la définition et la catégorie grammaticale de la vedette, ou le rapport entre la forme de la définition et le type d'objet décrit, sont précisément le genre de choses qu'on voudrait étudier à travers une exploration du dictionnaire. Si l'on veut fournir des outils pour une telle exploration, il faut par conséquent se fonder sur des critères plus élémentaires.

Nous proposons de nous inspirer des travaux menés dans le champ de la désambiguïsation, et notamment de l'algorithme de Lesk.

> *I am trying to decide automatically which sense of a word is intended (in written English) by using machine readable dictionaries, and looking for words in the sense definitions that overlap words in the definition of nearby words. […] To consider the exemple in the title [How to tell a « pine cone » from an « ice-cream cone »], look at the definition of pine in the* Oxford Advanced Learner's Dictionary of Current English*: there are, of course, two major senses, « kind of evergreen tree with needle-shaped leaves… » and « waste away through sorrow or illness… » And cone has three separate definitions: « solid body which narrows to a point… » « something of this shape whether solid or hollow… » and « fruit of certain evergreen trees… » Note that both evergreen and tree are common to two of the sense definitions: thus a program could guess that if the two words pine cone appear together, the likely senses are those of the tree and its fruits*. (Lesk, 1986)

Le problème ici abordé peut être formalisé de la façon suivante. Soit :

> un mot *M* qui connecte deux sens différents : s_1 et s_2

Imaginons rencontrer *M* dans deux co-textes :

> $C_1 = m_{1.1}, m_{1.2}, \ldots$ où les $m_{1.i}$ sont des mots
> $C_2 = m_{2.1}, m_{2.2}, \ldots$ où les $m_{2.i}$ sont des mots

En ce cas, nous sommes confrontés à des trigrammes, comme :

$m_{1,j}\ M\ m_{1,j+2}$ et $m_{2.k}\ M\ m_{2.k+2}$

La question est la suivante : comment déterminer le sens de *M* dans chaque séquence ? Si à chaque sens s_1 et s_2 de *M* correspond une définition D différente : D_1 et D_2, alors on cherchera la présence des trigrammes des co-textes dans les définitions, et on choisira le sens correspondant.

Cette solution présuppose une homologie entre l'emploi d'un mot (*i.e.* son occurrence dans un co-texte) et les mots de la définition qui préside à cet emploi. Ce présupposé mérite qu'on s'interroge : quel est le rapport existant entre les mots employés dans la définition et dans les exemples ? En réalité, ce rapport ne peut pas se réduire à un simple recouvrement, parce qu'une définition doit fournir un schéma (un modèle) qui permette de générer tous les emplois possibles.

Cependant et quoi qu'il en soit, la solution proposée par Michael Lesk nous offre une suggestion importante. Explorer un dictionnaire en langage naturel signifie : (i) prendre une entrée ; (ii) décomposer sa définitions en mots ; (iii) en envisageant chaque mot selon trois dimensions :

– occurrences (ou *tokens*) ;
– lemmes ;
– catégories morphosyntaxiques ;

et enfin (iv) regarder comment les séquences de mots se distribuent dans les définitions des autres entrées.

Le mode opératoire est le suivant : à partir d'une vedette, on décompose sa définition en atomes de sens, et on explore la façon dont ceux-ci se propagent dans les définitions des autres vedettes. Cela revient à explorer leurs échos sémantiques dans le dictionnaire. Il est à noter que cette exploration n'est pas statique et globale, mais dynamique et locale, car elle varie selon le point d'entrée choisi.

Remarquons tout d'abord que, d'un point de vue linguistique, la solution des « segments partagés » est grossière, car elle se fonde sur des séquences de mots sans présupposer aucune

structure. Il faut cependant noter que ce défaut est corrigé par un principe de coopération *à la Grice*. Nous supposons que les lexicographes sont des êtres rationnels et coopératifs, et donc que leurs définitions ne sont pas des suites de mots mis au hasard, mais bien des textes cohérents. Cette assomption autorise à émettre l'hypothèse que le critère de rapprochement des définitions, en pratique, peut se réduire simplement aux plus longs segments partagés. Nous faisons le pari que ces segments constituent des morceaux pertinents de la syntaxe des définitions. Le principe de coopération susmentionné joue un rôle fondamental en TAL, et la plupart des critiques qui reprochent au traitement automatique du langage un manque d'esprit linguistique (ou une perspective aveuglément extensionnelle) gomment ce point.

Il faut ensuite remarquer que l'écho sémantique que nous évoquions se différencie de la troisième investigation dictionnairique, envisagée par Jean Pruvost, en ce qui concerne le focus. Cette dernière est décrite de la façon suivante :

> La troisième approche est celle qui correspond à l'analyse des différents emplois du mot *norme* tout au long du dictionnaire : il s'agit d'établir un concordancier de l'usage du mot dans le corpus défini par tous les articles du dictionnaire où on trouvera le mot recherché. Ainsi apparaît l'usage dictionnairique du mot, au-delà de l'article qui lui est consacré, révélant par les co-textes de ce mot, c'est-à-dire ce qui le précède et ce qui le suit, une palette d'emplois, d'usages, propres à mieux en cerner la nature sémantique et syntaxique. Les agents de la norme que sont les dictionnairistes livrent ainsi à leur insu une illustration sémantique et syntaxique du mot qui complète heureusement l'article consacré à un mot. (Pruvost, 2005)

La troisième investigation considère le dictionnaire comme un corpus des environnements distributionnels du mot en examen : ce mot est le pivot du concordancier construit en explorant les définitions des autres mots. Dans l'approche de l'écho sémantique, en revanche, le pivot n'est pas constitué par la vedette examinée, mais bien par les mots de sa définition qui peuvent être communs avec celles des autres vedettes du dictionnaire.

L'articulation entre les deux niveaux lexicographiques susmentionnés et la perspective sur l'exploration dictionnairique implique que, concrètement, l'accès aux notions part d'un travail sur les mots eux-mêmes, et sur l'encadrement de ce travail.

Le résultat se présentera sous la forme d'un dictionnaire traditionnel doublé d'une représentation des relations sémantiques au sein d'un maillage partant des catégories épistémologiques et techniques pré-construites exposées plus haut, pour arriver à des champs sémantiques où la nature des relations envisagées sera représentée par des codes de couleurs. Ainsi, partant du domaine des sciences qu'est la « Médecine », l'utilisateur déroulera les champs constitutifs du domaine : Médecine, Anatomie, Botanique, Chirurgie.

Partant de « Botanique », un ensemble d'entités :

> Botanique → Objet de la science (phénomène observable, phénomène positif, phénomène négatif), instrument, corps constitué, partie du corps, action, état, événement, qualités (positives, négatives), caractères.

Enfin, de « phénomène observable » on retrouvera l'ensemble des champs sémantiques de chaque plante faisant apparaître des relations de synonymie entre les nomens et les noms français (*borrago*, *bourrache*) ou entre des unités lexicales dialectales (*borrache*, *bourache*). On peut ainsi imaginer établir des relations de méronymie ou d'hyperonymie entre des termes attestés dans le corpus, disparus du français moderne, représentant des *realia* non répertoriées aujourd'hui.

Parmi les premières conséquences d'une telle démarche, il y a la double possibilité de naviguer à travers une ressource et de la valider. En fait, plus les termes d'une ressource seront intégrés, plus il sera aisé d'en conclure que sa construction est valide. Une seconde conséquence renvoie à l'interopérabilité des ressources. Ainsi, les calculs de distance partant d'un lexique donné de l'ancien français, par exemple, pourront progressivement s'étendre à d'autres ressources de l'ancienne langue, voire à des ressources modernes qui intégreraient dès lors la similarité *gingembre* et *zédoaire*.

Protocole

Le dictionnaire

« L'investigation dictionnairique » désigne l'ensemble des moyens destinés à organiser la représentation du contenu d'un dictionnaire existant ou en cours de rédaction afin d'en faciliter l'accès à l'utilisateur final, ce dernier pouvant être (i) un lecteur humain ou (ii) une machine. Dans le premier cas, l'investigation dictionnairique consiste à construire une interface informatique de représentation herméneutique des contenus d'un dictionnaire guidant le lecteur du connu vers l'inconnu (Issac et Salvador, 2010), avec pour objectif de réduire la part d'inaccessible dans la culture médiévale pour un lecteur néophyte. Dans le second cas, l'investigation dictionnairique consiste à automatiser la création de ressources à partir d'un dictionnaire en cours de rédaction de manière à réduire la distance qui existe traditionnellement entre la ressource informatique, impropre à la consultation humaine, et le « dictionnaire papier », inexploitable. Les points importants liés au projet consistent à réfléchir à l'organisation herméneutique du projet lexicographique et de la représentation des données pour l'homme et la machine.

Il existe de nombreux dictionnaires en ligne (Caruso, 2011), 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 complexes. Le processus de constitution est totalement déconnecté du processus de consultation. Le principe – ou scénario – le plus fréquemment rencontré en termes d'interface est un calque – ou 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.

Édition collaborative

Un dictionnaire est le fruit du travail de plusieurs rédacteurs. Le recours à l'outil informatique, et plus particulièrement à sa dimension collaborative *via* le réseau, est dans ce cadre tout à fait pertinent. La notion même de travail collaboratif utilisant les nouvelles technologies fait l'objet de nombreuses expérimentations, notamment dans la réalisation de données dictionnairiques ou encyclopédiques. L'encyclopédie *Wikipédia* est à ce titre emblématique : elle fait l'objet de nombreuses études comme de controverses (voir par exemple Barbe, 2010 ou Endrizzi, 2008), alors même que son modèle éditorial est en perpétuelle évolution. L'édition collaborative dans le cadre de *Wikipédia* s'appuie sur un outil et des règles éditoriales. Celles-ci sont au nombre de cinq, et désignées comme « principes fondateurs » :

> Les principes fondateurs de *Wikipédia* fixent les grandes lignes qui définissent *Wikipédia* et les conditions de son élaboration. Ils constituent le fondement intangible de toutes les règles et recommandations du projet et sont au nombre de cinq : encyclopédisme, neutralité de point de vue, liberté du contenu, savoir-vivre communautaire et souplesse des règles[4].

L'outil utilisé est un *wiki*, qui répond au nom de *Mediawiki* et propose, outre un mode de saisie de l'article encyclopédique lui-même, un historique de ses versions et une page de discussion. Nous caractérisons les systèmes d'édition collaborative d'après trois critères : rédacteurs, structure et publication (RSP). Nous les expliciterons ici en les illustrant avec l'exemple de *Wikipédia*.

– Rédacteurs : Quelle est la politique en ce qui concerne l'identification des rédacteurs ? *Wikipédia* propose soit l'anonymat, soit une identification faible.
– Structure : Quelles sont les contraintes portant sur la production du rédacteur, et comment sont-elles exercées ? Le langage interne de *Mediawiki* propose un certain nombre de balises structurelles et de mise en forme, mais aucune contrainte

4. Extrait de : https://fr.wikipedia.org/wiki/Wikipédia:Principes_fondateurs [consulté le 29 décembre 2021].

n'est imposée. La normalisation est donc effectuée non pas par le système, mais par les mises à jour et les corrections successives des relecteurs.

– Publication : Quelle est la politique appliquée à la publication des productions des rédacteurs ? *Wikipédia* n'impose pas de relecture *a priori* ; toute production peut être directement publiée. Les éventuelles relectures et modifications sont effectuées *a posteriori*, aucun contrôle n'est fait sur le degré d'expertise, seul le consensus valide (ou invalide) un article.

Nos besoins en termes de diffusion et de travail collaboratif nous ont tout naturellement amenés à choisir une architecture client / serveur web. L'outil développé utilise donc les technologies standards XML / (X)HTML / CSS / JavaScript du côté du client, et un moteur de base de données du côté du serveur.

Les informations à manipuler étant d'une part de nature complexe, et d'autre part variables suivant la nature de l'information lexicographique répertoriée – les macrostructures des dictionnaires de spécialités doivent refléter la spécificité du domaine visé –, il est impossible de construire une structure capable de rendre compte de tous les cas de figure à l'aide d'un gestionnaire de base de données classique. À l'inverse, le langage XML permet une grande latitude dans la structuration de l'information ; c'est donc sur cette technologie que s'est porté notre choix. L'ensemble est rendu accessible par un serveur de base de données, BaseX[5], développé à l'université de Konstanz et structuré autour d'un moteur Xquery. Voilà un exemple de requête rendant accessible la nomenclature des vedettes du dictionnaire dont l'initiale est « A » :

```
for $x in distinct-values(//entry)[./form/orth contains text «^A»
using wildcards]
  order by $x collation «?lang=fr»
  return <a>{$x}</a>
```

On distingue principalement trois types d'utilisateurs : les lecteurs, qui n'ont aucun contrôle sur les contenus, les rédacteurs

5. En ligne : https://basex.org.

et les administrateurs. Les rédacteurs sont des lexicographes ou des spécialistes du domaine qui disposent d'une interface de saisie dont la structure est imposée. La saisie effectuée, le rédacteur a la possibilité de solliciter une validation de l'article auprès d'un administrateur. Ce dernier peut demander des modifications à l'auteur, ou publier l'article en l'état. Tant qu'une fiche n'est pas validée, sa version antérieure et sa nouvelle version cohabitent jusqu'à la validation finale, qui rejette l'ancienne version dans les limbes. Il existe ainsi trois états :

– état corrigé et validé,
– état en cours de rédaction,
– dans les limbes.

L'ensemble des fiches est au format XML, compatible TEI[6]. Les balises utilisées sont les suivantes :

– form : décrit la forme de l'entrée, *i.e.* la vedette (balise orth) ainsi que cela a été évoqué dans la section précédente,
– gramGrp : décrit les informations grammaticales (nature et genre) attachées à l'entrée,
– etym : donne la référence étymologique,
– sense : liste l'ensemble des sens possibles, chacun étant associé à un domaine (attribut n), une définition (balise def), un ensemble de citations (balise cit), un ensemble de liens (balise xr).

Voilà un exemple de fiche :

```
<entry>
  <form>
    <orth>ZEDOAIRE</orth>
  </form>
  <gramgrp>
    <gram type=”pos”>subst.</gram>
    <gram type=”gen”>masc.</gram>
  </gramgrp>
  <etym>
    <bibl><etymosrc>FEW XIX, 201b</etymosrc></bibl>
```

6. *Text Encoding Initiative*.

```
      <mentioned>Zadwar</mentioned>
    </etym>
    <sense n="BOT.">
      <def>Graine aromatique qui ressemble au gingembre mais qui
      est d'un goût moins âcre et de meilleure odeur.</def>
      <note id=»Struct»>
        <xr><ref>zédoaire,Curcuma,zedoaria</ref></xr>
        <xr><ref>Hyponyme</ref></xr>
        <xr><ref>Cohyponyme</ref></xr>
        <xr><gloss>Texte encyclopédique</gloss></xr>
      </note>
      <cit>
        <quote>Le foie confortent ces choses xilobalsamum, carvi,
        cubebes, zedoaire, alemandes, chastaingne en petit numbre
        [...]</quote>
        <bibl><author>WATERFORD, COPALE,</author>Le Secré
        des Secrés,
        <nonit>fin xii<sup>e</sup>s., p. 131, LIII</nonit></bibl>
      </cit>
    </sense>
  </entry>
```

Le corpus est constitué de l'ensemble des fiches dont la définition n'a pas systématiquement fait l'objet d'une validation. Le principe d'association s'effectue à partir d'un calcul de distance, l'idée étant de comparer chaque fiche avec la totalité du corpus. La nature même des liens qui seront calculés dépend bien évidemment de la manière dont les informations sont sélectionnées dans chacune des fiches. Ne pas effectuer cette sélection reviendrait à mettre au même niveau tous les éléments de la macrostructure : le résultat obtenu ne serait dès lors pas susceptible d'être interprété. Il s'agit donc non pas de concaténer, au prétexte que plus il y a d'information, plus le résultat serait pertinent, mais bien de choisir pour, le cas échéant, combiner.

Notre expérimentation s'est concentrée sur les définitions proposées par le dictionnaire CréaLScience dans sa version non aboutie. Nous avons donc extrait du corpus total un sous-corpus ne contenant que la vedette et les définitions qui lui sont rattachées :

<resultat>AIGUISER = Rendre aigu, intense</resultat>
<resultat>ALBUGINÉ = Humeur aqueuse de l'oeil</resultat>
<resultat>ANARCOSITÉ = Pouvoir narcotique</resultat>
<resultat>APLOMB = À l'aplomb, verticalement»</resultat>
<resultat>ADHÉRER = S'unir, se souder</resultat>
<resultat>ALPHOS = Ulcération de la peau</resultat>
<resultat>AIABLE = Propre à, apte à</resultat>
<resultat>ASCARIDE = Ver ascaride</resultat>
<resultat>ANNULEUX = Formé d'anneaux</resultat>
<resultat>ABONDANT = Abondant</resultat>
<resultat>APOSTOLICON = Onguent dit en lat. apostolicum</resultat>
<resultat>ADMINISTRATION = Action de faire absorber</resultat>
<resultat>ACCÈS = Accès d'une affection morbide, paroxysme d'une fièvre</resultat>
<resultat>ATORNER = Préparer</resultat>
<resultat>AGE = Âge, portion déterminée de la vie d'un homme Phase de la lune</resultat>
<resultat>ATEMPRANCE = Modération Équilibre de la complexion, du tempérament</resultat>
<resultat>AUDITIF = Qui sert à l'audition</resultat>
<resultat>ARGILLEUX = Argileux, de la nature de l'argile</resultat>

Similarités

Les définitions, on le voit, sont disparates, allant d'un simple mot redondant par rapport à l'entrée (comme pour ABONDANT), à un ensemble de définitions (comme pour ÂGE). Cet état de fait est attendu, le dictionnaire étant en cours d'élaboration / validation. L'un des résultats escomptés est justement la détection semi-automatique des incohérences ou des erreurs tant lexicographiques que définitionnelles.

Chaque définition est analysée (normalisation de la casse, découpage en mots[7]) et associée à sa vedette. Puis un étiqueteur morpho-syntaxique est appliqué de manière à obtenir pour chaque mot son lemme et sa catégorie grammaticale.

Plusieurs niveaux d'association ont été envisagés :

7. Nous utilisons ici le terme *mot* bien que les seuls critères retenus soient d'ordre typographique.

1. le lemme,
2. le mot,
3. l'unité polylexicale (mots simples et lemmes).

Les choix effectués déterminent là encore le sens donné, et par conséquent la grille d'analyse à utiliser. Ainsi la co-occurrence des termes de la recherche dans la définition, comme « plante » (subst.) et « à » (prép.), met en relation des vedettes de plantes individuelles (le géranium et la passiflore), alors que le même patron de recherche « plante à », en tant que lemme, met aussi en relation non plus un individu, mais des familles de plantes (plante(s) à feuilles caduques, plante(s) à feuilles lancéolées, plante(s) à feuilles persistantes). Il est ainsi possible d'envisager de combiner différents traits morphologiques ou sémantiques en imposant *a priori* certains motifs :

> [« à feuille »+ADJ+« utilisé dans »] où ADJ est un adjectif [ingrédient] ou [instrument]

Le résultat obtenu est un fichier XML représentant un graphe où les nœuds sont les vedettes (par exemple <li name=»LIEURE» num=»3»/>), et la nature du lien de similarité quand celui-ci est pertinent (par exemple <li name=»sans levain» num=»32» key=»key»>). Chaque nœud se voit associer un identifiant numérique qui est utilisé pour définir les liens. Ainsi <li deb=»1» fin=»21»/> relie l'entrée LIEURE avec FORSENERIE.

<li name=»LIEURE» num=»3»/>	<li deb=»1» fin=»21»/>
<li name=»LIBÉRALEMENT» num=»2»/>	<li deb=»1» fin=»9»/>
<li name=»CONTUSION» num=»3»/>	<li deb=»5» fin=»6»/>
<li name=»ALIS» num=»0»/>	<li deb=»0» fin=»5»/>
<li name=»INTÉGRAL» num=»7»/>	<li deb=»1» fin=»41»/>
<li name=»LEVAIN» num=»6»/>	<li deb=»1» fin=»10»/>
<li name=”CONTINU” num=”4”/>	<li deb=»1» fin=»20»/>
<li name=“SORBILE“ num=“8“/>	<li deb=»1» fin=»7»/>
<li name=“ANTHORA“ num=“9“/>	<li deb=»1» fin=»23»/>
<li name=”sans levain” num=”32” key=”key”>	<li deb=»1» fin=»34»/>
<li name=”levain” num=”5” key=”key”/>	<li deb=»1» fin=»30»/>
<li name=”INCESSIVEMENT” num=”11”/>	<li deb=»1» fin=»3»/>
<li name=”DESESPERÉ” num=”10”/>	<li deb=»1» fin=»14»/>
<li name=”INALTÉRÉ” num=”12”/>	<li deb=»1» fin=»25»/>
<li name=”INCORPOREL” num=”14”/>	<li deb=»1» fin=»36»/>
<li name=”STINCUS” num=”13”/>	<li deb=»1» fin=»35»/>
<li name=”ESTOILE” num=”15”/>	<li deb=»1» fin=»28»/>
<li name=”TRACE” num=”16”/>	<li deb=»1» fin=»19»/>
<li name=”TÉNESME” num=”18”/>	<li deb=»1» fin=»18»/>
<li name=”ESPERIT” num=”17”/>	<li deb=»1» fin=»16»/>
<li name=”SOURD” num=”20”/>	<li deb=»1» fin=»26»/>
<li name=“SOUFRE“ num=“19“/>	<li deb=»1» fin=»8»/>
<li name=“FORSENERIE“ num=“21“/>	<li deb=»0» fin=»1»/>
<li name=“AVEUGLE“ num=“23“/>	<li deb=»1» fin=»43»/>
<li name=“VIN“ num=“22“/>	<li deb=»1» fin=»22»/>
<li name=“ERRATIQUE“ num=“25“/>	<li deb=»1» fin=»2»/>
<li name=“DESESPOIR“ num=“24“/>	<li deb=»1» fin=»15»/>
<li name=“INSIPIDITÉ“ num=“27“/>	<li deb=»1» fin=»29»/>
<li name=“INSIPIDE“ num=“26“/>	<li deb=»1» fin=»13»/>
<li name=“MODÉRÉ“ num=“28“/>	<li deb=»1» fin=»4»/>
<li name=“SAUVAGE“ num=“29“/>	<li deb=»32» fin=»33»/>
<li name=“SEC“ num=“30“/>	<li deb=»1» fin=»12»/>
<li name=“VAIN“ num=“35“/>	<li deb=»1» fin=»39»/>
<li name=»MOULE» num=»34»/>	<li deb=»1» fin=»37»/>
<li name=»TINTINAILLE» num=»31»/>	<li deb=»1» fin=»42»/>
<li name=”PAIN” num=”33”/>	<li deb=»1» fin=»11»/>
<li name=”APPÉTIT” num=”36”/>	<li deb=»1» fin=»27»/>
<li name=”FAUCON” num=”39”/>	<li deb=»1» fin=»38»/>
<li name=”OPIATE” num=”38”/>	<li deb=»1» fin=»40»/>
<li name=”OLY” num=”37”/>	<li deb=»1» fin=»24»/>
<li name=”sans” num=”1” key=”key”/>	<li deb=»1» fin=»31»/>
<li name=»SAIN -2» num=»40»/>	<li deb=»1» fin=»17»/>
<li name=»ABISME» num=»42»/>	<li deb=»0» fin=»32»/>
<li name=»ÉTOILE» num=»41»/>	
<li name=»TENESME» num=»43»/>	

Ce résultat est traité par un script Javascript afin d’offrir un affichage et une navigation graphique dans n’importe quel navigateur (cf. *infra*).

Résultats, perspectives

Voici un premier exemple de graphe produit par le Dicoscope :

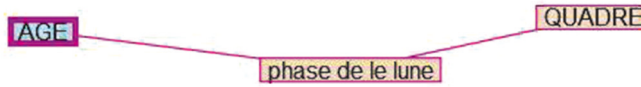


Fig. 1. Graphe ÂGE/QUADRE

Ce graphe révèle un lien entre ÂGE et QUADRE par le biais du segment « phase de la lune ». Ces deux mots connaissent chacun deux acceptions :

ÂGE
I. Âge, portion déterminée de la vie d'un homme
II. Phase de la lune

QUADRE
I. Aspect carré, ou quadrature (90.), jugé défavorable en astrologie
II. Phase de la lune, quartier

Le Dicoscope relie donc la deuxième acception de ÂGE avec la deuxième acception de QUADRE en relevant une intersection sémantique pointée par le lexicographe à travers le segment partagé analysé[8], la première étant un hyperonyme de la seconde. Le graphe détermine une relation de synonymie probable, la réciproque étant qu'il est possible d'exploiter le graphe pour effectuer une désambiguïsation entre les acceptions des lexèmes.

Dans un second temps, la figure suivante montre de manière détaillée les entrées liées à AZIMUT par le biais du segment lexicographique « de la sphère céleste ».

8. Par « analysé », nous évoquons le travail préalable de lemmatisation et de *tokenization* qui a permis la projection du Dicoscope. Nous rappelons que notre expérience est conduite sur le *Dictionnaire du français scientifique médiéval* (programme ANR CréaLScience) et que les segments sont lemmatisés.

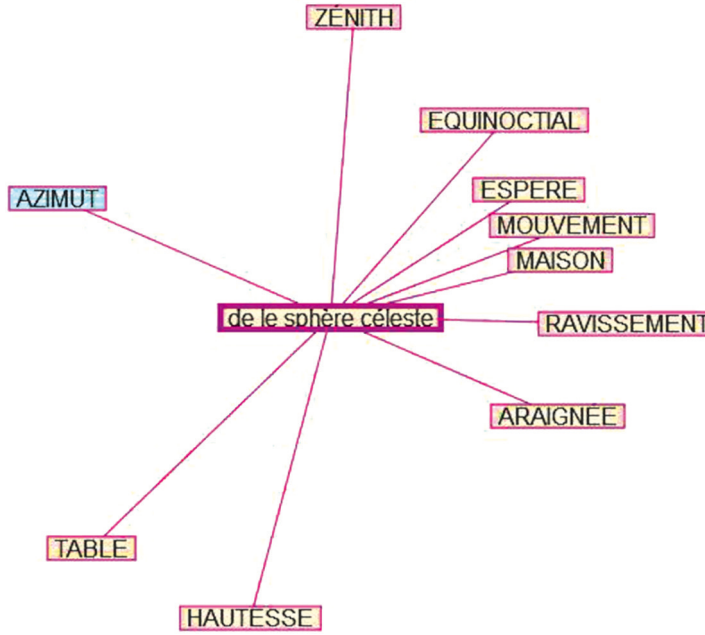


Fig. 2. Graphe «de la sphère céleste»

En l'occurrence, on obtient une liste de lemmes partagés par une série de fiches :

Céleste	(lié à SIGNAL, ACCIDENT, ECLIPSE…)
Cercle de le sphère céleste	(lié à AZIMUT MERIDIANE, ARIMILLE…)
Cercle de le	(lié à AZIMUT, ASCENSION)
Cercle	(lié à AZIMUT, HORIZON, CONE, CLIMAT…)
De le sphère céleste	(lié à AZIMUT ZENITH TABLE, ESPERE…)
De le sphère	(lié à AZIMUT, ARTICUS)
Le sphère	(lié à AZIMUT, COMETE, FIRMAMENT)
Passer par le zénith	(lié à AZIMUT, DIRECTION)
Passer par	(lié à AZIMUT, CENTRE, ARGUMENT…)
Passer	(lié à AZIMUT, SETON, ESPERIT…)
Sphère céleste	(lié à AZIMUT, ETOILE, INTELLIGENCE…)
Sphère	(lié à AZIMUT, POLE, APPROCHEMENT…)

À gauche, nous trouvons les segments partagés, et à droite les entrées qui les partagent, avec les informations concernant leur catégorie grammaticale et leur domaine. En construisant cette liste pour toutes les entrées du dictionnaire, nous obtenons l'ensemble des modules dictionnairiques. Or, si le dictionnaire peut être considéré comme un « trésor de la langue », le Dicoscope nous présente un trésor du dictionnaire, soit un méta-trésor. La perspective qui s'ouvre, ici, est celle d'une topographie du dictionnaire lui-même.

Il existe au moins deux façons de concevoir une topographie du dictionnaire. Tout d'abord, on peut étudier la distribution des segments selon les domaines, en explorant, par exemple, les moules caractéristiques d'un domaine, plutôt que d'un autre. Cela nous permettra de découvrir des regroupements ou des inclusions, fondés non pas sur une catégorisation *a priori*, mais bien sur les moules signifiants des définitions naturelles. Ainsi, nous pourrons évaluer, justement, la distribution des étiquettes de domaines accomplie par les lexicographes. D'ailleurs, dans le cas spécifique du *DFSM*, lorsque l'on dispose des noms des domaines anciens et modernes, on peut aussi envisager d'explorer leur évolution, en fournissant des données lexicographiques pour l'histoire des sciences. Ensuite, il sera intéressant d'explorer les segments partagés qui ne se révèlent attribuables à aucun domaine spécifique, mais qui manifestent un caractère trans-domaine. Ces segments-là permettront de faire émerger le socle du lexique indispensable pour construire tous les autres concepts. En ce qui concerne plus spécifiquement CréaLScience, nous attirons l'attention sur la possibilité d'investiguer les concepts qui sont restés constants à travers le lexique médiéval et ont perduré jusqu'à nos jours, c'est-à-dire le socle des concepts dont on ne peut pas se passer, au Moyen Âge comme aujourd'hui, pour définir tous les autres : des primitifs lexicographiques.

Observons enfin l'exemple suivant :

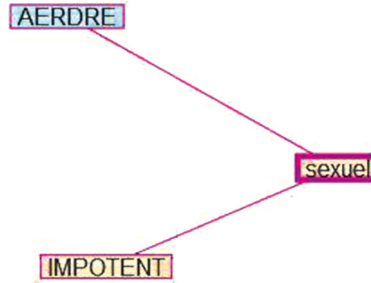


Fig. 3. Graphe « sexuel »

La définition de AERDRE proposée dans le *DFSM* est la suivante : « Être en contact lors de rapports sexuels ». Or, le Dicoscope montre ici une relation d'opposition, notamment avec « impotent », à travers le noyau commun partagé par les deux lexèmes. À la lumière de ce schéma qu'il est nécessaire de typer progressivement, probablement à la main, les liens mis en évidence par le Dicoscope permettent de réaliser un étiquetage des relations lexicales et conceptuelles du dictionnaire. Ainsi, nous aurons un outil efficace de validation du travail des équipes lexicographiques.

Finalement, l'outil mis en place peut aussi avoir tout simplement une fonction d'investigation dictionnairique qui obéisse précisément à la curiosité du lecteur, et qui le guide tout naturellement vers la connaissance de signifiants répertoriés difficilement accessibles. C'est le cas du lien entre le champ sémantique de HABITATION et « *avoir un rapport sexuel* », qu'un non spécialiste aura difficilement pu imaginer :

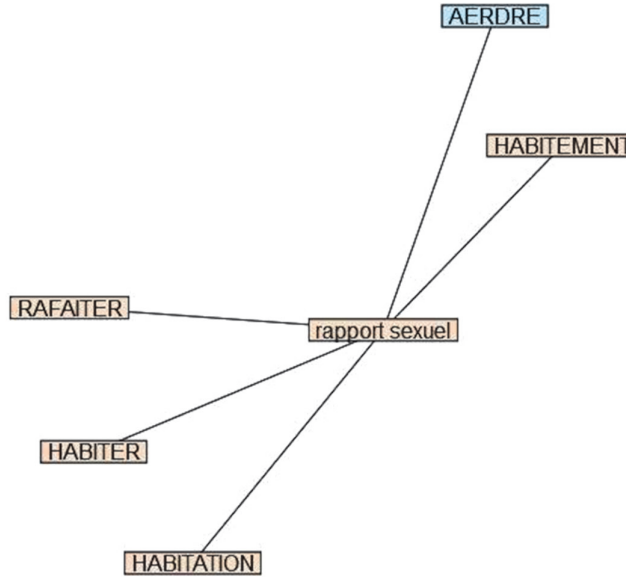


Fig. 4. Graphe « rapport sexuel »

Références bibliographiques

Aitchison, Jean, *Words in the Mind: An Introduction to the Mental Lexicon*, Oxford, Blackwell, 2003.

Austin, John Langshaw, « A plea for excuses » [1956], dans Urmson, James Opie et Warnock, Geoffrey James (dir.), *Philosophical papers*, London, Oxford University Press, 3e éd., 1979.

Authier-Revuz, Jacqueline, « Le guillemet, un signe de "langue écrite" à part entière », dans Defays, Jean-Marc, Rosier, Laurence et Tilkin, Françoise (dir.), *À qui appartient la ponctuation ?*, Louvain-la-Neuve, De Boeck/Duculot, 1998, p. 373-388.

—, *Ces mots qui ne vont pas de soi. Boucles réflexives et non-coïncidences du dire*, Paris, Larousse, 2 t., 1995.

Barbe, Lionel, « Wikipédia, un trouble-fête de l'édition scientifique », *Hermès*, n° 57, Paris, CNRS Éditions, 2010/1, p. 69-74.

Boudet, Jean-Patrice, *Entre science et nigromance. Astrologie, divination et magie dans l'Occident médiéval (XIIe-XVe siècle)*, Paris, Publications de la Sorbonne, 2006.

Caruso, Valeria, « Online Specialised Dictionaries: A Critical Survey », dans Kosem, Iztok et Kosem, Karmen (dir.), *Electronic Lexicography in the 21st Century: New Applications for New Users. Proceedings of eLesc 2011, Bled, Slowenien, 10-12 November 2011*, Ljubljana/Brighton, Trojína (Institute for Applied Slovene Studies)/Lexical Comuting Ltd., 2011.

Ducos, Joëlle, « Le lexique de Jean Corbechon : quelques remarques à propos des livres IV et XI », dans Van den Abeele, Baudoin et Meyer, Heinz (dir.), *Bartholomeus Anglicus, « De proprietatibus rerum », texte latin et réception vernaculaire*, Brepols, Turnhout, 2006, p. 101-115.

Endrizzi, Laure, « Le transfert des savoirs et le cas de Wikipédia », dans Schöpfel, Joachim (dir.), *La Publication scientifique. Analyses et perspectives*, Paris, Hermès/Lavoisier, 2008, p. 171-202.

Issac, Fabrice et Salvador, Xavier-Laurent, « Modèles théoriques inductifs et propositions d'applications aux données textuelles de l'ancien français », dans *JADT, Actes des 10 journées internationales d'analyse statistique des données textuelles*, Milano, LED, 2010.

Kocourek, Rostislav, *La Langue française de la technique et de la science. Vers une linguistique de la langue savante*, Wiesbaden, Oscar Brandstetter, 2e éd. augmentée, refondue et mise à jour avec une nouvelle bibliographie, 1991.

Lesk, Michael, « Automatic Sense Disambiguation. Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone », dans *SIGDOC Conference*, Toronto, Ontario, 1986.

Levelt, Willem J.W., *Speaking: From Intention to Articulation*, London, MIT Press, 1989.

Molinié, Georges, « Stylistique et tradition rhétorique », *Hermès*, n° 15, 1995, p. 119-128.

Pruvost, Jean, « Quelques concepts lexicographiques opératoires à promouvoir au seuil du XXIe siècle », *ELA. Études de linguistique appliquée*, n° 137, 2005/1, p. 7-37.

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

Estrella Pérez Rodríguez
Université de Valladolid

Une grande part de notre connaissance du passé nous a été léguée par l'intermédiaire des mots. C'est à travers des textes écrits à des époques particulières que nous pouvons connaître de nombreux faits qui sont survenus alors, et le Moyen Âge ne fait pas exception à cette règle de fait. Mais, pour que les textes constituent des témoins fiables de l'époque de leur rédaction, encore faut-il les comprendre de manière appropriée. Un instrument peut, sans nul doute, aider leur lecteur contemporain au regard de cette exigence : il lui faut disposer d'un dictionnaire de la langue dans laquelle ces textes sont rédigés.

En 1995, le professeur Maurilio Pérez (Université de León) a décidé d'entreprendre la dure et immense tâche d'élaborer un tel dictionnaire : le *Lexicon Latinitatis Medii Aevi Regni Legionis,* ou *LELMAL*[1], projet que je codirige actuellement. La première remise de ce travail a été publiée par la maison d'édition Brepols dans sa collection « Corpus Christianorum ». Il s'agit encore d'un *Lexicon Imperfectum*, car il comprend en l'état 3 020 entrées lexicographiques, ce qui représente sans doute le tiers, à peu près, de l'ensemble lexical établi par nos sources.

1. À la fin du présent article, le lecteur voudra bien trouver la liste de toutes les abréviations utilisées, ainsi qu'une bibliographie complète. Ce travail s'inscrit dans le cadre des projets FFI2015-64340-P (MINECO/FEDER) et VA027U14 (JCyL).

Cette contribution répond à un double objectif : en premier lieu, je souhaite exposer les caractéristiques fondamentales de ce dictionnaire. En second lieu, je développerai et commenterai quelques exemples intéressants qui démontrent l'importance de l'étude lexicographique afin de mieux connaître l'histoire de la langue d'un territoire donné.

Caractéristiques du dictionnaire

Sources et principes du travail lexicographique

Le *LELMAL* s'élabore actuellement à partir d'un corpus de sources fermé. Ce corpus est formé par les textes écrits principalement en langue latine sur le territoire du royaume des Asturies et de León, depuis la mer Cantabrique au Nord, jusqu'à Coria (Cáceres) au Sud. Ces textes s'inscrivent dans une période chronologique qui s'étend du VIII^e^ siècle, époque à laquelle les plus anciens d'entre eux ont été composés, jusqu'à 1230, année de l'union définitive des royaumes de León et de Castille. Pour des raisons linguistiques et également pratiques, la zone galicienne a été exclue du corpus. Celui-ci rassemble deux types de textes : d'un côté les chartes, qui ont un contenu juridique et sont très nombreuses (nous manipulons concrètement 47 cartulaires, qui contiennent un total de presque 10 000 actes) ; de l'autre côté les chroniques, qui sont au nombre de sept, et incluent le *Poème d'Almería*, qui fait partie intégrante de l'une d'entre elles.

Pour des raisons pratiques, nous travaillons sur des éditions publiées. Ce choix implique d'emblée un inconvénient réel, dont nous sommes et avons toujours été conscients, et qui tient à la différence de qualité de ces éditions. Aussi vérifions-nous à l'occasion par nous-mêmes quelques formes précises, en recourant à l'examen des actes originaux.

Depuis son origine, le travail a été organisé selon quatre principes fondateurs. En premier lieu, il requérait une équipe, cette équipe devant être aussi interdisciplinaire que possible, et associer des philologues latinistes, romanistes, germanistes, arabisants, et des historiens. Depuis le commencement, nous

avons également voulu que le travail se réalise avec l'aide inestimable des moyens informatiques, afin d'être assurés que nous disposions bien de toutes les occurrences de chaque terme dans les sources étudiées. S'agissant des textes historiographiques, la difficulté était résolue d'avance grâce aux concordances complètes des chroniques latines médiévales de l'Espagne déjà publiées (López Pereira *et al.*, 1993). En revanche, rien de semblable n'existait pour les chartes. En ce qui les concerne, le travail a commencé avec leur numérisation et la révision des textes ainsi obtenus afin d'éliminer les erreurs et les éléments étrangers au texte lui-même. Ainsi, autant de fichiers .txt que de cartulaires formant notre corpus ont été créés, et c'est avec eux que nous avons établi notre base de données. Nous travaillons sur cette base de données au moyen de deux programmes informatiques.

Le premier d'entre eux est le logiciel gratuit éditeur de textes ConTEXT, qui permet de réaliser des recherches dans toute la base de données, bien que l'on puisse seulement y rechercher des formes concrètes, par exemple le mot *arbusta* :

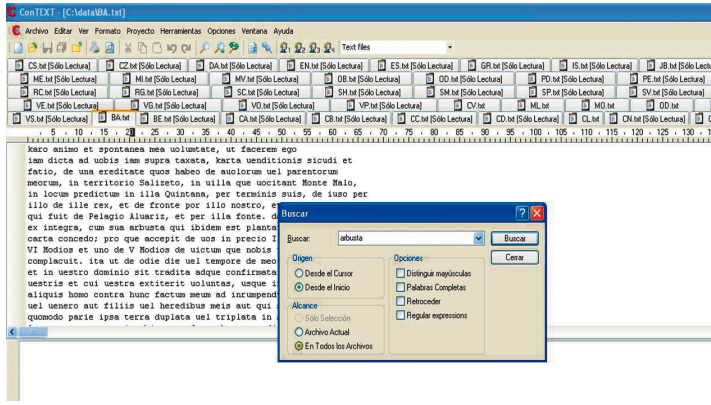


Fig. 1. Recherche du mot *arbusta* dans ConTEXT

Les résultats sont visibles dans un contexte restreint, et l'on peut aussi accéder au texte complet dans lequel l'une des occurrences observées apparaît :

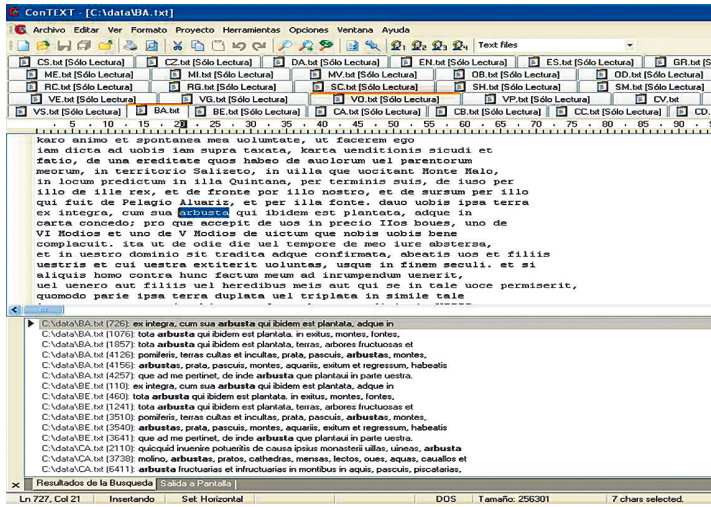


Fig. 2. Résultats de la recherche de *arbusta*

Le second programme utilisé est Microsoft Visual FoxPro, qui permet d'obtenir les concordances avec toutes les occurrences de chacune des formes incluses dans la totalité de la base de données. Pour les réaliser, la liste alphabétique de toutes les formes existantes dans la base de données est un prérequis nécessaire, car il nous faut indiquer, à partir de cette liste, le mot antérieur à la première forme que nous voulons rechercher, et la dernière forme qui nous intéresse. Ainsi pour obtenir les concordances du mot *cautio,-onis*, qui apparaît sous trois formes seulement dans nos sources (*cautione*, *cautionem* et *cautionis*), nous devons renseigner le mot qui précède la première de ces formes, c'est-à-dire *cauti*, et la dernière concernée (*cautionis*) :

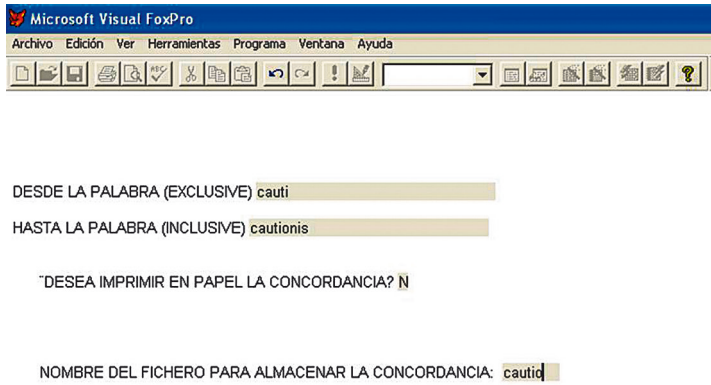


Fig. 3. Recherche des concordances du mot *cautio,-onis* avec Microsoft Visual FoxPro

Par la suite, le programme conçoit la concordance et la conserve dans un fichier auquel nous avons préalablement attribué un nom :

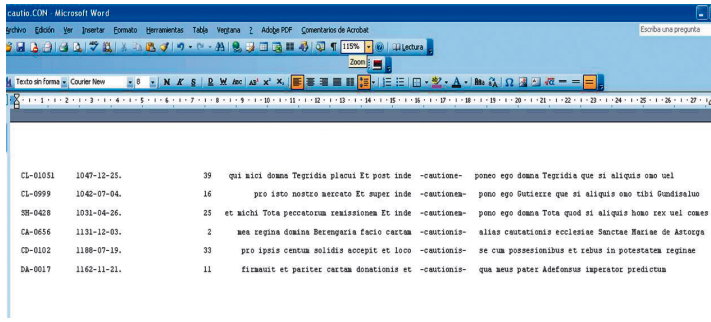


Fig. 4. Liste des concordances de *cautio* obtenues avec Microsoft Visual FoxPro

Dans ces tableaux de concordances, les formes recherchées sont précédées et suivies d'un contexte minimal. Devant ce contexte se trouve la référence de sa localisation : cartulaire, numéro et date du diplôme, numéro de la ligne où se trouve le vocable, etc. Dans l'exemple de concordance présenté ci-dessus, on peut constater, de plus, que deux lettres majuscules sont toujours attribuées à chaque cartulaire du corpus afin de l'identifier : ainsi *CL* correspond au cartulaire de la cathédrale

de León, *SH* à celui du monastère de Sahagun, *CA* à celui de la cathédrale d'Astorga[2], etc. Les chroniques sont quant à elles abrégées au moyen des trois ou quatre premières lettres de leurs titres (en minuscules): *Alb*. pour la *Crónica Albeldense*, *Adef*. pour la *Chronica Adefonsi imperatoris*, etc.[3], et on peut ainsi les distinguer des chartes dès le premier coup d'œil. Ces abréviations sont également utilisées dans le dictionnaire[4].

La numérisation et le « nettoyage » des textes a constitué une tâche longue et laborieuse, qui s'est achevée en 2001, motif pour lequel la base de données est composée par les chartes publiées jusqu'alors, qui, cela dit, constituent la majorité du corpus existant. Cependant, cela implique que les nouvelles éditions de quelques cartulaires comme ceux de Villaverde de Sandoval (Herrero Jímenez, 2003), d'Eslonza (Ruiz Asencio et Ruiz Albi, 2007) ou de Villanueva de Oscos (Álvarez Castrillón, 2011) ne soient pas utilisées, bien qu'elles soient plus correctes que les anciennes[5]. Une seule exception est à noter : le second volume des chartes d'Otero de las Dueñas, publié en 2005 (Fernández Flórez et Herrero, 2005), a été intégré par la suite.

Nous parvenons ainsi au troisième principe dirigeant le travail : nous avons décidé d'étudier et de réunir dans le *LELMAL* tous les termes qui apparaissent dans les sources sélectionnées, sans aucun genre de limitations (ni en raison de leur origine linguistique, ni à cause de leur ancienneté, ni pour aucun autre motif). Autrement dit nous n'excluons pas les termes du latin classique, ni les termes romans, ni les arabismes, ni les mots de toute autre origine.

Le quatrième principe recteur concerne la dimension lexicographique. Depuis le début de nos travaux, nous n'avons pas suivi le processus traditionnellement appliqué aux dictionnaires

2. Dans cette contribution, nous utiliserons également ces abréviations afin de citer les chartes. À la fin de l'article, le lecteur trouvera la liste des sources notariales mentionnées, avec l'indication des éditions correspondantes.
3. Les mêmes ont été utilisées dans López Pereira *et al.* (1993).
4. Une liste complète de toutes les sources du dictionnaire (et des abréviations qui leur correspondent) se trouve dans le *LELMAL*, p. xviii-xxii.
5. Néanmoins, nous les consultons lorsque cela est nécessaire.

de latin classique ou médiéval, qui consiste à effectuer les entrées lexicographiques par ordre alphabétique. Nous avons décidé, au contraire, de procéder par champs sémantiques ou lexicaux, un système qui présente divers avantages : d'un côté, chaque rédacteur peut se spécialiser dans un ou plusieurs de ces champs ; de l'autre, la comparaison des valeurs et des usages d'un terme avec ceux d'autres termes semblables est facilitée, ce qui garantit d'obtenir une précision et une profondeur plus grandes dans les résultats. Cependant, il faut reconnaître que le concept de champ sémantique a parfois été employé de manière assez souple, si bien que les rédacteurs ont fréquemment plutôt travaillé, en réalité, sur des ensembles de termes qui les intéressaient plus particulièrement pour une raison quelconque.

Caractéristiques externes

Afin que les utilisateurs puissent tirer du *LELMAL* le maximum de profit, il est nécessaire qu'ils connaissent bien ses caractéristiques externes, c'est-à-dire celles de sa structure, car il s'agit d'une œuvre complexe, dans laquelle de multiples éléments ont été pris en compte. Je vais ici résumer une à une les plus importantes de ses particularités externes[6].

Chaque article lexicographique peut être composé de six parties différentes :

1) *Lemme ou entrée lexicale* : Il apparaît toujours en caractères gras. Comme l'une des caractéristiques de la langue écrite au Moyen Âge est la grande variété graphique avec laquelle les mots sont reportés – et cela devient spécialement évident dans les textes notariaux – nous utilisons comme entrée la forme graphique la plus fréquente dans nos sources : par exemple *afinitas,-atis* ; *castanedo* ; *coniermana,-e*.

2) *Variantes formelles* : Aussitôt après l'entrée, les variantes graphiques du lemme sont indiquées par ordre alphabétique. Les variantes avec une seule occurrence sont signalées au moyen

6. Le *LELMAL en offre u*ne description plus détaillée (en espagnol, anglais et français), p. vii-lxxvii.

d'un point d'exclamation, qui les précède[7]. Après les variantes, on ajoute les dérivés du terme s'il y en a, et on précise si ce sont des noms propres (toponymes ou anthroponymes[8]).

3) *Étymologie* : La ligne immédiatement inférieure à celle de l'entrée et des variantes est dédiée à l'étymologie du terme, sauf si celui-ci dérive directement de l'indo-européen. L'étymologie s'exprime toujours de façon brève, hormis dans les cas où l'origine du mot est polémique, et où il faut faire place à l'exposé du débat et même à l'indication de la bibliographie appropriée.

4) *Définition* : C'est la partie centrale de l'article lexicographique. En premier lieu, il faut remarquer que notre intention était de ne pas multiplier inutilement les significations des termes, en nous contentant d'énoncer les usages basiques et en réservant les cas problématiques pour les exemples. Toutes les valeurs présentes dans nos sources se recueillent sans exception. Elles sont numérotées et ordonnées lorsque c'est possible, par ordre décroissant d'ancienneté. Celles qui existaient déjà dans l'Antiquité classique, c'est-à-dire dans la période qui s'étend jusqu'au IIIe siècle apr. J.-C., sont précédées d'un astérisque. Au moyen de deux barres verticales en caractères gras, on sépare les acceptions concrètes d'une signification plus générale[9].

Chacune des définitions est suivie de quelques exemples d'emploi du terme. Nous avons veillé à ce que les exemples soient les plus variés possibles, que leur contexte soit suffisant pour une compréhension parfaite, et que toutes les variantes graphiques du terme soient représentées avec eux. L'exemple le plus ancien et le plus récent de chacune des variantes est systématiquement donné. Les exemples se situent selon l'ordre suivant : d'abord ceux provenant des chroniques, par ordre alphabétique de leurs abréviations ; puis ceux provenant des actes, par ordre

7. Par ex., les variantes de *afinitas,-atis* sont : *!adfinitatibus*, *affinit-*. Celles de *castanedo* : *castacnedo*, *castagnedo*, *!castaneto*, *castaniedo*, *castanieto*, *castanned-*, *castanneto*, *!kastanedo*, *!kastanieto*. Celles de *coniermana,-e* : *coermana*, *!cogermanas*, *coiermana*, *congermana*, *cormana*, *quarmana*, *quermana*, *!quoiermanis*.
8. Par ex., le mot *auellano*. Var. : *!auelano*. Dériv. : *auelaned/t-*, *auellaned/t-*, *aulaneda*.
9. Par ex., *afinitas,-atis* a les significations suivantes : 1. *Proximité, voisinage. 2. *Affinité, parenté. ‖ Parent. 3. Limite. De celles-ci, uniquement 1 et 2 sont classiques.

chronologique. Chaque exemple est toujours précédé de sa référence exacte, indiquant la page et la ligne de l'édition où se trouve le vocable étudié, dans le cas des chroniques ; dans le cas des chartes, le cartulaire, le numéro du document et de la ligne concernée, et à la suite la date entre parenthèses. Dans tous les exemples, le vocable analysé apparaît écrit *in extenso* et en caractères gras[10].

Si le vocable a également dans les sources un usage toponymique ou anthroponymique, ce dernier emploi est indiqué après deux barres verticales et est aussi accompagné d'exemples[11].

5) *Noms des auteurs* : Le *LELMAL* constitue une œuvre d'équipe, mais n'est pas anonyme pour autant. Chaque article lexicographique est signé des initiales de son (ou de ses) rédacteur(s), situées à la fin de l'article, près de la marge de droite[12].

6) *Notes* : L'une des caractéristiques propres au *LELMAL*, car peu fréquente dans les dictionnaires, est la présence de notes. Les notes sont optionnelles, c'est le rédacteur de l'entrée qui décide de les ajouter ou non, ainsi que de l'étendue de leur texte. Les notes contiennent une information additionnelle : historique, linguistique, bibliographique, précisant la fréquence et le lieu

10. Voir l'exemple développé dans la note 12.
11. Dans le cas de *auellano* (voir note 8), indiqué comme ci-après : « ‖ *Sust. (casi siempre fem.) usado como topónimo con la adición del sufijo* -eto/a > -edo/a*: CL 1306.17 (c.1100)* Id sunt : … De **Auelanedo**, modio*; CN 4.11 (1120)* Do etiam uobis **Auellaneta** cum omni hereditate sua*; CO 143.33 (1122)* ecclesiam Sancti Vincentii et **Auelanetam***; SH 1461.5 (1191)* illam hereditatem quam habeo del Escouio de **Auellanedo***; RC 106.14* dedit ad Corias pro anima sua medietatem de illa uila de **Aulaneda** cum suis pertinenciis*; RC 151.30* cum medietate de Sancto Iohanne de Godan et **Aulaneda** integra ».
12. S'ils sont deux ou davantage, les initiales apparaissent par ordre alphabétique. Voici un exemple complet d'un terme tel qu'il se trouve dans le *LELMAL* :
« **audacter**. *Var.* : !audaciter.
Adv. en -(i)ter *formado sobre el adj.* audax,-cis.
1. **Con audacia, valientemente* : *Adef. 197.16* septimo uero die **audacter** uiri bellatores Christiani eruperunt de ciuitate per portas ad occasum solis*; Sil. 121.5* Idem uero qui eum tam **audaciter** percussit,…ab opidanis incolumis receptus est*; CL 442.8 (975)* **audacter** superuia inflati et malitia subducti atque diabolico furore accensi*; CL 1343.16 (1113)* precipio ut omnis quicumque sine uestra iussione intrauerit **audaciter** infra terminos… calumnia quam intus fecerit pariat. *MPG* ».

d'usage du terme, etc. Ces notes sont indiquées dans le texte de l'article lexicographique à l'aide d'une lettre minuscule qui figure en exposant, et sont développées à la fin de l'article. La minuscule peut être répétée si la même note concerne plus d'un élément de l'article.

Pour en finir avec la description des caractéristiques externes du dictionnaire, il faut évoquer ici une catégorie spéciale de termes. Il s'agit des mots inconnus et des « mots fantômes ». On désigne par cette expression des mots sans réalité linguistique, en général issus d'erreurs, qu'elles soient du fait du scribe médiéval ou bien du transcripteur moderne. Pour que le lecteur du dictionnaire puisse immédiatement percevoir qu'il se trouve devant l'un de ces termes, le lemme est présenté en petites majuscules, et l'article obéit à un ordre spécifique : tout de suite après l'entrée arrivent le (ou les) exemple(s) et, ensuite seulement, la définition possible du terme[13].

Il faut remarquer, de même, qu'à toutes les variantes graphiques des vocables correspond une entrée, classée dans le dictionnaire à la place qui lui revient selon le principe alphabétique, et au niveau de laquelle, au moyen d'un *videatur*, on renvoie à l'article lexicographique qui leur est attribué.

Nous souhaitons par ailleurs que le dictionnaire soit d'un usage pratique et d'un contenu précis, suffisant sans pour autant se révéler prolixe ; il importe que sa consultation soit aisée. Nous ne l'avons pas conçu comme le résultat final de nombreuses recherches, sinon comme leur matrice, le point de départ de possibles travaux institutionnels, linguistiques, historiques, littéraires, etc. Nous cherchons à ce que le *LELMAL* devienne indispensable aux historiens, utile aux philologues

13. Par ex., « ADEIGNAUIMUS : *CL 910.11 (1032)* ut faceremus ad uobis kartula uendiccionis… de terra nostra propria…, determinata ipsa terra :… quarta parte karrale qui discurre ad Legione, ubi termini determinauimus et coram testis **adeignauimus**.
Forma fantasma producto de un error, bien del escriba medieval bien del editor moderno, que afecta solo a la letra e, *que en realidad debería ser una* s, *es decir* adsignauimus, *verbo que aparece en ese mismo contexto en otros diplomas : por ej. CL 922.15 (1033)* in karale qui discure ad a Treualiolum, per ubi terminu delimitauimus et coram testis **adsignauimus**. *EPR* ».

latinistes et romanistes et précieux pour tous. Cependant, son but premier reste de faciliter la lecture et la compréhension des textes médiévaux rédigés en langue latine du Moyen Âge ou de la période romane.

Le travail de rédaction du *LELMAL* est aujourd'hui encore en cours, avec l'élaboration de nouveaux termes (jusqu'à présent presque 4 000) et la correction de la partie déjà publiée. Notre aspiration est de pouvoir achever le dictionnaire du royaume de León d'ici trois à quatre années, et de l'enrichir par la suite avec les textes castillans.

Exemples

Voyons maintenant les exemples. Nous en avons choisi trois relatifs au roman, plus un « mot fantôme ».

Ventrescas

Le premier terme retenu est *uentresca*, un substantif roman dérivé du latin *uenter* (« ventre »), particulièrement intéressant du fait qu'il est à peine attesté en castillan avant le XVIII^e^ siècle (Leiva, 1999). À l'époque médiévale, l'examen des sources de la zone castillane ne permet d'isoler que deux attestations seulement du vocable. On peut le lire sous la forme *ventresga* dans le *Fuero de Zurita de los Canes* (*ca*. 1218-1250) :

> *Et sabedera cosa es que el pellegero no a de retener alguna cosa delas* UENTRESGAS*, delas pelleias, ni de otras taiaduras*[14]

et, sous la forme *ventrisca*, dans les *Cortes de Jerez* de 1268 (Alfonso X[15]) :

> *la dosena delas gorias delas nutrias seys mrs.; la dosena delas* UENTRISCAS *delas nutrias dose mrs.*

Il semble qu'il soit également attesté au Moyen Âge dans la zone catalano-aragonaise : plus précisément, Miguel Gual

14. Ureña y Smenjaud (1911). Texte trouvé dans Real Academia Española (RAE), Banco de datos (CORDE) : *Corpus diacrónico del español*, en ligne : http://www.rae.es [consulté le 03 février 2015].
15. Avec la précédente citation, ce sont les deux seules occurrences médiévales de ce terme que nous avons trouvées dans Real Academia Española (RAE), en ligne, voir note 14.

Camarena[16] l'a repéré dans les ordonnances de courtiers de Barcelone remontant à 1271 et dans les taxes de « *reua* » de Perpignan datées de 1284[17], où sont évoquées des *ventresques de lú(d)ries* ; et, sous la forme *bentresqua*, il s'utilise en Aragon au XV^e^ siècle[18]. Dans tous ces usages, son signifié est « peau de la zone ventrale des animaux[19] », des loutres, en particulier, dans les exemples catalans. D'après le *DCECH* (*s.v.* « vientre »), la forme castillane provient du français[20]. Cependant, en français ancien, on trouve *ventresches/kes* dans un document provenant d'Orléans et daté de 1340, et dans plusieurs textes du XV^e^ siècle, avec le signifié de « ventrière (pièce du harnais) » ou de « filet du poisson (tranche du flanc)[21] ». Le dictionnaire de français ancien de Godefroy (*s.v.* « ventresche »[22]) le signale également avec le signifié de « peau du ventre » dans un texte du XIII^e^ siècle. En italien, il est présent fréquemment dès le XIV^e^ siècle[23]. Il semble donc que, dans toutes ces zones géographiques, y compris la région castillane, ce substantif soit documenté depuis le XIII^e^ siècle. Ce panorama, brièvement décrit ici, peut être complété grâce au travail réalisé dans le cadre du *LELMAL*. Le vocable se trouve en effet attesté dans les sources du *LELMAL* sous la forme *uentrescas*[24]. Nous disposons de deux occurrences dans autant

16. Gual Camarena (1968), *s.v.* « lúdria ».
17. Les textes se trouvent dans Gual Camarena (1968), p. 126-135 et 142-147, respectivement.
18. Sesma Muñoz et Líbano Zumalacárregui (1982), *s.v.* « bentresqua ».
19. Voir Corominas, *Diccionari etimològic i complementari de la llengua catalana*, *s.v.* « ventre ».
20. Leiva (1999, p. 156) propose, d'un côté, que le castillan aurait introduit ce vocable comme un emprunt du catalan ou de l'italien et, de l'autre, affirme que le castillan *ventrecha* dérive de l'influence du français *ventresche*.
21. D'après le *DMF* [consulté le 03 février 2015].
22. [consulté le 03 février 2015]. Nous n'avons trouvé le terme dans aucun des autres dictionnaires de français ancien inclus dans : http://www.micmap.org/dicfro/dictionaries.
23. D'après Leiva (1999, p. 149). Dans les bases de données qui incluent l'*Opera del vocabolario italiano* nous n'avons trouvé qu'une seule fois *ventresche*, dans un texte daté de 1234, le *Libro di Mattasalà di Spinello* (« *penello dele ventresche di madona Moschada* »). En ligne : http://www.ovi.cnr.it/II-Vocabolario.html [consulté le 13 février 2015]. Le terme n'est toutefois pas mentionné dans le *TLIO*.
24. Nous ne l'avons pas localisé dans les textes latins de la Galicie : *Corpus Documentale Latinum Gallaeciae* (*CODOLGA*), *versión* 10 (2013, en ligne : http://corpus.cirp.es/codolga [consulté le 13 février 2015]), ni dans ceux de la Catalogne : *Corpus Documentale Latinum Cataloniae* (*CODOLCAT*), *versión* 3 (2014, en ligne : http://gmlc.imf.csic.es/

de chartes du XIII^e^ siècle, originaires de deux monastères des actuelles provinces de León et de Zamora, respectivement :

> CD 172.6 (1202) *Verum tamen circa ordinis susceptionem haec erit dispensatio, quod neque cocullam nec uellum cogetur suscipere donec ipsa uellit, sed lineis et pannis albis ac pardis, et pellibus agninis et eis quae uulgo dicuntur* UENTRESCAS ;
>
> SM 123.33 (1221) *Et uxori uestre unum mantum et una garnacha de stanforte de quatuor in quatuor annos et mantum habeat pennam albam de* UENTRESCAS *et garnacham similiter.*

Le premier texte évoque une dispense accordée aux religieuses du monastère de Carracedo au moment de prendre le voile : on ne les oblige pas à se vêtir de la cape à capuche ni du voile tant qu'elles ne le désirent pas, elles peuvent choisir de porter des vêtements de lin ou d'étoffe blancs ou de couleur foncée et des peaux d'agneaux, ou encore celles que l'on appelle vulgairement « *uentrescas* ». Le second extrait est tiré d'un document testamentaire léguant à une femme un manteau et un autre vêtement, une garnache d'estanfort : tous deux sont taillés dans une peau blanche de « *uentrescas* ».

Le problème que posent ces deux textes tient au fait qu'ils sont uniquement conservés dans des copies datant du XVIII^e^ siècle. Il est impossible de savoir avec une totale certitude si le mot qui nous intéresse provient bien de l'original, mais c'est toutefois probable parce que, comme nous avons pu l'observer, il existe d'autres témoins du terme dans les textes de la péninsule Ibérique du même siècle. S'il en était ainsi, nous serions devant les attestations les plus anciennes provenant de ce territoire. Dans les textes de León, le vocable se rapporte à des vêtements féminins doublés avec des peaux d'animaux et a la même signification que dans les autres textes péninsulaires antérieurement cités : « peau de la zone ventrale d'un animal », laquelle, comme le soulignent certains de ces textes, a une plus grande valeur que la peau du cou[25]. Ici également, il s'utilise au

codolcat [consulté le 13 février 2015]), ni non plus dans le *Corpus de documentos españoles anteriores a 1700* (*CODEA*), *versión* 2011 (en ligne : http://demos.bitext.com/codea [consulté le 13 février 2015]).

25. Voir Gual Camarena (1968), *s.v.* « lúdria », p. 353.

pluriel. Le premier exemple cité nous apprend, de plus, qu'il s'agit d'un mot de la langue parlée, au moyen de l'expression *uulgo dicuntur*. Ce fait, ainsi que la date précoce de cette occurrence, antérieure encore à celle des témoignages catalans et français connus, semble indiquer une origine patrimoniale[26]. De plus, à en juger d'après les données dont nous disposons, l'usage du terme dans l'Espagne médiévale a été faible et, après le XV^e^ siècle, il a cessé d'être utilisé. Cependant, il se réintroduit au XVI^e^ siècle par l'influence italienne, avec une nouvelle signification[27].

Ieras

Le second exemple retenu est un mot relatif à la façon de mesurer les terres. Il s'agit du terme *ieras*, que nous avons uniquement trouvé dans une charte provenant du monastère cistercien de Moreruela (Zamora) :

> MO 39.20 (1195) *In alio loco damus uobis tres terras que fuerunt Pelagii Citiz. Et in alio loco quantum ad nos pertinet de Pelagio Velloso et unam terram que fuit uxoris Pelagii Saluadoriz. Et*

26. Sa dérivation de la forme diminutive du latin vulgaire **uentriscula*, telle que la proposait Meyer-Lübke (*REW* 9211), paraît impossible, mais peut-être pas celle de *uenter* avec le suffixe pan-roman *-esco*, lequel, bien que sans trop de profusion, est également présent en castillan médiéval, surtout dans des adjectifs : par ex. *grezisco, morisco*, les deux témoignés dans nos textes, mais aussi le substantif *parentesco*, dont nos sources présentent également un cas (CL 1391.4 [1129]). Au sujet de ce suffixe, voir Malkiel (1972).

27. Le terme, d'après Leiva (1999, p. 141-142), s'emploie dans deux œuvres castillanes du XVI^e^ siècle de forte influence italienne : *Retrato de la Loçana Andaluza* (1524) de F. Delicado et le livre de cuisine *Libro de Arte de Cocina* (1599) de D. Grando, avec le signifié de « ventre du porc » (voir aussi Leiva, 2001, p. 544-546). Et au XVIII^e^ siècle, Martín Sarmiento (*Lettre sur les thons*, 1757) désigne le terme comme italien ou portugais, et déjà référant au thon : « *Ya hemos llegado al Atun hecho. De el se preparan dos alimentos uno de imus venter latino Hypogastrion Griego, y Atun de hijada Castellano. A esto los Italianos llaman* ventresca*, y los Portugueses* ventrisca » (de Salas et García Solá, 1876, p. 157). La situation que nous percevons par rapport au terme *ventresca* dans la zone castillane concorde avec ce qu'observe Malkiel (1972) pour le suffixe *-esco* à l'échelle européenne : il s'emploie dans les formes médiévales du français, du provençal, du catalan, de l'espagnol et du portugais, mais à partir de la Renaissance et en raison de la grande influence italienne que subit tout l'Occident, il revient dans ces langues comme un italianisme. Une telle situation tend à confirmer, de même, le fait que le terme se trouvait dans les chartes léonaises originales, et qu'il ne s'agissait pas d'une introduction des copistes du XVIII^e^ siècle, époque à laquelle, comme l'indique le témoignage de Martín Sarmiento cité plus haut, ce substantif s'utilisait avec un autre signifié. Sous la forme *ventrecha*, ce vocable est indiqué dans le *DRAE*, pour la première fois dans son édition de 1817, avec le signifié de « ventre des poissons ».

in alio loco XV IERAS *et unam cupam. Hoc itaque concambium damus uobis pro hereditate quam a uobis accepimus.*

Le *LHP* (*s.v. «* iera ») affirme que *iera* dérive du terme du latin tardif *hera < area* et a la signification de « friche », inspirée de celle proposée par Du Cange pour *(h)era* (*s.v.*) : « le champ ou le lieu qui n'est ni cultivé ni labouré ». Cependant, une telle signification ne convient pas à notre texte, car *iera* y apparaît accompagné d'un adjectif numéral, alors que *area /era* ne s'utilise jamais avec un adjectif semblable dans les sources dont nous disposons[28]. Ce qui est affirmé dans le *DCECH* au sujet du substantif roman *jera* (*s.v.*) nous semble plus approprié. Il y est décrit comme un vocable léonais dérivé du collectif latin *diaria* (« labeur, salaire journalier ») avec le signifié de « surface de terre labourée en une journée par une paire de bœufs, labeur d'une journée », bien que ses auteurs ne le trouvent pas attesté avant 1627. Il s'agit d'un substantif employé jusqu'à nos jours aux Asturies, dans les régions de León, Salamanque, Estrémadure et Zamora[29], territoire qui appartient au royaume de León. Zamora est aussi la zone de provenance de la charte à laquelle nous nous reportons. En galicien, il existe encore *geira*[30], également avec le signifié de « mesure agraire : arpent de terre », un signifié similaire à celui que, croyons-nous, prend *ieras* dans son unique occurrence parmi les sources léonaises médiévales, que nous pourrions définir comme : « arpent de terre : mesure de terre équivalente à la surface labourée en une journée par une paire de bœufs[31] ». Notre texte indique que sont livrés « une terre qui a appartenu à la femme de Pelayo Salvadorez, et dans un autre endroit 15 arpents de terre, et un tonneau[32] ».

28. Voir *LELMAL*, *s.v. «* area,-e ».
29. Le Men, *s.v.* « jera », reprend ses usages dans toutes ces provinces, ainsi qu'en Galicie et au Portugal.
30. Pour le Portugal, voir aussi Santa Rosa de Viterbo, *s.v.* « geira », où il est défini comme « labeur ».
31. Le Men cite un exemple très similaire au nôtre, originaire d'Olivenza (Estrémadure) : « *He comprado 20* geras *de olivar* ».
32. Il faut comparer le texte avec cet autre, où un autre terme de mesure agraire est utilisé à la place de *ieras* : IS 198.30 (1214) *insuper* (sc. *concedo*) *quatuor aranzadas de uinea et unam cupam de XX modiis*.

Combo *et* recombo

L'adjectif *combo* peut aussi nous informer sur la langue parlée dans la zone étudiée. Il apparaît uniquement dans deux exemples tirés des sources asturiennes, concrètement du cartulaire de la cathédrale d'Oviedo :

> CO 23.46 (926) *per terminos suos : per Selia, per regum de Argandoe et per pando et per forkata de illa Ornia et per petra ficta et per petra longa et per forkata et per arbor* COMBO, *per serra de Ossilis ;* CO 88.14 (1084) *per suos terminos pernominatos : per flumen Seliam et per Telluiam, quę est integra mea, et per pinnam Forcatam et per arborem* COMBUM *et per busto Arnales et per Riumuode usque in Seliam.*

Aucun original de ces deux textes n'est conservé, nous ne disposons que de copies incluses dans le problématique *Liber Testamentorum* de l'évêque Pelayo[33], réalisé dans le premier tiers du XII^e^ siècle. Plus précisément, le premier d'entre eux est tiré d'un acte du roi Ramiro II en faveur de l'église d'Oviedo qui présente, paraît-il, de nombreux traits caractéristiques de la documentation de Ramiro III (961-985), ce qui laisse à penser que le copiste a utilisé comme modèle une donation dudit roi[34]. Une copie de ce document est également conservée dans la *Regla Colorada*[35], laquelle diffère seulement quant à certains éléments graphiques qui n'affectent pas notre vocable.

Le deuxième texte, en revanche, est uniquement conservé dans la copie du *Liber Testamentorum*. Entre les deux documents, il y a, vraisemblablement, un peu plus d'un siècle et demi d'écart, mais l'adjectif apparaît dans les deux cas dans un contexte identique : il qualifie le substantif *arbor* à l'intérieur de la démarcation des limites d'une propriété. Étant donnée la grande probabilité que la partie du texte où se trouve l'adjectif soit une interpolation ou un ajout du copiste du *Liber*, nous ne pouvons pas retrouver ce terme au-delà de la date de composition du *Liber*, c'est à dire dans les années 1120.

---

33. Valdés (2000), p. 522.27 et 616.6, respectivement.
34. Comme l'affirme Fernández Conde (1971), p. 185-189. Voir aussi Valdés (2000), p. 129-130.
35. Rodríguez Díaz (1995). Ce cartulaire a été réalisé à Oviedo en 1384.

En ce qui concerne l'origine du mot, il faut dire qu'il n'existait pas dans le latin de Rome. Cet adjectif appartient à la famille du substantif *comba* (« convexité, concavité »), terme d'origine celtique[36]. Ce nom est considéré dans le *DCECH* (*s.v.* « comba ») comme un vocable dialectal, léonais ou mozarabe, probablement apparenté avec le terme gallo-latin *cumba* > français *combe* (« vallée étroite et profonde, petite vallée encaissée[37] »), mot qui se trouve attesté à la fin du XII^e^ siècle en Gaule et qui a laissé une trace en occitan (*comba)* et en catalan (*coma)*. Avec la même signification, le substantif *cumba* s'utilise fréquemment dans les diplômes latins médiévaux de Catalogne : on l'y trouve attesté pour la première fois en 844[38], et aussi en Angleterre au début du XII^e^ siècle[39] ; en revanche, on ne trouve pas l'emploi adjectival. D'après les auteurs du *DCECH*, l'adjectif *combo* est beaucoup plus « incertain » car, selon leurs informations, il n'est pas signalé avant la fin du XVIII^e^ siècle, où il apparaît dans un dictionnaire, et n'est pas, non plus, confirmé par les dialectes. De ce fait, il leur semble que « si l'adjectif *combo* n'est pas un mot fantôme, déduit par les lexicographes de *comba* et de *combado*, il serait, tout au plus, un adjectif dérivé secondairement de ces mots ». Cette opinion ne peut plus être tenue face aux exemples tirés des sources asturiennes. Conformément à ceux-ci, nous pouvons affirmer que l'adjectif existe au moins depuis le commencement du XII^e^ siècle, et que ce doit être, en effet, un terme spécifique et exclusif de la langue parlée dans la zone asturienne[40], spécialement utilisé pour désigner les arbres au tronc courbé de façon très exagérée.

Le substantif *comba* est également attesté dans notre corpus, bien que, dans ce cas, uniquement sous la forme d'un nom propre, toponymique et patronymique :

36. De la racine celtique *kumbo- (« courbé ») ; en gallois *cwmm* (« vallée »). Voir à ce sujet la bibliographie citée dans le *GMLC*, *s.v. «* cumba », n. 2.
37. Définition du *Dictionnaire du moyen français* (1330-1500), 2012, en ligne : http://www.atilf.fr/dmf [consulté le 03 février 2015].
38. *GMLC*, *s.v. «* cumba ». Néanmoins, la première attestation de ce substantif dans le reste de l'Espagne est datée, suivant le *DCECH*, de 1573.
39. *DMLBS*, *s.v. «* cumba ».
40. Toujours en usage actuellement, puisqu'il figure dans le *DLLA*, *s.v. «* cumbu, -a, -o ».

> RC 97.15 *dedit ei iure hereditario illos montes et de Arganza et de Porcinero et de* CONUA ; RC 98.8,9 *De Sancto Mamete de* CONUA. *Sancto Mamete de* CONUA *fuit de Zalone ;* IS 222.23 (1225) *Guilielmus de la* CONBAS.

Le toponyme se trouve employé de manière similaire uniquement dans la documentation asturienne, ce qui confirme le caractère local de l'adjectif. Son dérivé *recombo*, présent dans un unique document, encore conservé exclusivement dans la copie du *Liber Testamentorum* (587.12), le confirme à nouveau :

> CO 117.62 (1100) *Ex alia parte per riuulum quem dicunt Bullera ad sursum et per quoto Pennino et per illa aquilera et per Penna de Rege et per illa uerruga et per illo trabe et per illa spelunca et per arbor* RECOMBO *et per illo scouio quem dicunt Pede de Mula.*

Comme on peut l'observer, il s'emploie dans le même contexte et avec un signifié identique au terme simple. Il semble clair que ces mots sont deux vocables de la langue parlée exclusivement dans la zone asturienne du royaume, qui s'utilisent ici dans un souci de précision propre aux exigences de la profession notariale[41].

Plentum

Un dernier exemple permet d'illustrer un autre type de difficultés procédant des chartes lorsqu'un travail lexicographique est réalisé à partir de leurs données. Il s'agit de la forme insolite *plentum*, que nous trouvons dans le texte suivant :

> CL 704.3 (c.1011) *Nodicia de hereditates et uineas qui sunt in Villa Citi Rege post parte Sancti Iacobi apostoli, fratris Domini. Id sunt: cortes et solares, hereditates et uineas qui fuerunt de domna Sinduara abbatissa siue ab* PLENTUM *et parentum quomodo et terras suas ganantias ab omni integritate.*

C'est une forme unique dans notre documentation et de surcroît inconnue en latin, aussi bien antique que médiéval, ce qui nous laisse à penser qu'elle relève d'une confusion du scribe médiéval. Il s'agirait d'un de ces mots que nous nommons « fantôme ». Il se trouve dans un contexte évoquant des possessions, dans lequel est également mentionnée l'origine de ces dernières,

41. Dans les sources du *LELMAL* on trouve quelques autres régionalismes, voir Pérez Rodríguez (2014).

avec le génitif *parentum* et avec *ganantias*. Dans les chartes, ces contextes sont habituels, et les expressions suivantes sont communément utilisées :

> CA 24.10-11 (923) *donis nostris quos auemus de auibus nostri set de parentibus nostris uel de nostras ganantias ;* CA 297.7-8 (1044) *uilla nostra propria que abemus de nostros auios et de nostros parentes et de nostras ganantias ;* CR 24.5-6 (1125) *nostra ereditate que abeo de parentorum meorum, sibi de ganantias quomodo de suas comparaturas ;* IS 3.20-21 (1043) *ereditates quos uiro meo… abuit de abios et parentibus et illas ganantias.*

Il ne semble pas cependant que *plentum* puisse être le résultat d'une déformation qui affecte *auios/auibus*. Au contraire, nous observons que ce terme rime avec le terme *parentum* qui le suit, ce qui se produit avec une certaine fréquence dans les formules documentaires, et l'altération découle peut-être précisément de cette recherche de l'assonance. Nous estimons, par conséquent, que c'est probablement le résultat d'une adultération de la forme *plenum* ou *plenius*, qui apparaît fréquemment dans tels contextes dans les diplômes asturiens léonais ; parfois, comme dans notre texte, dans l'expression adverbiale *ad plenum/ plenius* (« totalement, complètement, entièrement »), par exemple : SH 1612.20 (1218), « *uobis* AD PLENVM *restituat* » et – beaucoup plus proche de celle qui inclut *plentum* – CL 405.23 (967), « *terras ab omni integritate quos comparaui…, paludes, carrizales, terras* AT PLENIVS *habundanter…* » Le sens du texte serait approximativement : « Nouvelle des domaines et vignes de Villacid del Rey, propriété de l'apôtre Santiago, frère du Seigneur. Ce sont des cours et des terrains, des propriétés et des vignes qui ont appartenu à l'abbesse Mère Sinduara au complet, ou bien hérités de ses parents ou bien des terres gagnées par elle dans toute leur intégrité ». C'est l'hypothèse qui nous semble la plus plausible.

Nous pourrions continuer notre démonstration avec davantage d'exemples encore, mais ceux qui ont été exposés dans cette contribution, précisant et complétant considérablement (ou bien corrigeant) les affirmations avancées jusqu'à ce jour

au sujet de ces termes, peuvent, du moins nous l'espérons, se révéler suffisamment représentatifs de l'intérêt des apports rendus possibles par le projet LELMAL.

Références bibliographiques

Dictionnaires

Corominas, Joan, *Diccionari etimològic i complementari de la llengua catalana*, 9 vol., Barcelona, Curial Edicions Catalanes, 1983-1991.

DCECH = Corominas, Joan et Pascual, José Antonio, *Diccionario crítico etimológico castellano e hispánico*, 6 vol., Madrid, Gredos, 1980-1991.

DLLA = Diccionariu de la llingua asturiana, Uviéu, Academia de la Llingua Asturiana, 2000.

DMF = Dictionnaire du moyen français (1330-1500).
En ligne : http://www.atilf.fr/dmf, 2012.

DMLBS = Dictionary of Medieval Latin from British Sources, Oxford, Oxford University Press, 1975-2014.

DRAE = Diccionario de la lengua española.
En ligne : http://www.rae.es.

Du Cange, *Glossarium mediae et infimae latinitatis* [Niort, L. Favre, 1883-1888], 5 vol., Graz, Akademische Druck, 1954.
En ligne : http://ducange.enc.sorbonne.fr/

GMLC = Glossarium Mediae Latinitatis Cataloniae. Voces latinas y romances documentadas en fuentes catalanas del año 800 al 1100, Barcelona, Universidad de Barcelona, 1960-.

Godefroy, Frédéric, *Dictionnaire de l'ancienne langue française et de tous ses dialectes du IX^e^ au XV^e^ siècle*, Paris, Vieweg/Bouillon, 1880-1895.
En ligne : http://micmap.org/dicfro/search/dictionnaire-godefroy.

LELMAL = Pérez González, Maurilio et Pérez Rodríguez, Estrella (dir.), *Lexicon Latinitatis Medii Aevi Regni Legionis (s. VIII-1230) Imperfectum. Léxico latinorromance del reino de León (s. VIII-1230)*, Turnhout, Brepols, 2010.

Le Men, Janick, *Léxico del leonés actual*, 6 vol., León, Centro de Estudios e Investigación San Isidoro, 2002-2012.

LHP = Léxico Hispánico Primitivo (siglos VIII al XII). Versión primera del Glosario del primitivo léxico iberorrománico, projet inité et initialement dirigé par Ramón Menéndez Pidal, établi par Rafael Lapesa avec la collaboration de Constantino García, sous la direction de Manuel Seco, Madrid, 2003.

REW = Meyer-Lübke, Wilhelm, *Romanisches etymologisches Wörterbuch*, Heidelberg, Winter, 1972.

Santa Rosa De Viterbo, Joaquim, *Elucidário das palavras, termos e frases que em Portugal antigamente se usaram*, 2 vol., éd. Mário Fiúza, Porto/Lisboa, Livraria civilização, 1966.

TLIO = Tesoro della Lingua Italiana delle Origini.
En ligne : http://tlio.ovi.cnr.it/TLIO.

Sources citées

CA = Cavero, Gregoria et Martín, Encarnación (dir.), *Colección documental de la catedral de Astorga*, 2 vol., León, Centro de Estudios e Investigación San Isidoro, 1999-2000.

CD = Martínez, Martín (dir.), *Cartulario de Santa María de Carracedo*, t. I, *992-1500*, Ponferrada, Instituto de Estudios Bercianos, 1997.

CL = *Colección documental del archivo de la catedral de León (775-1230)*, vol. I, dir. Sáez, Emilio ; vol. II, dir. Sáez, Emilio et Sáez, Carlos ; vol. III, dir. Ruiz Asencio, José M. ; vol. IV, dir. Ruiz Asencio, José M. ; vol. V, dir. Fernández Catón, José Maria ; vol. VI, dir. Fernández Catón, José Maria ; León, Centro de Estudios e Investigación San Isidoro, 1987-1990.

CO = Larragueta, Santos García (dir.), *Colección de documentos de la catedral de Oviedo*, Oviedo, CSIC, 1962.

IS = Martín, Encarnación (dir.), *Patrimonio cultural de San Isidoro. Documentos de los siglos X-XIII*, León, Centro de Estudios e Investigación San Isidoro, 1995.

MO = Alfonso, Isabel (dir.), *La Colonización cisterciense en la meseta del Duero. El dominio de Moreruela (siglos XII-XIV)*, Zamora, Diputación de Zamora, 1986.

RC = García Leal, Alfonso (dir.), *Registro de Corias*, Oviedo, CSIC, 2000.

SH = *Colección diplomática del monasterio de Sahagún (Siglos IX y X)*, vol. I, dir. Mínguez, José Maria. ; vol. II et III, dir. Herrero, Marta ; vol. IV, dir. Fernández Flórez, José A. ; vol. V, dir. Fernández Flórez, José A. ; León, Centro de Estudios e Investigación San Isidoro, 1976-1994.

SM = González, Ángel Rodríguez (dir.), *El tumbo del monasterio de San Martín de Castañeda*, León, Centro de Estudios e Investigación San Isidoro, 1973.

Autres ouvrages

Álvarez Castrillón, José Antonio (dir.), *Colección diplomática del monasterio de Santa María de Villanueva de Oscos (1139-1300)*, Oviedo, Real Instituto de Estudios Asturianos, 2011.

De Salas, Javier et García Solá, Francisco, *Memoria sobre la industria y legislación de pesca: que comprende desde el año 1870 al 1874*, s.n., Madrid, 1876.

Fernández Conde, Francisco J., *El Libro de los Testamentos de la catedral de Oviedo*, Roma, Iglesia Nacional Española, 1971.

Fernández Flórez, José Antonio et Herrero, Marta (dir.), *Colección documental del monasterio de Santa María de Otero de las Dueñas*, t. II, *1109-1300*, León, Centro de Estudios e Investigación San Isidoro, 2005.

Gual Camarena, Miguel, *Vocabulario del comercio medieval. Colección de aranceles, aduaneros de la corona de Aragón (siglos XIII y XIV)*, Tarragona, Excelentísima Diputación Provincial, 1968.

Herrero Jímenez, Mauricio (dir.), *Colección documental del monasterio de Villaverde de Sandoval (1132-1500)*, León, Centro de Estudios e Investigación San Isidoro, 2003.

Leiva, Francisca, « *Ventresca* y sus parientes : *ventrisca, ventresco* y *ventrecha* », *Analecta Malacitana*, nº 22-1, 1999, p. 139-157.

—, *Vocabulario cordobés de la alimentación (s. XV y XVI)*, Córdoba, Universidad de Córdoba : Servicio de publicaciones, 2001.

López Pereira, José E., Díaz de Bustamante, José Manuel, Vázquez Buján, Enrique et Lage Cotos, María Elisa, *Corpus Historiographicum Latinum Hispanum Saeculi VIII-XII : Concordantiae*, Hildesheim, Olms/Weidmann, 1993.

Malkiel, Yakov, « The Pan-European Suffix *-esco*, *-esque* in Stratigraphic Projection », dans Valdman, Albert (dir.), *Papers in Linguistics and Phonetics to the Memory of Pierre Delattre*, Den Haag/Paris, Mouton, 1972, p. 357-387.

Pérez Rodríguez, Estrella, « Localismos en el reino asturleonés entre el s. VIII y 1230 : algunos ejemplos del léxico agrícola », dans Fernández, Ángel Martínez *et al.* (dir.), *Ágalma. Ofrenda desde la Filología Clásica a Manuel García Teijeiro*, Valladolid, Universidad de Valladolid, 2014, p. 425-431.

Rodríguez Díaz, Elena (dir.), *El libro de la Regla Colorada de la catedral de Oviedo. Estudio y edición*, Oviedo, Real Instituto de Estudios Asturianos, 1995.

Ruiz Asencio, José Manuel et Ruiz Albi, Irene (dir.), *Colección documental del monasterio de San Pedro de Eslonza*, León, Centro de Estudios e Investigación San Isidoro, 2007.

Sesma Muñoz, José Ángel et Libano, Ángeles, *Léxico del comercio medieval en Aragón (siglo XV)*, Zaragoza, Institución Fernando el Catolíco, 1982.

Ureña Y Smenjaud, Rafael de (dir.), *El fuero de Zorita de los Canes según el códice 217 de la Biblioteca nacional (siglo XIII al XIV) y sus relaciones con el fuero latino de Cuenca y el romanceado de Alcázar*, Madrid, Establecimiento Tipografico de Fortanet, 1911.

Valdés, José Antonio (dir.), *El* Liber Testamentorum Ovetensis. *Estudio filológico y edición*, Oviedo, Real Instituto de Estudios Asturianos, 2000.

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

Elisa Guadagnini*
Istituto Opera del vocabolario italiano
CNR, Firenze / KU Leuven

> Avec les techniques modernes de traitement mécanique de textes sont mis à notre disposition de grandes masses de matériaux bruts, et ceci en peu de temps et à peu de frais. Il revient à celui qui crée ces matériaux de les traiter selon les règles de l'art, afin de collaborer vraiment à l'analyse des sources historiques écrites et à l'augmentation du savoir.
>
> *Déclaration de Heidelberg*, 2001[1]

> *Compilare il presente Vocabolario parve la più alta, e vera maniera, fra tutte l'altre, di benificare questo idioma*.
>
> Préface « A' lettori », *Vocabolario della Crusca*, 1612

* Cette contribution transcrit de manière assez fidèle ma communication orale. S'agissant d'une présentation générale, et somme toute assez sommaire, du passé et des activités présentes de l'OVI, je n'ai ajouté ici que très peu de notes, jugeant inopportun d'alourdir le texte avec des références bibliographiques ponctuelles qui seraient pourtant nécessaires pour accompagner chacun des arguments que je ne fais qu'aborder de manière superficielle : j'ai cru pouvoir conserver à l'écrit le caractère vulgarisateur que présentait mon oral. Je tiens à souligner aussi le fait que je présente ma vision personnelle de l'OVI : le tableau qui en est brossé est donc certainement partiel et partial, et borné par les limites mêmes de mon expérience de travail. À mon expérience propre, par ailleurs, se rattachent les projets *DiVo* et *ReMediA*, que je cite ici parmi d'autres existants, et qui eussent assurément tout autant mérité d'être mentionnés.

1. *L'héritage culturel européen et la lexicologie du XXIe siècle : l'avenir de la lexicographie historique. L'exemple du* Dictionnaire étymologique de l'ancien français, 28-30 juin 2001, Internationales Wissenschaftsforum der Universität Heidelberg.

L'Opera del vocabolario italiano (OVI) est un institut de recherche dédié à la rédaction du vocabulaire historique de l'italien ancien, le *Tesoro della lingua italiana delle origini* (*TLIO*).

L'OVI, qui fait partie du CNR (Consiglio nazionale delle ricerche), est constitué en institut depuis 2001, après avoir fonctionné de 1985 jusqu'à cette date sous le statut de centre d'études. Son activité cependant est plus ancienne[2] : s'il a toujours été financé par le CNR, pendant ses vingt premières années l'activité et l'existence même de l'OVI ont été liées à l'Accademia della Crusca[3], dont l'institut partage aujourd'hui encore le siège (la villa médicéenne de Castello, à Florence). C'est en effet à l'Accademia della Crusca que l'on situe, en 1964, le début officiel des travaux pour la réalisation d'un vocabulaire historique décrivant, selon le projet de l'époque, le lexique italien des origines à nos jours[4].

Dès les années 1960, l'idée fut conçue de se consacrer d'abord à la rédaction d'un *Tesoro delle origini* : au fil des années, la décision fut prise de ne s'intéresser qu'aux origines, et de repousser les textes des époques moderne et contemporaine à une phase de travail ultérieure. Aujourd'hui encore, l'équipe de l'OVI travaille à cette première tranche historique du vocabulaire, tout en gardant pour le futur le projet de rédiger les articles relatifs aux séquences chronologiques suivantes.

De longues discussions ont amené les équipes à considérer comme « période des origines » la phase qui va des premiers

2. Pour une histoire de l'institut de sa fondation à la direction Beltrami (1992), voir Vaccaro (2013). Je renvoie à ce travail fondamental pour toutes les informations concernant l'histoire de l'institut et de ses projets, contenues dans les premières pages de l'ouvrage.
3. L'Accademia della Crusca est l'institution lexicographique italienne par excellence depuis 1612, année de la publication de la première édition du *Vocabolario della Crusca* : quatre rééditions du *Vocabolario* se sont succédées depuis, et la cinquième est restée incomplète, car sa rédaction a été interrompue en 1923, après plus d'un siècle de travail.
4. L'*OVI* fut dirigé depuis sa fondation en 1965 et jusqu'en 1972 par Aldo Duro, puis par Giovanni Nencioni (jusqu'en 1974), et ensuite par d'Arco Silvio Avalle. Quand l'OVI est devenu un centre d'études du CNR, en 1985, la direction a été assurée par Carlo Alberto Mastrelli, qui est demeuré en fonction jusqu'en 1992. Pietro G. Beltrami fut le directeur de l'OVI de 1992 à 2013 ; Paolo Squillacioti a ensuite assuré les fonctions de directeur par intérim jusqu'à l'arrivée, le 1er octobre 2014, de l'actuel directeur, Lino Leonardi.

documents écrits conservés (le plus ancien d'entre eux est l'*Indovinello veronese*, que l'on date du IXe siècle) à la fin du XIVe siècle : il s'agit là d'une périodisation classique pour l'histoire de la littérature et de la langue italienne, qui place au XVe siècle – avec l'Humanisme – le début de l'époque moderne.

Dès le début du projet, la décision fut prise de réunir dans la partie ancienne du vocabulaire l'ensemble des variétés italo-romanes dont sont restés des vestiges écrits, et non la seule variété toscane, ou plus spécifiquement florentine ancienne, qui est à la base de l'italien contemporain[5]. Il s'agissait dans les années 1960 d'un choix lexicographique révolutionnaire, mais à vrai dire le tableau diatopique restitué par les dictionnaires allait s'élargir sous peu – et indépendamment des travaux de l'OVI : bien avant la publication du premier article du *TLIO*, il faut rappeler que le *Grande dizionario dell'italiano*, fondé par Salvatore Battaglia et complété sous la direction de Giorgio Barberi Squarotti, cite parmi les textes médiévaux des œuvres provenant de l'Italie du Nord et de l'Italie médiane, offrant un aperçu géographique plus vaste que celui retenu par les grands dictionnaires italiens qui l'avaient précédé, c'est-à-dire les cinq éditions du *Vocabolario della Crusca* et le *Dizionario della lingua italiana* de Niccolò Tommaseo et Bernardo Bellini[6]. Mais le premier essai d'une lexicographie de l'ancien italien qui ne fût pas le rassemblement des glossaires des éditions critiques – une option de description du lexique des « origines », qui, d'ailleurs, fut prise en compte, à un moment donné, à l'*OVI* aussi[7] – fut l'expérience du *Glossario degli antichi volgari italiani* (*GAVI*), que Giorgio Colussi a rédigé à partir des années 1980 et jusqu'à sa mort[8] : pour la première fois, l'on tentait une description lexicographique qui utilisait

5. On peut lire une excellente réflexion sur le concept d' « italien ancien » selon les critères chronologiques et diatopiques dans Tomasin (2013).
6. Le *Vocabolario della Crusca*, tout comme le Tommaseo-Bellini, cite en majorité absolue des textes toscans (à l'exception près des poèmes de Iacopone de Todi).
7. Le projet d'un *Glossario dei glossari* (« Glossaire des glossaires ») fut entrepris dans les années 1970 et poursuivi pendant quelques années, avant d'être abandonné en faveur d'une autre idée de dictionnaire : *voir* Vaccaro (2013), p. 363 et *passim*.
8. Le premier volume du *GAVI* date de 1983, le dernier (XX/2) de 2006 : les volumes publiés couvrent les lettres A, B, C, D, S, U, V et Z.

directement les textes médiévaux (en principe jusqu'à 1321, année de la mort de Dante, mais sont cités aussi des textes ultérieurs offrant même quelques ouvertures sur l'époque moderne), visant à restituer pour le vocabulaire des « origines » un tableau lexicologique, et non purement glossographique, en s'appuyant sur l'entière documentation disponible.

Le *Tesoro della lingua italiana delle origini* (*TLIO*), le vocabulaire que l'OVI est en train de rédiger, couvre donc l'ensemble des variétés italo-romanes dans lesquelles sont conservés des textes écrits avant l'an 1400.

Dès les années 1960, le choix fut fait de ne pas s'appuyer sur la lexicographie précédente mais de rédiger un vocabulaire de première main utilisant comme source directe les textes, en entendant par là les éditions modernes existantes, et non les manuscrits ni les *editiones principes*. La décision ayant été prise d'accomplir un dépouillement exhaustif des sources médiévales publiées, un premier problème apparut, avec le recensement des textes et des éditions. À la suite notamment d'un voyage à Nancy, où s'élaborait un autre « Trésor », celui « de la langue française », le directeur du projet, Aldo Duro, créa l'*Ufficio filologico* (le « Bureau philologique »), auquel fut confiée la tâche de mener à bien l'élaboration de la table des textes à citer ainsi que le repérage et l'évaluation des éditions existantes. Quand, le premier janvier 1965, Domenico De Robertis devint le directeur du bureau philologique nouvellement né, il entreprit avec ses collaborateurs et ses élèves un immense travail de recherche et d'étude, mais aussi de contrôle et de révision des textes : l'équipe du bureau philologique ne se borna pas, en effet, à recenser les éditions et à choisir les meilleures d'entre elles, mais travailla beaucoup à l'amélioration des textes critiques. Pour ne donner qu'un exemple des activités du bureau, pour tous les documents à tradition unique l'on procéda à la collation intégrale de l'édition avec le manuscrit ; il en fut de même pour toutes les éditions que l'on déclarait fondées sur un manuscrit déterminé.

Quand d'Arco Silvio Avalle, en 1974, prit la direction de l'*OVI*, il vit dans l'activité du bureau philologique un risque

réel de paralysie du vocabulaire : dix ans plus tard, en 1983, le CNR stigmatisa – avec les mots de Scevola Mariotti – le « philologisme » de l'*Opera del vocabolario*, qui paraissait nocif pour la réalisation du vocabulaire[9]. Ce contraste entre la position du CNR, qui voulait une assurance quant aux délais et exigeait que les progrès du projet soient visibles, et celle de l'Accademia della Crusca, qui défendait sa méthode philologique, perdura plusieurs années et créa de multiples tensions. Pendant les trois premières décennies de son activité, l'équipe de l'OVI ne rédigea pas un seul article du vocabulaire, mais réalisa un énorme travail de préparation des textes qui allaient constituer la base de données pour la rédaction : la documentation relative à ce travail est aujourd'hui librement consultable en ligne, grâce à un projet mené à terme par Pär Larson et Zeno Verlato[10].

Entre-temps, par ailleurs, l'informatique a vite évolué. Dès 1965, à l'OVI, l'équipe oeuvra à la constitution d'un archivage des sources dépouillées de façon électronique et lemmatisées : cet archivage devait être – dans la vision portée par Aldo Duro – un outil autonome, disponible à la consultation pour l'ensemble des chercheurs. À la suite des échanges associant Duro et le père Roberto Busa, qui, avec le support technique d'IBM, était en train de publier à Gallarate l'*Index Thomisticus*, pendant plusieurs années l'OVI confia l'informatisation de ses sources à IBM : à l'époque le support utilisé était la carte perforée. Ensuite, avec l'essor de logiciels capables de gérer des textes longs et non seulement une courte chaîne de caractères, le projet d'une archive de contextes évolua naturellement vers le projet d'une archive digitale de textes entiers : en 1988, l'OVI, depuis trois ans centre d'études du CNR, adopta un logiciel – écrit par Eugenio Picchi, de l'Istituto di linguistica computazionale (CNR, Pise) – qui répondit ensuite à l'appellation « DBT » et se trouva assez répandu en

9. Se référer à Vaccaro (2013), p. 371.
10. Le projet *LIVS – Lingua italiana e Vocabolario storico : Metodi antichi e moderni*, financé par la région Toscane, a existé du 9 février 2011 au 8 septembre 2013. Les archives des matériaux inventoriés sont disponibles en ligne : http://tlio.ovi.cnr.it/LIVS/livsdoc/.

Italie dans les années 1990[11]. Les données informatisées pendant les années précédentes furent récupérées et converties dans le format DBT : le corpus ainsi constitué par récupération de matériaux plus anciens et par ajout de nouvelles acquisitions atteignit une dimension convenable pour la rédaction en 1995[12]. Ce premier corpus était partiellement lemmatisé, car DBT permettait une lemmatisation manuelle, réalisée texte par texte. En 1998, une version simplifiée du corpus (sans la lemmatisation) fut publiée sur le réseau grâce au consortium Italnet[13]. La même année devint opérationnel le logiciel GATTO, dédié, conçu et écrit spécifiquement pour la base de données de l'OVI par Domenico Iorio-Fili, un informaticien travaillant à l'époque à l'institut. Parmi les avantages que présentait l'adoption de ce logiciel, il y avait le fait que l'on pouvait dès lors procéder à une lemmatisation qui s'effectuait sur le corpus entier[14], ce qui permettait un contrôle optimal des matériaux qui jusqu'à cette date ne pouvaient être analysés que contexte par contexte (à l'époque des cartes perforées) ou lemmatisés forme par forme à l'intérieur d'un seul texte (DBT). En 2005 fut publié en ligne le corpus lemmatisé, qui offre aux usagers des possibilités de recherche très complexes grâce au logiciel GattoWeb, écrit lui aussi par Domenico Iorio-Fili.

Aujourd'hui, l'OVI teste Gatto4, un nouveau logiciel écrit par Domenico Iorio-Fili juste avant sa retraite, survenue en août 2014.

---

11. Il s'agit notamment du logiciel de la *LIZ – Letteratura Italiana Zanichelli*, de Pasquale Stoppelli et Eugenio Picchi, dont la première édition date de 1993.
12. La récupération des matériaux codifiés dans des formats précédant le format DBT a été réalisée par Rosalba Cigliana et Valentina Pollidori. La première version du *corpus* est due au travail des informaticiens Eugenio Picchi et Elisabetta Marinai et de la chercheuse Valentina Pollidori, qui a géré le *corpus OVI* jusqu'à sa mort, en 2004. Pour la première phase des travaux qui ont conduit à la réalisation du *corpus*, voir Avalle (1979), De Robertis (1985) et Duro (1985).
13. La base de données consultable sur le site *Italnet* a été constituée par Theodore J. Cachey, Mark Olsen et Christian Dupont (en ligne : http://artfl-project.uchicago.edu/content/ovi). Cette base de données a eu le grand mérite de faire connaître le *corpus OVI* à la communauté scientifique – elle a été en effet beaucoup utilisée par les chercheurs –, mais est aujourd'hui obsolète, sa dernière mise à jour datant de 2005.
14. La lemmatisation du *corpus OVI* a été effectuée jusqu'en 2006 par Roberta Cella et ensuite par Elena Artale. Pour la méthode de lemmatisation, *se référer à* Esperti (1979). Elena Artale est en train d'accomplir la révision et la mise à jour des normes de lemmatisation : un premier résultat de ce travail, daté de 2013 et non publié, est réservé à l'usage interne.

Ce logiciel, outre qu'il présente de nouvelles fonctions de gestion des textes, constitue la tentative de répondre au piège de la « complexité » informatique contre lequel met en garde Robert Martin[15] : alors que, jusqu'à maintenant, le corpus OVI utilisait une évolution du « vieux » codage DBT, Gatto4 adopte le marquage international XML (format TEI), ce qui va rendre possible – dans un futur proche – l'interopérabilité des données du corpus au moyen d'un langage informatique largement partagé.

Quant au vocabulaire, une esquisse d'article lexicographique, qui anticipe le modèle adopté par le *TLIO*, fut élaborée en 1989[16] : l'on choisit – notamment et définitivement – d'utiliser de façon exclusive les ressources et le support informatiques, que ce soit pour la rédaction ou pour la publication du vocabulaire.

L'élaboration d'un modèle d'article pour le *TLIO*, tout comme la conception – ou peut-être la *vision* – de ce que devait être cette œuvre lexicographique (et de ce qu'elle est effectivement), après maintes années de théories et de réflexions préliminaires, date cependant des années 1990 et toutes deux sont dues à Pietro Beltrami, qui prit la direction de l'OVI en 1992 et la conserva jusqu'à 2013. Le premier article du *TLIO*, rédigé en 1996 par Beltrami lui-même, fut celui correspondant au lemme *abaco* ; en novembre 1996 fut publié en ligne un premier ensemble regroupant 122 articles, qui atteignit le seuil des 1 000 articles à la fin de 1998 : depuis lors, la rédaction progresse en moyenne de 2 000 articles par an.

Fidèle à sa conception originaire, le *TLIO* se compose d'articles rédigés à partir du dépouillement direct du corpus, qui est la

15. Je me réfère à la communication que Robert Martin a donnée à l'occasion de ce congrès : « Les réussites (et les pièges) de la lexicographie électronique ». Dans cette communication, il a souligné l'extrême complexité du langage informatique, qui rend souvent difficile le passage des données à de nouvelles versions du *software*, tout comme le passage de consignes entre l'informaticien qui a géré le logiciel jusqu'alors et l'informaticien qui va devoir en prendre la relève ; voir *supra*, p. 19-20.

16. Il s'agit de *COVIRED* : *se référer à* Ceccoli, Lorenzi et Pollidori (1989), *COVIRED : una procedura automatica come ausilio per la redazione del Tesoro della lingua italiana delle origini*, Firenze, Centro di Studi CNR Opera del vocabolario italiano – texte à usage interne, non publié mais refondu en partie dans Ceccoli, Lorenzi et Pollidori (1991), p. 67-84.

source principale (et en principe unique) de la documentation : le rédacteur, qui travaille seul[17], rédige chaque article après avoir lu et interprété toutes les occurrences du lexème présentes dans le corpus, qu'il peut repérer facilement puisque la lemmatisation fournit la liste des formes graphiques que revêt le lexème dans la base de données. Ce parti pris de rédaction implique une analyse *a posteriori* du matériel : en théorie, le rédacteur travaille en mettant de côté sa compétence de locuteur italien et en extrapolant les acceptions du lexème à partir de l'interprétation de chaque contexte[18]. Il organise ensuite ces acceptions dans une structure purement sémantique – les aspects grammaticaux, notamment, ne constituent pas un principe de structuration du matériel : s'il y a coïncidence sémantique, il est possible de réunir sous une même définition « cumulative », par exemple, l'adjectif et le substantif ou l'adverbe, ou – pour les verbes – l'emploi transitif et celui intransitif ou impersonnel.

Chaque article présente, en tête, des informations générales sur le lexème : la liste de ses formes graphiques présentes dans le corpus, l'étymon, la première attestation absolue et la

17. En 2005 a été instituée et confiée à Rossella Mosti la « prérédaction », une sorte d'étape intermédiaire entre la collecte des données résultant du *corpus* et la rédaction proprement dite : *voir* Mosti (2012). L'un des principaux résultats de cette activité a été la compilation de la liste des articles prévus pour le *TLIO*, qui a été complétée en 2014 (*cf. infra*).

18. Le fait que la rédaction soit effectuée *a posteriori* à partir du *corpus* ne signifie cependant pas qu'elle adopte une procédure inductive inspirée des méthodes de la linguistique statistique. Comme l'a si bien exprimé Diego Dotto : « *Qualsiasi proposta di descrizione di una varietà linguistica antica, una grammatica o un vocabolario, presuppone l'esistenza di un corpus. […] Di volta in volta la discriminazione tra ciò che non si trova [nel corpus] per una lacuna casuale e ciò che non si trova per una regola della lingua è un'operazione di estrema delicatezza. L'evidenza del principio si scontra infatti con la competenza inevitabilmente parziale di chi si propone di ordinare, descrivere o addirittura spiegare una costruzione o un lessema di una varietà linguistica antica, per la quale manca la possibilità di elicitare le regole grazie all'introspezione o all'interrogazione diretta di un parlante nativo. Il margine d'errore è sensibilmente più ampio. Un atteggiamento prudente è quindi giustificabile, ma la cautela non deve sfociare nella rinuncia programmatica a formulare ipotesi ; è vero invece che l'ipotesi deve essere evidente in sé e soprattutto sottoponibile in qualsiasi momento alla verifica su (nuovi) dati empirici o su (nuove) ipotesi che la potranno confermare o falsificare. Da altro punto di vista, importa rimarcare che la limitazione all'attestato senza alcuna distinzione tra frequenza e regola rimane essa stessa un'ipotesi al pari della ricostruzione, da valutare, di nuovo, caso per caso, alla luce della sua maggiore o minore plausibilità* » (Dotto 2012, p. 344-345).

« distribution géographique » du lexème, qui elle-même indique la première attestation pour chaque variété diatopique[19].

Le *TLIO* vise à décrire un état de langue, et non à constituer un glossaire des sources. Destiné à reconstruire l'architecture du lexique italien ancien, il se tient depuis toujours à la règle imposant de citer de préférence des textes documentaires plutôt que des œuvres littéraires : cela est dû à la conviction que ce genre de texte restitue un usage des mots autant que possible dénotatif et spontané, à l'opposé de l'usage littéraire. Le *TLIO* donne en outre la priorité à la citation d'exemples ayant valeur de glose, qui sont d'ailleurs marqués par un sigle spécifique dans le vocabulaire (Gl : « glose »).

Qui plus est, toujours parce qu'il vise à décrire un état de langue, le *TLIO* s'impose de « regarder au-delà de la donnée matérielle[20] », qui n'est qu'un reflet de ce que l'on cherche à voir et à décrire. Tout en ayant muni chaque texte d'une fiche bibliographique qui fournit les informations philologiques essentielles, et en attribuant une place privilégiée au témoignage des textes documentaires, pour lesquels l'œuvre coïncide souvent avec le témoin, le *TLIO* se fonde sur trois présomptions philologiques fondamentales : 1- que l'édition restitue l'œuvre ; 2- que la datation correspond à celle (connue ou présumée) de la composition de l'œuvre ; 3- que la localisation linguistique dépend, elle, de l'histoire de la tradition du texte. La localisation décrit donc la couleur linguistique du texte tel qu'il est restitué par l'édition, conformément à la distinction philologique classique entre critique des leçons et critique des formes.

Né dans les années 1960 italiennes et aujourd'hui encore fidèle à cette tradition et à cette école, l'OVI fonde sa mission scientifique sur l'idée que, comme l'a écrit Gianfranco Contini, la « réalité » du document ne constitue pas, en soi, une « vérité »

19. Pour une description sommaire d'un article du *TLI*, se référer à « Avvertenze per la consultazione », disponible en ligne sur la page : http://tlio.ovi.cnr.it/TLIO/.

20. Voir Beltrami (2010), p. 245 : « *per la datazione del lessico non si può cessare di guardare al di là del dato materiale* ».

textuelle[21]. Considérant qu'il est essentiel d'éviter tout fétichisme des données matérielles qui amènerait à faire de la leçon attestée un absolu, entre deux abstractions possibles – trouver le texte dans des éditions ou dans des manuscrits, étant tous deux des hypothèses de travail –, le *TLIO* a choisi de se fonder sur des éditions, tout en sachant qu'une édition n'exprime qu'une tentative d'approcher la vérité du texte, tentative qui ne saurait donner des résultats absolus ni définitifs. L'application de ce principe informe les données d'attestation du lexème et la gestion de la documentation. Dans les articles, les premières attestations qui dérivent de corrections ou de conjectures de la part de l'éditeur critique sont signalées en tant qu'interventions éditoriales, mais sont en principe acceptées. L'application de ce principe implique aussi le fait que le *corpus* n'inclut que le texte critique, et ne dépouille pas les apparats.

Aujourd'hui le *corpus OVI dell'italiano antico*, librement consultable sur la toile[22], recueille 2 318 textes, pour un total de 23 173 538 occurrences de 467 548 formes graphiques (*tokens*) ; dans le *corpus* sont présents 116 596 lemmes, pour un total de 3 654 946 occurrences lemmatisées[23].

Le *corpus OVI* constitue aujourd'hui une ressource essentielle pour toute étude qui s'intéresse à l'ancien italien ou aux textes italiens médiévaux, compte tenu de son ampleur et du prestige philologique dont il jouit – le nom de Domenico De Robertis en est, en quelque sorte, le garant. Le *corpus OVI* est de fait devenu dès sa publication sur le réseau un outil de travail fondamental pour tout historien de la langue italienne, bien que sa conception originelle ne le vouât qu'à servir de base à la rédaction du *TLIO* : la promotion du *corpus* à source autorisée, quasiment

21. Il s'agit d'une idée continienne bien connue : qu'il suffise ici de citer l'article « Filologia » de l'*Enciclopedia del Novecento*, 1977, Roma, Istituto della Enciclopedia Italiana fondata da Giovanni Treccani (en ligne : http://www.treccani.it/enciclopedia/filologia(Enciclopedia-del-Novecento)/), récemment republié avec un commentaire de Lino Leonardi (voir Contini, 2014).
22. En ligne : http://gattoweb.ovi.cnr.it.
23. Les données se réfèrent au *corpus* mis à jour le 5 décembre 2014. Le *corpus OVI* est géré depuis 2006 par Elena Artale et Pär Larson ; pour une description et un bilan récent, *voir* Artale et Larson (2012).

incontournable, de données linguistiques ou – plus encore – la considération du *corpus* comme véritable autorité en matière de données linguistiques ne sont pas exemptes de risques.

Un *corpus*, comme n'importe quel outil, est construit avec une finalité spécifique : sa qualité et sa valeur se mesurent par rapport à son efficacité à répondre au besoin pour lequel il a été conçu et réalisé. Or, le *corpus OVI* a été construit pour être la source d'un vocabulaire de l'ancien italien : il se porte garant, pour le dire ainsi, de la fiabilité lexicale de ses données (et, là encore, il requiert tout de même de ses usagers une certaine finesse d'interprétation). Dans la base de données sont présents des textes pratiques tout comme des textes littéraires, des éditions d'autographes comme des éditions de textes transmis par des copies (par un seul témoin ou par plusieurs témoins, manuscrits ou imprimés), des éditions critiques récentes comme des éditions du XVIII^e^ ou XIX^e^ siècle, des *testi di lingua* (c'est-à-dire des textes représentatifs d'une variété linguistique déterminée) et des textes qui témoignent d'une langue peu caractérisée du point de vue diatopique. En généralisant et en simplifiant beaucoup, il est possible de dire que l'on trouve dans le *corpus* trois situations ecdotiques typiques, qui sont restées telles quelles depuis les années 1960[24] :

1) la poésie lyrique, spécialement celle du XIII^e^ siècle, jouit depuis toujours d'une position de prestige et a donc profité de l'attention de l'avant-garde de la philologie italienne – qu'il suffise de rappeler ici la publication des *Poeti del Duecento* de Gianfranco Contini (1960) –, ce qui fait qu'elle se lit dans des éditions critiques souvent récentes et très soignées d'un point de vue ecdotique (« néolachmannisme » italien) ;

24. À l'époque, tout comme aujourd'hui, une documentation conséquente était disponible pour l'ancien italien, mais de qualité très différente et surtout éditée selon des critères et pour des finalités très différents. Giorgio Pasquali, en 1941, évaluant la disponibilité de textes en vue du projet d'un vocabulaire, écrivait : « *Certo, parecchi testi della nostra letteratura, perfino di quelli del periodo più antico di essa, sono, per nostra vergogna, inediti. Questo è un caso eccezionale ; molto più frequentemente non si possono leggere se non in edizioni insufficienti…* » (*voir* Pasquali, 1941).

2) pour ce qui concerne les statuts et les documents pratiques, l'on dispose de beaucoup d'éditions qui sortent de la grande école italienne d'histoire de la langue (les *Testi fiorentini del Dugento e dei primi del Trecento* de Alfredo Schiaffini datent de 1926 ; les *Nuovi testi fiorentini del Dugento* d'Arrigo Castellani datent de 1951-1952 ; les *Testi veneziani* de Alfredo Stussi datent de 1965) : les *testi di lingua* sont jugés comme des témoins fiables d'une variété particulière et diatopique, et leurs éditions constituent la base documentaire sur laquelle s'appuie la description linguistique (notamment phonétique) de cette variété. Les éditions sont scrupuleuses et très conservatives, allant chez l'école Castellani jusqu'à maintenir, dans certains cas, la segmentation des mots dans la chaîne graphique telle que la présente le manuscrit (qui est bien sûr, souvent, un original et un autographe) ;

3) à côté, pour ainsi dire, de ces deux traditions ecdotiques, scientifiques et contemporaines se trouvent beaucoup de textes notamment en prose (des textes littéraires, des traductions, des encyclopédies et des traités) : ces textes sont bel et bien édités, mais ils l'ont été en majeure partie pendant le XIX^e^ siècle[25]. À cette époque, en effet, l'on assiste en Italie à une véritable course à l'édition de textes du « bon siècle », c'est-à-dire du XIV^e^ : celle-ci est l'une des voies qu'emprunte l'engagement politique du *Risorgimento*, le combat nationaliste pour l'indépendance italienne. L'existence d'une langue « nationale » qui connut pendant le XIV^e^ siècle sa période de splendeur est l'un des mythes fondateurs de l'unité italienne, et c'est dans ce sens que beaucoup de textes furent édités par divers érudits et savants engagés. Bien sûr, les critères ecdotiques sont préscientifiques : habituellement ces éditions suivent le témoignage d'un

25. Parmi eux bien sûr figure l'édition établie par Michele Barbi de la *Vita nova* de Dante (révision parue en 1932 d'une première édition datant de 1907), qui est considérée comme le point de départ de la philologie italienne, mais cette édition est restée en quelque sorte un cas isolé. Dans les années 1950, la parution des *Volgarizzamenti del Due e Trecento* de Cesare Segre (1953) et de la *Prosa del Duecento* de Cesare Segre et Mario Marti (1959) était censée stimuler la production d'éditions critiques, qui cependant ne se développèrent pas vraiment.

manuscrit, toujours corrigé au nom de l'orthopédie formelle (orthographique, mais aussi morphologique et syntaxique) et parfois corrigé aussi sur le plan de la « pureté » textuelle – les termes triviaux, par exemple, sont quelquefois censurés. Or, ces éditions sont bel et bien présentes dans le *corpus OVI*, parce qu'en général elles restituent fidèlement – à l'exception près de quelques champs lexicaux[26] – les lexèmes attestés par le manuscrit édité, même si elles affichent des interventions très lourdes du point de vue de la forme.

Si cette variété de typologies ecdotiques, donc, n'affecte et n'affaiblit pas pour autant la valeur des données du point de vue lexical, qui est le seul qui intéresse la rédaction du vocabulaire, il faut être conscient qu'elle pourrait compter bien davantage si l'on s'intéressait à d'autres points de vue, et tout spécifiquement à l'histoire de la langue pour la phonétique ou la morphologie. Le repérage de lexèmes et de formes à partir d'une base textuelle informatisée induit facilement un nivellement a-philologique des données, alors qu'une interprétation historique et critique attentive est nécessaire, pour éviter tout malentendu. Pour guider les chercheurs dans la consultation du *corpus*, et tout particulièrement ceux qui s'intéressent à l'histoire de la langue du point de vue phonétique ou morphologique, les textes affichant une couleur linguistique déterminée sont marqués par le sigle « TS » (« texte significatif »), qui permet de construire des sous-*corpus* de textes bien caractérisés du point de vue de la diatopie linguistique.

26. Pour le lexique des *realia* dans les textes de traduction, voir p. ex. Guadagnini (2015) : « *La tendenza a restituire la terminologia "corretta" (dal punto di vista dell'intelligenza del testo latino), alterando le rese originarie attestate nelle traduzioni medievali, interessa generalmente il lessico che, all'interno delle testimonianze medievali, potremmo definire "storico", vale a dire tutti quei vocaboli che fanno riferimento a una realtà del passato non proseguita o radicalmente mutata nella contemporaneità degli scriventi : si tratta insomma di un lessico che è insieme erudito, poco attestato e tendenzialmente concentrato nei testi di traduzione, come appunto gli etnici ma anche i nomi di vesti, monete, misure, cariche pubbliche, ecc. Dal punto di vista lessicografico, per la documentazione di questa tipologia lessicale è presumibilmente assai rilevante l'effetto che ha sortito sui dati l'attività editoriale moderna.* »

Quant au vocabulaire, le *TLIO* compte aujourd'hui presque 30 000 articles publiés sur la toile[27] : grâce à la liste complète des lemmes récemment rédigée par Rossella Mosti, l'on prévoit que le vocabulaire complet en rassemblera environ le double[28]. Le segment A-F est quasiment complet, mais on compte déjà beaucoup d'articles rédigés pour les segments alphabétiques suivants, jusqu'à Z : dans les dernières années surtout, l'équipe des rédacteurs du *TLIO* a eu tendance à adopter une stratégie de rédaction non pas alphabétique, mais plutôt onomasiologique ou étymologique au sens large, choisissant par exemple de rédiger tous les articles qui partagent le même hyper-étymon latin ou le même préfixe.

Depuis le 1er octobre 2014, l'OVI a un nouveau directeur, Lino Leonardi : une nouvelle phase s'ouvre pour l'institut et pour ses projets.

Le progrès du *TLIO*, ou peut-être tout simplement son passage de la théorie préventive à la pratique, a signifié l'abandon de certains aspects de l'article lexicographique qui étaient prévus encore au début des années 2000, comme la description des constructions verbales ou le recueil des synonymies et des antonymies : l'OVI, confiant dans les possibilités de la « lexicographie évolutive », espère pouvoir compléter ses articles avec ces informations dans le futur.

Bien qu'il ait quelque peu simplifié ses objectifs, il reste vrai que le *TLIO* est toujours ouvert à de nouvelles acquisitions ou améliorations qui modifieraient les articles publiés en ligne. Pour déjouer le « piège de l'instabilité[29] », le *TLIO*, doté d'un numéro ISSN qui lui est propre en tant que publication périodique, associe à chaque article deux dates : la date de la rédaction de la première version publiée, qui se trouve au point 0.8 de l'en-tête et qui se réfère à l'entrée de l'article dans le *TLIO*, et la date de la

27. En ligne : http://tlio.ovi.cnr.it/tlio/. Il s'agit de 29 422 articles exactement (mise à jour du 17 février 2015). Pour un bilan assez récent du vocabulaire, on peut se reporter à Beltrami (2009) ou à Squillacioti (2012).
28. La liste complète des lemmes est consultable en ligne : http://reddyweb.ovi.cnr.it.
29. Je cite à nouveau les mots de Robert Martin.

dernière mise à jour, qui se trouve à la fin de l'article et identifie la version que l'on est en train de consulter.

Le progrès du *TLIO* emmenant l'approfondissement des connaissances lexicologiques, au sein de l'OVI a mûri le désir de développer certains aspects du travail : les normes de rédaction étant désormais fixées et stables, de nouvelles voies s'ouvrent pour la documentation.

Je mentionnerai brièvement deux projets actuellement en cours de réalisation et qui développent la base de données principale, le *corpus OVI*. Le plus ancien est le projet *DiVo* (Dizionario dei volgarizzamenti, « Dictionnaire des traductions vernaculaires[30] »). Né de l'expérience de rédaction du *TLIO*, le *DiVo* vise à analyser la langue des traductions médiévales des classiques et des ouvrages latins de l'antiquité tardive, qui constituent une partie majeure de la documentation des origines – cet aspect n'a toutefois pas été pris en compte par la lexicographie italienne des deux siècles derniers, bien qu'il ait été discuté dans les quatre premières éditions du *Vocabolario della Crusca*[31]. Comme l'a bien résumé Cosimo Burgassi, « le *DiVo* cherche [...] à mettre en lumière les relations linguistiques et, plus généralement, culturelles, qui sont établies entre le modèle classique et la tradition littéraire médiévale à travers la traduction. [...] Ce plan d'investigation [...] enregistre les connexions culturelles entre la romanité et le Moyen Âge par la lexicologie[32] ». Le projet s'appuie sur deux bases de données qu'il a lui-même créées : le *corpus DiVo*, qui recueille les traductions vernaculaires associées par paragraphes au texte latin traduit, et le *corpus CLaVo*, qui recueille les textes latins traduits associés par paragraphes à la traduction vernaculaire[33]. Financée par

30. Hébergé par l'OVI et la Scuola Normale Superiore de Pise, dirigé par Giulio Vaccaro et moi-même, le projet *DiVo* a officiellement pris naissance en mars 2012 et se développera pendant quatre ans. Pour une description du projet, voir Guadagnini et Vaccaro (2014) et Burgassi (2014).
31. Se référer à Guadagnini (2013).
32. Se référer à Burgassi (2014).
33. Le *corpus DiVo* (en ligne : http://divoweb.ovi.cnr.it) recueille exhaustivement les traductions médiévales de classiques latins – de Cicéron à Boèce, ce dernier étant considéré comme limite conventionnelle – et une sélection plutôt riche de traductions

l'État italien il y a cinq ans, l'équipe du projet a complété les bases de données et travaille actuellement à la deuxième phase du projet, qui se propose d'isoler les particularités de la langue des *volgarizzamenti* par rapport aux autres sous-codes de l'ancien italien[34].

Le second projet, *ReMediA* (*Repertorio di Medicina Antica*, « Répertoire de médecine ancienne »), dirigé par Ilaria Zamuner et Elena Artale, représente quant à lui la première tentative de création d'un *corpus* plurilingue à l'OVI. Le *corpus ReMediA*, dont une première version a été publiée sur la toile le 29 juillet 2014[35], est conçu pour restituer la spécificité de la terminologie technique et en particulier de la langue de la médecine et de la pharmacopée médiévale, dont Elena Artale et Ilaria Zamuner sont spécialistes : cette terminologie est à la fois un sous-ensemble assez clos du lexique vernaculaire et un domaine que l'on comprend mieux dès lors que l'on considère le contexte plurilingue dans une perspective comparative[36]. Le *corpus* va recueillir les principaux traités médicaux latins et romans, qui sont souvent des traductions d'un texte latin préexistant : la possibilité de comparer le latin et les diverses langues vernaculaires va permettre de mieux comprendre les textes du point de vue linguistique (lexical, bien sûr, mais aussi syntaxique) et également du point de vue historique. Le récent travail d'Ilaria Zamuner portant sur les résultats romans du lexème latin *aranea* (*tunica*)[37] fournit un premier exemple des résultats que permet d'obtenir l'application de cette méthode comparative.

---

d'ouvrages grecs (qui ont été traduits à partir d'un texte latin intermédiaire, tel l'*Éthique à Nicomaque* d'Aristote) et des pères de l'Église, composées en n'importe quelle variété de l'italien. À côté de ces traductions, le *corpus DiVo* regroupe également les éléments para-textuels (les gloses marginales et interlinéaires, les commentaires et les glossaires éventuellement associés à la traduction). Le *corpus CLaVo* (en ligne : http://clavoweb.ovi.cnr.it) recueille, comme on l'a exposé, les textes latins, ordonnés du plus ancien au plus récent suivant la chronologie des traductions vernaculaires associées. *Voir* Dotto (2013).

34. Pour les premiers résultats, voir Burgassi et Guadagnini (2014), Dotto (2015) et Guadagnini et Vaccaro (2011).
35. En ligne : http://remediaweb.ovi.cnr.it.
36. Voir par exemple Artale (2014). Pour une description synthétique du projet, consulter : http://www.sifr.it/ricerca/remedia.pdf.
37. Voir Zamuner (2015).

Ici se termine ce tour d'horizon de l'activité passée et présente de l'OVI, qui a permis d'évoquer quelques moments cruciaux et quelques aspects de ses travaux : l'activité de l'OVI est certes centrée sur la rédaction du *TLIO*, mais l'institut a aussi développé et développe, tout autour de son dictionnaire, des outils et des méthodes interconnectés, visant à approfondir toujours plus les connaissances déposées dans son œuvre majeure et les possibilités de recherche mises à la disposition des chercheurs. Pour cela, le pari sur l'informatique – qui, pris il y a un demi-siècle, témoigne d'un esprit véritablement pionnier – s'est révélé un choix des plus heureux.

Références bibliographiques

Artale, Elena et Larson, Pär, « Il punto sui corpora dell'Opera del Vocabolario Italiano », *Dizionari e ricerca filologica. Atti della Giornata di studi in memoria di Valentina Pollidori, Firenze, 26 ottobre 2010 (Supplemento III al Bollettino dell'Opera del Vocabolario Italiano)*, Alessandria, Edizioni dell'Orso, 2012, p. 25-40.

Artale, Elena, « Testi medici antichi e banche dati informatizzate. L'indicizzazione come risorsa ecdotica ed esegetica », dans Garavelli, Enrico et Suomela-Härmä, Elina (dir.), *Atti del XII Congresso SILFI (Helsinki, 18-20 giugno 2012)*, Firenze, Franco Cesati Editore, 2014, p. 43-50.

Avalle d'Arco Silvio, *Al servizio del vocabolario della lingua italiana*, Firenze, Accademia della Crusca, 1979.

Beltrami, Pietro G., « The Lexicography of Early Italian : Its Evolution and Recent Advances », dans Bruti, Silvia, Cella, Roberta et Foschi Albert, Marina (dir.), *Perspectives on Lexicography in Italy and Europe*, Newcastle upon Tyne, Cambridge Scholars Publishing, 2009, p. 27-53.

—, « Lessicografia e filologia in un dizionario storico dell'italiano antico », dans Ciociola, Claudio (dir.), *Storia della lingua italiana e filologia. Atti del VII Convegno ASLI (Pisa-Firenze, 18-20 dicembre 2008)*, Firenze, Cesati, 2010, p. 235-248.

Burgassi, Cosimo, « Le projet DiVo et ses corpus : une base de données italo-latine de traductions médiévales », *Bulletin du centre d'études médiévales d'Auxerre*, n° 18, 2014/1.
En ligne : http://cem.revues.org/13423.

Burgassi, Cosimo et Guadagnini, Elisa, « Prima dell' "indole", Latinismi latenti dell'italiano », *Studi di Lessicografia Italiana*, n° 31, 2014, p. 1-39.

Ceccoli, A., Lorenzi, F. et Pollidori, V., « Un programma per la redazione del Vocabolario Storico della Lingua Italiana assistita dal calcolatore », dans *Récit et informatique. Actes de la journée d'études, C.R.L.L.I., Université de Paris X - Nanterre, 9 décembre 1989*, Claude Cazalé Bérard (éd.), La Garenne-Colombes, Éditions de l'Espace européen, 1991, p. 85-106.

Contini, Gianfranco, *Filologia*, Bologna, Il Mulino, 2014.

De Robertis, Domenico, « L'ufficio filologico dell'Opera del vocabolario, il suo impianto, il suo lavoro », dans Alfieri, Gabriella *et al.*, *La Crusca nella tradizione letteraria e linguistica italiana*, Firenze, Accademia della Crusca, 1985, p. 443-451.

Dotto, Diego, « Note per la lemmatizzazione del corpus DiVo », *Bollettino dell'Opera del vocabolario italiano*, n° 17, 2012, p. 336-364.

—, « Notizie dal DiVo. Un primo bilancio sulla costituzione del corpus », dans Larson, Pär, Squillacioti, Paolo et Vaccaro Giulio (dir.), « *Diverse voci fanno dolci note », L'Opera del vocabolario italiano per Pietro G. Beltrami*, Alessandria, Edizioni dell'Orso, 2013, p. 71-83.

—, « Esercizi sul contributo del lessico di traduzione in lessicografia : dal TLIO al DiVo », dans Buchi, Éva, Chauveau, Jean-Paul et Pierrel, Jean-Marie (dir.), *Actes du XXVII^e Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013)*, Strasbourg, Société de linguistique romane/ ÉliPhi, 2015.

Duro, Aldo, « L'impianto del nuovo vocabolario: profilo storico », dans Alfieri, Gabriella *et al.*, *La Crusca nella tradizione letteraria*

e linguistica italiana, Firenze, Accademia della Crusca, 1985, p. 431-442.

ESPERTI Piero, « Grammatichetta della lingua italiana ad uso del calcolatore », dans AVALLE, d'Arco Silvio (dir.), *Al servizio del vocabolario della lingua italiana*, Firenze, Accademia della Crusca, 1979, p. 123-187.

GUADAGNINI, Elisa, « Notizie dal DiVo. Parole tradotte e lessicografia dell'italiano », dans LARSON, Pär, SQUILLACIOTI, Paolo et VACCARO, Giulio (dir.), « *Diverse voci fanno dolci note », L'Opera del vocabolario italiano per Pietro G. Beltrami*, Alessandria, Edizioni dell'Orso, 2013, p. 59-70.

—, « Variazioni aborigene : note di lessicografia dell'italiano antico », *Bollettino dell'Opera del vocabolario italiano*, 2015.

GUADAGNINI, Elisa et VACCARO, Guilio, « "*Nom de pays : le nom…*" Parole, paesi e popoli nel corpus DiVo », dans LUBELLO, Sergio (dir.), *Volgarizzare, tradurre, interpretare nei secc. XIII-XVI. Atti del Convegno internazionale di studio : Studio, archivio e lessico dei volgarizzamenti italiani (Salerno, 24-25 novembre 2010)*, Strasbourg, Éditions de linguistique et de philologie, 2011, p. 267-281.

GUADAGNINI, Elisa et VACCARO, Guilio, « Un contributo allo studio del "volgarizzare e tradurre" : il progetto DiVo », dans PACCAGNELLA, Ivano et GREGORI, Elisa (dir.), *Lingue, testi, culture : L'eredità di Folena, vent'anni dopo. Atti del XL Convegno Interuniversitario (Bressanone, 12-15 luglio 2012)*, Padova, Esedra editrice, 2014, p. 91-105.

MOSTI, Rossella, « Tra lemma e voce: ruolo della preredazione nel *Tesoro della lingua italiana delle origini* », *Dizionari e ricerca filologica. Atti della Giornata di studi in memoria di Valentina Pollidori, Firenze, 26 ottobre 2010 (Supplemento III al Bollettino dell'Opera del Vocabolario Italiano)*, Alessandria, Edizioni dell'Orso, 2012, p. 85-99.

PASQUALI, Giorgio, « Per un tesoro della lingua italiana », *Atti della R. Accademia d'Italia. Rendiconti della Classe di scienze morali e storiche*, s. 7, II, 1941, p. 490-521.

Squillacioti Paolo, « Uno sguardo al *Tesoro della Lingua Italiana delle Origini*: procedure e prospettive del vocabolario storico dell'italiano antico », *Dizionari e ricerca filologica. Atti della Giornata di studi in memoria di Valentina Pollidori, Firenze, 26 ottobre 2010 (Supplemento III al Bollettino dell'Opera del Vocabolario Italiano)*, Alessandria, Edizioni dell'Orso, 2012, p. 74-84.

Tomasin, Lorenzo, « Qu'est-ce que l'italien ancien ? », *La Lingua Italiana*, n° 9, 2013, p. 1-18.

Vaccaro, Giulio, « Veniamo da molto lontano e andiamo molto lontano. Documenti per la storia dell'Opera del Vocabolario Italiano dalle origini al 1992 », *Bollettino dell'Opera del Vocabolario Italiano*, n° 18, 2013, p. 277-390.

Zamuner, Ilaria, « *'Aranea' tunica* e la lessicografia medico-scientifica romanza », *Cultura Neolatina*, n° 75, 2015/1.

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

Ana Gómez Rabal
CSIC
Institution Milá y Fontanals (Barcelone, Espagne)

La réalisation de tout projet lexicographique peut susciter l'intérêt plus ou moins vif des spécialistes, mais aussi des personnes intéressées par des matières diverses. Nous voulons présenter ici un projet fondé et réalisé par des philologues (spécialistes de la langue latine, et en particulier médiévale) dont le but dépasse le cercle de la philologie *stricto sensu*. Il s'agit d'un dictionnaire, le *Glossarium Mediae Latinitatis Cataloniae ab anno DCCC usque ad annum MC* (*GMLC*), résultat de l'étroite collaboration entre deux institutions, le Consejo superior de investigaciones científicas (CSIC) – le « Conseil supérieur de la recherche scientifique » – et l'université de Barcelone, et réalisé dans le centre du CSIC qui répond au nom d'Institution Milá y Fontanals, à Barcelone (Espagne). Ce dictionnaire, remarquons-le, veut intéresser non seulement les philologues, mais aussi les historiens, les juristes et toute personne attirée par le Moyen Âge, d'où le désir de faire de cette œuvre un instrument de référence indispensable tout aussi bien, par exemple, pour le latiniste qui aborde l'étude d'un champ lexical déterminé du latin médiéval hispanique, que pour le romaniste qui examine des problèmes de la grammaire historique ou pour le juriste qui analyse les formes d'élaboration des testaments.

La réalisation d'une ambition lexicographique comme celle qu'incarne le *GMLC* s'inscrit dans un milieu, le milieu européen, où fleurissent de nombreux dictionnaires, soit nationaux soit

régionaux, de latin médiéval et où s'est forgé un projet qui se maintient dans toute sa vigueur, celui d'un dictionnaire européen commun de latin médiéval. Le *GMLC* suit, donc, les voies tracées en Europe par la philologie latine médiévale et la lexicographie et doit affronter, en toute logique, les mêmes défis que d'autres projets jumeaux.

Pour ces deux raisons – parce que le *GMLC* veut être un instrument utile aux médiévistes quel que soit leur champ d'études, et aussi parce que la trajectoire du glossaire est parallèle à celle d'autres projets européens – il nous semble indispensable de donner, tout d'abord, un aperçu rapide de l'histoire du glossaire et de revenir, ensuite, sur les phases de travail actuellement développées.

Présentation

Le *Glossarium Mediae Latinitatis Cataloniae ab anno DCCC usque ad annum MC* prétend, nous l'avons indiqué, offrir aux philologues, historiens, juristes et, en général, à toute personne intéressée par le Haut Moyen Âge, la documentation latine écrite en Catalogne du IX^e^ au XII^e^ siècle. C'est ce que souligne le sous-titre de l'édition imprimée du glossaire : *Voces latinas y romances documentadas en fuentes catalanas del año 800 al 1100*[1], « Mots latins et romans documentés à partir de sources catalanes de l'année 800 à l'année 1100 ».

Ces « mots latins et romans » sont recueillis, étudiés et systématisés à partir, essentiellement, de la lecture et du dépouillement de documents notariés, à savoir des actes de donation, de dotation, de vente, d'acquisition, des testaments, des serments de fidélité et des litiges, dont la production est une constante très riche dans les territoires correspondant aux domaines linguistiques du catalan à partir du IX^e^ siècle, pendant le Haut Moyen Âge[2].

1. Voir Mariano Bassols et Joan Bastardas (1960-1985) et Joan Bastardas (2006).
2. Concernant les éditions de ces chartes et des autres œuvres correspondant à l'époque et aux territoires cités, voir Ana Gómez Rabal (2008).

On doit souligner cette abondance face à la presque inexistence dans la Catalogne de l'époque d'autres types de textes légaux, les textes de création législative, à une seule exception près, celle des textes légaux et des textes de droit coutumier locaux qui s'inspirent du *Liber Iudicum* wisigothique[3].

Il faut considérer également cette abondance de chartes juridiques au regard de la rareté de tout autre genre de « littérature » dans le sens le plus large du terme, c'est-à-dire celui qui englobe les belles lettres, les œuvres historiques, philosophiques, les traités de rhétorique ou les livres techniques. Parmi ces rares exceptions avant le milieu du XII^e^ siècle, nous devons citer les écrits littéraires de l'abbé Oliba (*ca*. 970-1046),[4] la *Vita Beati Petri Vrseoli* et l'*Epistola Garsiae monachi Cuxanensis*.[5] Une autre exception remarquable : les textes techniques et scientifiques écrits par Lupitus, archidiacre de Barcelone, ou attribués à lui[6].

3. Le *Liber Iudicum*, ou *Liber Iudiciorum*, est le recueil sur lequel s'est basée la pratique juridique haut-médiévale catalane et qui a servi de modèle pour l'élaboration des textes légaux et des textes de droit coutumier locaux. Voir Josep Maria Font i Rius (1969 et 1983). Sur l'application du *Liber Iudicum* en correspondance avec les canons conciliaires et les capitulaires carolingiens, se référer à Font i Rius (2003, en particulier p. 71-72).

4. Entre eux, toujours dans un style soigné, on compte une assez riche collection de lettres, dont seulement sept pièces ont été conservées, deux d'entre elles adressées à ses moines de Ripoll, deux à Sancho III de Pampelune (1000-1035), une à Gauzlin, archevêque de Bourges et abbé de Fleury, et une encyclique à d'autres monastères annonçant la mort (en 1020) de son frère le comte Bernard ; et on compte aussi des poèmes, parmi lesquels un éloge historique de Ripoll (*ca*. 1032), intégrant un résumé de son passé rédigé en hexamètres, des louanges à certains membres de la famille comtale des Oliba, et un poème élogieux également adressé à son ami l'abbé de Fleury. Pour l'édition de ces textes, se référer à Eduard Junyent i Subirà (1992).

5. La première de ces deux œuvres, composée entre 1075 et 1100, a été éditée par Jean Mabillon (1737). Pour une édition de la deuxième œuvre, écrite entre 1040 et 1046, voir Petrus de Marca (1688).

6. Ce petit corpus technico-scientifique contient diverses œuvres portant sur un sujet concret, l'astrolabe, et sur d'autres questions de géométrie ou d'arithmétique. Pour l'édition de ces traités, voir José Maria Millàs Vallicrosa (1931) ; en particulier sur l'astrolabe, se référer aux p. 271-275 (Lupitus Archidiaconus, *Prologus in libellum de astrolabio*, *ca*. 980, manuscrit du XI^e^ siècle), p. 275-293 (*Sententie astrolabii*, traduction d'un original arabe du X^e^ siècle attribué, par conjecture, à Lupitus de Barcelone), p. 293-295 (Anonymus, *De mensura astrolapsus*, X^e^ siècle, manuscrit du XI^e^ siècle), p. 304-305 (Anonymus, *Regulae de astrolabio*, X^e^ siècle, manuscrit du XI^e^ siècle) et p. 308-315 (Anonymus, *De astrolabio*, X^e^-XI^e^ siècles).

Ce sont tous ces textes, qu'ils relèvent d'un type ou d'un autre (juridique, littéraire, scientifique), produits entre les années 800 et 1100, qui ont été dépouillés et insérés dans les fichiers du *GMLC,* et qui forment le corpus des textes servant de base pour l'élaboration du dictionnaire. Mais c'est surtout la documentation qui reproduit les actes juridiques – la documentation notariée – qui concentre l'intérêt des lexicographes du *GMLC*, parce qu'elle se convertit en un témoignage écrit privilégié de la langue latine médiévale ; expression du langage juridique, ecclésiastique, institutionnel et curial du Haut Moyen Âge, époque où affleurent des innovations lexicales importantes et des indices très clairs, de nature phonétique ou morpho-syntaxique, d'une langue romane qui commence à percer. Ce latin documentaire présente souvent, même pour un bon latiniste, de grandes difficultés quant à l'interprétation, car deux phénomènes s'entrecroisent : le latin subit l'influence plus ou moins forte de la langue romane et, parallèlement, la langue romane se latinise, et apparaît fréquemment comme le fruit d'un effort de latinisation d'expressions et de formes déjà pleinement romanes. Le *GMLC*, constitué d'articles lexicographiques élaborés surtout à partir des documents notariés que nous venons d'évoquer, cherche à aplanir le chemin ouvert à ceux qui souhaitent affronter les difficultés de cette langue.

Histoire

Le Glossarium Mediae et Infimae Latinitatis *de Charles Dufresne, sieur Du Cange, et le* Novum Glossarium Mediae Latinitatis

Le *Glossarium Mediæ et Infimae Latinitatis* de Charles Dufresne, sieur Du Cange, est assurément le premier parmi les ouvrages de référence que tout médiéviste, quel que soit le domaine de recherche qu'il a choisi, s'est immanquablement vu dans la nécessité de consulter. Cette œuvre lexicographique essentielle – connue comme « le Du Cange » –, publiée en 1678, et dont les rééditions successives jalonnent les XVIII^e^ et XIX^e^ siècles,

constitue un point de départ inéluctable pour toute tentative de compilation minutieuse de la langue latine médiévale.

La nécessité de créer un nouveau dictionnaire du latin du Moyen Âge répondant à des critères scientifiques modernes devint, au début du XX^e^ siècle, de plus en plus évidente. Ce vœu fut formulé au cours d'un congrès d'histoire tenu à Londres en 1913 ; bientôt le projet se dessina et, en 1920, il reçut l'appui de l'Union académique internationale. Répondant à la désignation officielle de *Novum Glossarium Mediæ Latinitatis* (NGML), il reste connu par les philologues et les historiens comme le « nouveau Du Cange ».

Le travail de compilation, de lecture et de dépouillement des textes commença en 1924, simultanément dans six pays : la Belgique, la France, le Royaume-Uni, l'Italie, les Pays-Bas et la Pologne. Chacun d'eux constitua peu à peu son propre fichier, et le premier travail fut rendu public en 1936 avec le premier fascicule (*a-agradior*) du *Latinitatis Italicæ Medii Aevi Lexicon Imperfectum*, dirigé par Francesco Arnaldi et publié dans l'*Archivum Latinitatis Medii Ævi* (ALMA). En 1953 le comité polonais fit paraître le premier fascicule du *Lexicon Mediæ et Infimæ Latinitatis Polonorum*, dirigé par Marian Plezia ; en 1975, le comité britannique publia à son tour la première livraison de son *Dictionary of Medieval Latin from British Sources*, dirigé par Ronald Edward Latham. Dans les années 1930, de nouvelles équipes de travail s'incorporèrent au projet, en Suède, en Hongrie et en Tchécoslovaquie, et les académies de Munich et Berlin, en 1959, entreprirent la publication de leur *Mittellateinisches Wörterbuch*, à la charge d'Otto Prinz. D'autres pays se joignirent à l'entreprise, tels la Finlande, le Danemark, l'Irlande ou la Yougoslavie.

Mais l'apparition en 1957 du premier fascicule du *Novum Glossarium Mediae Latinitatis*, correspondant à la lettre L, à la charge de Franz Blatt, entérina définitivement la marche de cette entreprise commune qui maintient aujourd'hui tout son dynamisme. Franz Blatt précisait dans son « Avis au lecteur » le sens et la portée de cette œuvre :

> Le nouveau dictionnaire du latin médiéval ne prétend nullement remplacer le *Glossarium mediae et infimae latinitatis conditum a C. du Fresne domino* Du Cange, celui-ci étant à la fois une encyclopédie et un dictionnaire ; nous avons, par conséquent, réduit au maximum les explications d'ordre historique et technique. Le point de vue purement lexicologique prévaut.

Et, un peu plus bas :

> Le nouveau dictionnaire contient [...] non seulement les néologismes médiévaux, mais aussi les mots et sens classiques encore vivants, le but du dictionnaire étant de donner une description sinon complète, du moins fouillée de la langue latine au Moyen Âge.

Actuellement le Comité Du Cange, dont le siège est à Paris, poursuit la rédaction du *NGML*. Le comité est régi par l'Académie des inscriptions et belles-lettres, par le Centre national de la recherche scientifique (CNRS) et par l'École pratique des hautes études (EPHE, 4e section de sciences historiques et philologiques), et constitue la section de lexicographie latine de l'Institut de recherche et d'histoire des textes (CNRS)[7].

Le Glossarium Mediae Latinitatis Cataloniae ab anno DCCC usque ad annum MC

À Barcelone, un groupe de professeurs réunis sous la direction de Mariano Bassols de Climent, et parmi lesquels se trouvait Joan Bastardas, décida au cours de l'année scolaire 1952-1953 de se joindre au projet du *Novum Glossarium* (*NGML*). Dans le but de donner davantage de matière au *NGML*, le groupe entama un travail de dépouillement de textes latins médiévaux hispaniques – spécialement, mais non exclusivement catalans – et d'élaboration des fiches correspondantes. Tout cela se fit à partir d'éditions des textes documentaires bien que, lorsque la rigueur scientifique l'exigeait, on eut recours à la lecture directe des documents.

7. Concernant l'histoire, l'organisation et le but scientifique poursuivi par l'équipe de lexicographie latine médiévale, voir en ligne : http://www.irht.cnrs.fr/fr/recherche/sections/lexicographie-latine, et : http://www.aibl.fr/travaux/moyen-age/article/le-bureau-du-cange?lang=fr [consultés le 21 juin 2017].

Bientôt cependant le groupe envisagea la possibilité de publier, en mettant à profit le travail de dépouillement réalisé et qui continuait à se réaliser, son propre glossaire. C'était l'occasion tout d'abord de recueillir le lexique qui pourrait être exclu d'un dictionnaire plus général, car il était logique qu'entre les rédacteurs du *NGML* primât le critère de choisir des exemples communs à tout l'Occident européen ; d'où la perte de certaines nuances quant aux différences et caractères particuliers du latin des divers territoires, où peu à peu apparaissaient des traits socioculturels et linguistiques spécifiques.

D'autre part, proposer aux responsables du *NGML* un corpus non seulement élaboré, mais aussi publié semblait beaucoup plus fructueux que de leur remettre des fiches qui pourraient présenter des difficultés d'interprétation. Autrement dit, on considéra qu'il était souhaitable d'offrir au *NGML* des matériaux révisés par des chercheurs connaissant la langue, l'histoire, les conditions sociales, économiques, institutionnelles, la culture et les coutumes du territoire.

Enfin, au cours de l'année scolaire 1956-1957, époque où l'École de philologie de Barcelone du Conseil supérieur de la recherche scientifique disposa, enfin, de locaux appropriés, il fut décidé d'entreprendre la publication d'un glossaire à partir des sources documentaires catalanes, surtout notariées, et dont les limites chronologiques iraient de l'année 800 à l'année 1100. Les premiers articles furent soumis à des romanistes et latinistes européens qui conseillèrent à l'équipe de poursuivre ses travaux, et celle-ci publia en 1960 le premier fascicule (*a-aragalius*) du *GMLC*.

C'est le professeur Mariano Bassols de Climent qui fut à l'initiative de la création et de la première organisation du *GMLC*, mais l'entreprise et la réalisation des articles lexicographiques ont été, dès le début, de la responsabilité du professeur Joan Bastardas.

Dans les années 1960, le projet initié par l'École de philologie latine de Barcelone paraissait donc parfaitement viable, puisque la documentation latine notariée des siècles compris entre le IX^e^ et le XII^e^ est en Catalogne, nous l'avons souligné, abondante, riche et suffisamment variée.

Mais l'intérêt de l'entreprise était aussi conforté par d'autres tentatives, antérieures, d'étude et de systématisation des caractéristiques du latin du Haut Moyen Âge en Catalogne. L'antécédent le plus important, déjà lointain mais source d'une inspiration réelle pour le nouveau projet, était le corpus de matériel lexicographique rassemblé par Josep Balari i Jovany (1844-1904). Les collections de fiches (16 000 fiches) écrites et compilées par Balari i Jovany furent la base sur laquelle ce chercheur composa son œuvre *Orígenes históricos de Cataluña* (*Les Origines historiques de la Catalogne*, Barcelone, 1899). En 1956, le professeur Joachim Carreras i Artau hérita de ces fiches, et les transmit au professeur Bastardas. Elles sont un point de repère exceptionnel pour l'équipe du *GMLC*.

Réalisation du *Glossarium Mediae Latinitatis Cataloniae*

L'équipe de rédaction dirigée par le professeur Bastardas poursuivit, toujours dans la même orientation, son travail jusqu'à compléter la publication du premier volume en 1985. Ce volume, qui correspond aux lettres A, B, C et D, contient une préface, une introduction générale, une bibliographie et plus de 1 000 articles offrant une ample information linguistique sur le latin médiéval et le catalan pré-littéraire, articles enrichis par des notes à caractère archéologique, historique et culturel.

Au cours des années 1990, les possibilités qu'offrirent les nouveaux moyens techniques ouvrirent la voie à l'élaboration d'un fichier informatisé. Ce nouveau pas fut jugé opportun et utile. En vue de la rénovation du programme – continuer le travail de dépouillement des documents et de rédaction des articles et parallèlement commencer le travail d'informatisation du fichier manuel constitué par 50 000 fiches papier – les accords établis entre l'Institution Milá y Fontanals du CSIC et l'université de Barcelone et, depuis 1985, entre ces deux centres et l'Institut d'études catalanes se sont renforcés. Il faut ajouter à cela les subventions ministérielles obtenues sous l'impulsion de Pere J. Quetglas Nicolau, professeur du Département de

philologie latine de l'université de Barcelone et successeur du professeur Bastardas comme directeur de l'équipe.

État actuel du projet

La réalisation du *GMLC* est actuellement structurée autour de deux phases de travail. La première correspond à sa rédaction, et la seconde, à sa numérisation.

Pour cette seconde phase, la direction de l'équipe a tout d'abord pensé, nous l'avons dit, à informatiser le fichier manuel, mais dès les années 2001-2002 le groupe de travail prit une décision différente et s'assigna un autre but : celui de numériser la base documentaire constituée par les éditions des textes, sans négliger les critères de qualité dans la sélection des éditions. On projette de créer ainsi, sur un support digital, un fichier riche, inédit, ouvert aux variations et aux élargissements (exigés, sans doute, par le dépouillement de nouvelles publications documentaires), cohérent, intelligible et facile à manier. Le fait de numériser permet, en outre, la création et l'utilisation de concordances – ou index alphabétiques de mots, d'unités lexicales, d'un texte avec leur contexte significatif – qui confirment, complètent et élargissent l'information recueillie dans les fiches sur lesquelles travaillent les rédacteurs du *GMLC*. Le corpus de textes corrigés à partir d'éditions et de transcriptions est constitué actuellement par plus de 23 000 documents.

Les deux phases de travail – rédaction et numérisation – sont complémentaires et indissociables. Ce tout s'est matérialisé, d'une façon encore très partielle, par la publication du fascicule 11, qui correspond à la lettre F (2001) ; d'une façon un peu plus complète, par la rédaction et la publication du fascicule 12, qui correspond à la lettre G (2006) ; d'une façon beaucoup plus systématique, par la préparation de la réédition des lettres A-D (prêtes pour publication papier en 2010) et, surtout, par la préparation des articles du fascicule E[8].

8. Rédaction achevée en 2016.

Mais, en ce qui concerne cette phase de numérisation, nous devons souligner surtout que l'équipe est en train de convertir le fichier numérique, d'usage interne jusqu'en 2011, en un fichier accessible à toute personne intéressée par la documentation latine produite tout au long du Haut Moyen Âge en Catalogne. En tant que coordinatrice de ce projet, et sous la direction du professeur Quetglas, je me suis spécialement chargée de diriger une partie des efforts de l'équipe vers le but annoncé, c'est-à-dire le dépouillement, l'édition et la publication de notre corpus de textes ; pour le dire autrement, vers la publication d'une base de données lexicale de consultation publique, le *Corpus Documentale Latinum Cataloniae* (CODOLCAT)[9].

Ce projet de création d'un nouveau service de consultation externe de la base de données lexicale numérisée a obéi au désir d'utiliser les avantages des nouvelles technologies et s'inscrit dans la lignée d'un autre projet espagnol : le *Corpus Documentale Latinum Gallaeciae* (CODOLGA), dirigé par le docteur José Eduardo López Pereira, professeur à l'université de la Corogne[10]. Ce projet pionnier a servi d'inspiration et de modèle pour la conception du CODOLCAT. Le CODOLCAT est une plate-forme multilingue qui offre la possibilité de procéder à des requêtes simples ou d'affiner la recherche par type de producteur du document, par localisation, selon un éventail chronologique, etc. ; les résultats sont affichés sous la forme de concordances. Nous avons publié cinq versions du CODOLCAT. La première d'entre elles (v. 1, 2012) permettait l'accès à environ 1 000 actes, correspondant à 4 cartulaires ; la deuxième version (v. 2, 2013) était constituée par environ 2 400 actes, correspondant à 10 cartulaires ; la troisième version (v. 3, 2014), par environ 3 600 actes, correspondant à 14 cartulaires ; la quatrième (v. 4, 2015) met à la portée du lecteur 5 200 actes, correspondant à 17 cartulaires ; la cinquième (v. 5, 2016), 6 300 actes de 21 cartulaires. Notre intention est de continuer à publier une version annuelle avec un incrément approximatif de 1 000 chartes, au minimum, par an.

9. En ligne : http://gmlc.imf.csic.es/codolcat.
10. En ligne : http://corpus.cirp.es/codolga (v. 12, 2015) [consulté le 18 octobre 2016].

Il existe actuellement en Espagne un troisième projet d'élaboration et de publication d'un corpus numérisé suivant le même modèle : il s'agit du *Corpus Documentale Latinum Valencie* (CODOLVA), sur la documentation du royaume de Valence et, au Portugal, du *Corpus Documentale Latinum Portucalense* (CODOLPOR), de l'université de Lisbonne. Il existe d'autre part en Espagne une autre équipe de lexicographie médio-latine, qui s'occupe de la documentation des royaumes de León et de Castille et a déjà publié son *Lexicon Latinitatis Medii Aevi Regni Legionis imperfectum (s. VIII-1230)*, sous la direction des professeurs Maurilio Pérez González (université de León) et Estrella Pérez Rodríguez (université de Valladolid).

Pour l'équipe du *GMLC*, la continuation de la préparation et de la publication du glossaire ainsi que la poursuite de l'édition et de la publication annuelle du CODOLCAT constitueront toujours deux objectifs inséparables.

Défis pour le *GMLC* et pour la lexicographie latine médiévale

L'équipe du *GMLC* prépare la publication numérique des fascicules déjà rédigés[11]. L'accès à cette publication s'effectuera librement par internet. La préparation de cette publication constitue un enjeu concret pour notre groupe en raison de la compatibilité des deux outils que nous voulons offrir : la base de données lexicale (le CODOLCAT) et le dictionnaire lui-même. Voyons les points sur lesquels va se structurer cette compatibilité recherchée :

1. Dès le service de requête du dictionnaire, toutes les données contenues dans le CODOLCAT (lexicales ou autres) doivent être récupérables ; et, à l'inverse, toute l'information balisée et structurée – donc indexée – que contient le *GMLC* devra être récupérable quand l'usager interrogera le CODOLCAT.
2. La plate-forme sera flexible : nous sommes en train de créer des formulaires d'introduction des données destinés aux

11. La préparation de l'édition numérique a été commencée à partir des articles lexicographiques revus pour la réédition papier des lettres A-D.

rédacteurs. Ces formulaires rendront plus aisées l'édition et la publication des articles de l'œuvre.

3. S'agissant de l'apparition des exemples, un dictionnaire est forcément sélectif. Dans le CODOLCAT, au contraire, le lecteur pourra accéder à tous les exemples.
4. Il sera possible de réaliser des requêtes et des recherches qui ne seront pas strictement lexicales. Quand nous introduisons les données dans le CODOLCAT, nous recueillons beaucoup d'informations sur les archives ou les bibliothèques où se conservent les chartes et les manuscrits, sur le type d'acte juridique concerné, sur l'écriture, etc. Il sera possible de répondre à des questions posées par des historiens, à des interrogations prosopographiques, géographiques, etc.

Mais notre objectif va plus loin encore : il veut dépasser cette cohérence acquise entre la base de données et le dictionnaire. Notre souhait, notre ambition est d'obtenir la compatibilité avec les autres dictionnaires de latin médiéval et de pouvoir offrir – comme les créateurs du *Novum Glossarium* l'exprimaient – un moyen de consultation commune pour le latin médiéval dans toute sa diversité, géographique, stylistique, même chronologique[12]. Il ne s'agit nullement de créer un serveur unique, mais une plate-forme commune où chaque groupe déposerait les données qu'il déciderait lui-même de partager. La technologie nous offre maintenant cette possibilité. Le besoin, l'exigence même d'une collaboration entre les équipes – et cela malgré les différences dans les buts déjà atteints et malgré la diversité des méthodes et des moyens employés – correspondent désormais à un désir objectivement réalisable[13].

12. Si le *Novum Glossarium* a pour limites chronologiques les années 800 et 1200, les divers projets européens de lexicographie latine du Moyen Âge correspondant au domaine linguistique d'une langue (romane ou non romane) ont des marges chronologiques très différentes : ainsi, pour ne donner que deux exemples, le *Dictionary of Medieval Latin from British Sources*, de l'année 400 à l'année 1450 ; le CODOLVA, de l'année 1200 à l'année 1350.
13. Cet article a été rédigé au sein du projet « Informatización del *Glossarium Mediae Latinitatis Cataloniae* (8) » : « Ampliación y desarrollo de la base de datos *Corpus Documentale Latinum Cataloniae* (2) » (FFI2016-77831-(2-1-P)), financé par le Ministère espagnol de l'Économie.

Références bibliographiques

Textes

Bassols, Mariano et Bastardas, Joan (dir.), *Glossarium Mediae Latinitatis*, t. I (A-D), Barcelona, CSIC/Universidad de Barcelona, 1960-1985.

Bastardas, Joan (dir.), *Glossarium Mediae Latinitatis Cataloniae*, fasc. 11 (F) et fasc. 12 (G), Barcelona, CSIC, 2006.

Font i Rius, Josep Maria, *Cartas de población y franquicia de Cataluña*, 2 t., vol. I: *Estudio. Diplomatario. Presentación monográfico-local e índices*, Madrid/Barcelona, CSIC, 1969.

—, *Cartas de población y franquicia de Cataluña*, t. II, *Estudio. Apéndice al vol. I*, Madrid/Barcelona, CSIC, 1983.

—, « L'escola jurídica de Barcelona », dans Alturo, Jesús, Bellès, Joan, Font i Rius, Josep Maria *et al.*, *Liber Iudicum popularis. Ordenat pel jutge Bonsom de Barcelona*, Barcelona, Generalitat de Catalunya, 2003, t. I, p. 67-100.

Gómez Rabal, Ana, « En torno a las ediciones de la documentación latina catalana altomedieval », *Archivum Latinitatis Medii Aevi (Bulletin Du Cange)*, n° 66, 2008, p. 355-366.

Junyent i Subirà, Eduard, *Diplomatari i escrits literaris de l'abat i bisbe Oliba* (éd. A.M. Mundo), Barcelona, Institut d'Estudis Catalans, 1992.

Mabillon, Jean, *Acta Sanctorum Ordinis Sancti Benedicti in saeculorum classes distributa (Saeculum V)*, Venetiae, 1737, vol. VII, p. 851-860.

Millàs Vallicrosa, José Maria, *Assaig d'història de les idees Fisiques i matemàtiques a la Catalunya medieval*, Barcelona, 1931.

Pérez González, Maurillo (dir.), *Lexicon Latinitatis Medii Aevi Regni Legionis imperfectum (s. VIII-1230)*, Turnhout, Brepols, 2010.

Petrus de Marca, *Marca Hispanica sive limes Hispanicus, hoc est geographica et historica descriptio Cataloniae, Ruscinonis et circumjacentium populorum ab annu 817 ad annum 1258*, Paris, 1688, ap. 222, cd. 1072-82.

Ressources en ligne

Corpus Documentale Latinum Cataloniae (CODOLCAT), Quetglas, Pere J. (dir.), Gómez Rabal, Ana (coord. éd.). En ligne : http://gmlc.imf.csic.es/codolcat (V.S, 2016).

Corpus Documentale Latinum Gallaeciae (CODOLGA), López Pereira, José Eduardo (dir.). En ligne : http://corpus.cirp.es/codolga (v. 12, 2015) [consulté le 18 octobre 2016].

ALMUNIA, Prædii rustici species apud Hispanos. Charta Sanctii Regis Aragonum æræ 1132. apud Martinezium in Hist. Pinnatensi lib. 3. cap. 9 : *Concedo prædicto Cœnobio,... illam meam Almuniam, vocatam Daymus, quæ afrontat ex una parte cum Torredellas, etc.* Alia Alfonsi VI. Regis æræ 1133. apud Anton. de *Yepez* in Chronico Ord. S. Benedicti tom. 6 : *In quo loco incipit alia via, per quam descendunt usque in viam publicam super Almuniam Regis, etc.* Observantiæ Regni Aragon. lib. 5. tit. de Jure dotium, § 4 : *Tamen si Miles vel Infantio habent unam Almuniam vel turrim, censetur una hæreditas cum toto hæreditamento adjuncto illi Almuniæ, vel turri.* Adde Colmenarezium in Hist. Segobiensi cap. 16. § 4.

☞ In his tribus exemplis, excepto forte postremo, per *Almunia* intelligi potest hortus, ut intelligendus est dubio procul in sequentibus ; Testamentum I. Adephonsi Regis Hispaniæ apud Marten. tom. 1. Collect. Ampliss. col. 546. C : *Et ut hi qui in eadem Ecclesia suprascripta permanserint, supplementum aliquod victui habere possint, offero illis villam unam nomine Hukeka, et unam Almuniam, quam nos Latine vocamus Ortum, qui est prope illam Ecclesiam S. Servandi.* Et in Archivo S. Victoris Massiliensis armar. Hispan. n. 115 : *Almunia Regis, quam nos Latine vocamus Ortum.* [✠ Lusitanis *Almuinha*. Conf. S. Rosa de Viterbo Elucidarii tom. 1. pag. 102. et Appendic. pag. 7. A verbo latino *Alimonia* originem trahere hanc vocem scribit et exemplis multis e documentis Lusitan. adductis firmare conatur, ea non tantum *hortum*, sed quodvis prædium haud longe ab urbe situm, significari.]

Fig. 1. *Almunia* : *Glossarium Mediae et Infimae Latinitatis, conditum a Carolo du Fresne, domino Du Cange* ; nouvelle éd. par L. Favre, t. I-X, Niort, 1883-1887

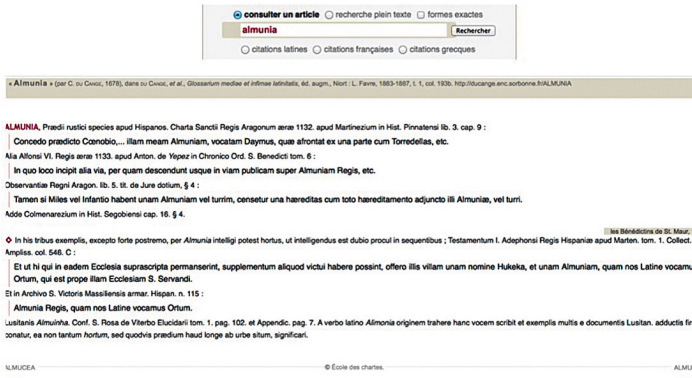
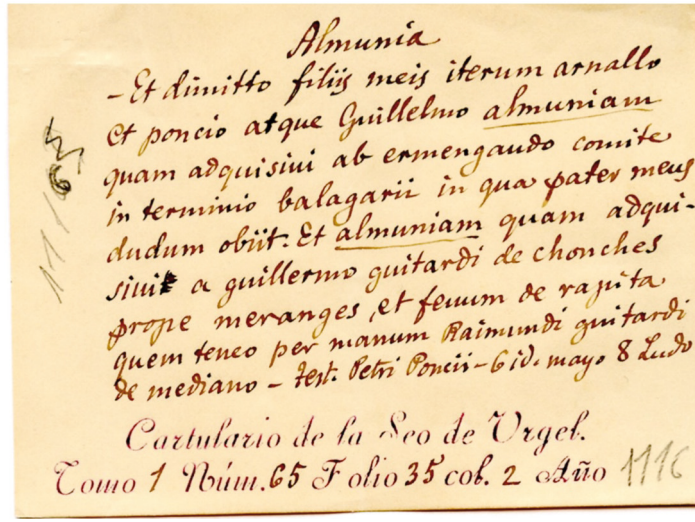
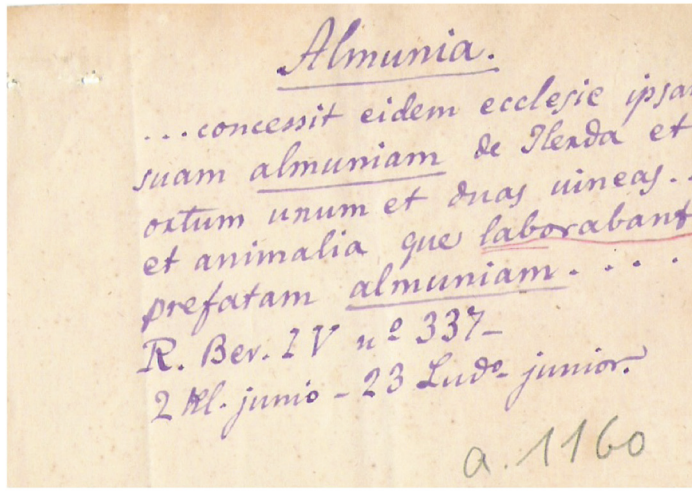


Fig. 2. *Almunia* : Du Cange *et al.*, *Glossarium Mediae et Infimae Latinitatis*, édition numérique, en ligne : http://ducange.enc.sorbonne.fr/almunia



Almunia
– Et dimitto filiis meis iterum arnallo et poncio atque Guillelmo almuniam quam adquisiui ab ermengaudo comite in terminio balagarii in qua pater meus dudum obiit. Et almuniam quam adquisiui a guillermo guitardi de chonches prope meranges, et feuum de rapita quem teneo per manum Raimundi guitardi de mediano – test. Petri Poncii – 6 id. mayo 8 Ludo

Cartulario de la Seo de Urgel.
Tomo 1 Núm. 65 Folio 35 col. 2 Año 1116



Almunia.
...concessit eidem ecclesie ipsam suam almuniam de Ilerda et ortum unum et duas uineas.. et animalia que laborabant prefatam almuniam.....
R. Ber. IV nº 337–
2 Kl. junio – 23 Ludo. junior.
a. 1160

Fig. 3. et 4. *Almunia*, fiches de Josep Balari i Jovany.
Barcelone, Institution Milá y Fontanals, CSIC, bibliothèque de l'équipe du *GMLC*

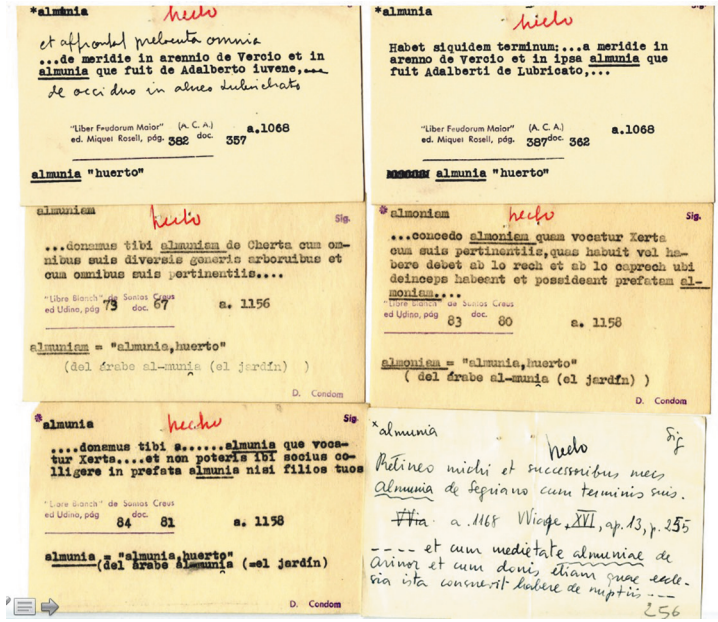


Fig. 5. *Almunia*, fiches du *GMLC*. Barcelone, Institution Milá y Fontanals, CSIC, bibliothèque de l'équipe du *GMLC*

almunia, -mon- [*ár.* munya, *ár. hisp.* '*huerto, jardín vasto*'] *huerta o granja:* **1068** LFeud. I 362, p. 387: habet siquidem terminum . . . a meridie in arenno de Vercio et in ipsa almunia que fuit Adalberti de Lubricato. **1068** LFeud. I 357, p. 382: et affrontat prelocuta omnia . . . de meridie in arennio de Vercio et in almunia que fuit de Adalberto iuuene, de occiduo in alueo Lubricato. **1079** LFeud. I 165, p. 174: et transeunt (*sc.* termini) ab occidente per transuersum per almuniam grossam usque ad terminum comitatus Vrgelli. **1091** CTabernoles 95, f. 50r: I turrem que uocant Bellcarre . . . cum duabus almuniis qui in circuitu eius sunt. Affrontat . . . de meridie in ipsa almunia de Alfiz et de Filela. **1092** CTabernoles 66, f. 36r (VViage XII, ap. 1, p. 211-212): insuper addo ipsam almuniam quae fuit de Iuceph Caualer, quae est ultra collo de Portella in ipso plano . . . et sic uadit per sumitatem de ipsa serra, et ipsas almunias paruulas includit, quae sunt de ipsa almunia iam dicta. **1094** (Urgell) Miret, *Castellbó*, ap. VIII, p. 362: Et addo ad hoc donum omnes meschitas que sunt infra muros ciuitatis Balagarii cum omnibus terris et uineis et hortis et arboribus et tendis et almuniis et omnia que illorum sunt uel esse debent. **1116** ACUrgell, *Cart.* I 65, f. 35, col. 2: dimitto filiis meis . . almuniam quam adquisiui ab Ermengaudo comite in terminio Balagarii in qua pater meus dudum obiit. Et almuniam quam adquisiui a Guillermo Guitardi de Chonches prope Meranges.

1129 ACUrgell, *Cart.* I 80, f. 40, col. 1: et in castrum de Castellione unam almuniam qui uocant de Auimfadida. **1156** CSCreus 67, p. 73: donamus tibi . . . almuniam de Cherta cum omnibus suis diuersi generis arboribus . . . et cum omnibus suis ad se pertinentibus aquis. **1158** CSCreus 80, p. 83: concedo almonia quam uocatur Xerta . . . ab lo rech et ab lo caprech. **1158** CSCreus 81, p. 84: donamus tibi . . . almunia que uocatur Xerta . . . ut deinceps habeas et possideas prefatam almuniam . . . et non poteris ibi socius colligere in prefata almunia nisi filios tuos. **1160** ACA Ramón Berenguer IV, n.º 337: concessit eidem ecclesie ipsam suam almuniam de Ilerda et ortum unum . . . et animalia que laborabant prefatam almuniam. **1168** (Lérida) VViage XVI, ap. 13, p. 255: retineo michi . . . almunia de Segriano. p. 266: cum medietate almuniae de Arinor.

Fig. 6. *Almunia*, première édition du *GMLC*, fascicule 1, *a - aragalius* (1960)

almunia, -mon-

[*ar.* munya, *ar. hisp.*'*hort, jardí vast*' | '*huerto, jardín vasto*' | '*vegetable garden, big garden*']

horta o granja | *huerta o granja* | *vegetable garden or farm:*

1068 LFeud. I 362, p. 387: habet siquidem terminum ... a meridie in arenno de Vercio et in ipsa almunia que fuit Adalberti de Lubricato.

1068 LFeud. I 357, p. 382: et affrontat prelocuta omnia ... de meridie in arennio de Vercio et in almunia que fuit de Adalberto iuuene, de occiduo in alueo Lubricato.

1079 LFeud. I 165, p. 174: et transeunt (*sc.* termini) ab occidente per transuersum per almuniam grossam usque ad terminum comitatus Vrgelli.

1091 CTavèrnoles 95, f. 50^{r} (Soler 47): I turrem que uocant Bellcaire ... cum duabus almuniis qui in circuitu eius sunt. Affrontat ... de meridie in ipsa almunia de Alfiz et de Filela.

1092 CTavèrnoles 66, f. 36^{r}, Soler 49 (VViage XII, ap. 1, pp. 211-212): insuper addo ipsam almuniam quae fuit de Iuceph Caualer, quae est ultra collo de Portella in ipso plano ... et sic uadit per sumitatem de ipsa serra, et ipsas almunias paruulas includit, quae sunt de ipsa almunia iam dicta.

1094 (Urgell) Miret, *Castellbó* 8, p. 362: et addo ad hoc donum omnes meschitas que sunt infra muros ciuitatis Balagarii cum omnibus terris et uineis et hortis et arboribus et tendis et almuniis et omnia que illorum sunt uel esse debent.

1116 ACUrgell, *Cart.* I 65, f. 35, col. 2: dimitto filiis meis ... almuniam quam adquisiui ab Ermengaudo comite in terminio Balagarii in qua pater meus dudum obiit. Et almuniam quam adquisiui a Guillermo Guitardi de Chonches prope Meranges.

1129 ACUrgell, *Cart.* I 80, f. 40, col. 1: et in castrum de Castellione unam almuniam qui uocant de Auimfadida.

1156 CSCreus 67, p. 73: donamus tibi ... almuniam de Cherta cum omnibus suis diuersi generis arboribus ... et cum omnibus suis ad se pertinentibus aquis.

1158 CSCreus 80, p. 83: concedo almonia quam uocatur Xerta ... ab lo rech et ab lo caprech.

1158 CSCreus 81, p. 84: donamus tibi ... almunia que uocatur Xerta ... ut deinceps habeas et possideas prefatam almuniam et non poteris ibi socius colligere in prefata almunia nisi filios tuos.

1160 ACA Ramon Berenguer IV, n. 337: concessit eidem ecclesie ipsam suam almuniam de Ilerda et ortum unum ... et animalia que laborabant prefatam almuniam.

1168 (Lleida) VViage XVI, ap. 13, p. 255: retineo michi ... almunia de Segriano. *ibid., p. 256:* : cum medietate almuniae de Arinor.

Fig. 7. *Almunia*, édition numérique du *GMLC* (prototype)

amarello [*cf. esp.* amarillo *del bajo lat. hisp.* amarellus '*amarillento, pálido*', *dim. de* amarus] [1] *amarillo:* **1069** ACUrgell, *Cart.* I 407, f. 134, col. 1 (Miret, BRABL VI, p. 385): ad ipsa canonica . . . dimisit . . . unum barril de X quinals simul cum ipso superlito de palio que habebat in ipsa sede et I feltro amarello qui ibidem erat.

[1] Amarellus *no ha dejado rastro en el dominio del catalán.*

Fig. 8. *Amarello*, première édition du *GMLC*, fascicule 1, *a - aragalius* (1960)

amarellus, -a [*cf. esp.* amarillo < *lat. med. hisp.* amarellus, *'esgrogueït, pàl·lid'* | *'amarillento, pálido'* | *'yellowish, pale', dim. ab* amarus][1] *groc* | *amarillo* | *yellow:* **924** DipOsona 283, p. 263 (*donació dels comtes Sunyer i Riquilda a l'altar de S. Salvador del monestir de Ripoll* | *donación de los condes Suñer y Riquilda al altar de S. Salvador del monasterio de Ripoll* | *donation made by the Counts Suñer and Riquilda at the altar of St. Salvador in the monastery of Ripoll*): planeta I colore amarella diocodrina, kappa I uermilia ex diorodono. **998** DipGirona 612, p. 518: ad domum Sancti Petro Gallicantu uacas III et uitulos III et somero I et caldaria I et oues XXX, mapes I, tuuallia I, tapitto I et feltro I amarello et capizalo I. **1069** ACUrgell, *Cart.* I 407, f. 134, col. 1 (*ed.* Miret, *Aplech*, p. 385): ad ipsa canonica … dimisit … unum barril de X quinals simul cum ipso superlito de palio que habebat in ipsa sede et I feltro amarello qui ibidem erat. **1068-1071** Sanahuja, *Àger* 27, p. 348: uestimento ... optimo de oztorino amarel.

[1] Amarellus *no ha deixat rastre en el domini del català.*

[1] Amarellus *no ha dejado rastro en el dominio del catalán.*

[1] Amarellus *has left no trace in the Catalan area.*

Fig. 9. *Amarellus*, deuxième édition du *GMLC* (2010)

amarellus, -a

[*cf. esp.* amarillo < *lat. med. hisp.* amarellus , *'esgrogueït, pàl·lid'* | *'amarillento, pálido'* | *'yellowish, pale', dim. ab* amarus[1]]

groc | *amarillo* | *yellow:*

924 DipOsona 283, p. 263 (*donació dels comtes Sunyer i Riquilda a l'altar de S. Salvador del monestir de Ripoll* | *donación de los condes Suñer y Riquilda al altar de S. Salvador del monasterio de Ripoll* | *donation made by the Counts Suñer and Riquilda at the altar of St. Salvador in the monastery of Ripoll*): planeta I colore amarella diocodrina, kappa I uermilia ex diorodono.

998 DipGirona 612, p. 518: ad domum Sancti Petro Gallicantu uacas III et uitulos III et somero I et caldaria I et oues XXX, mapes I, tuuallia I, tapitto I et feltro I amarello et capizalo I.

1069 ACUrgell, *Cart.* I 407, f. 134, col. 1 (*ed.* Miret, *Aplech*, p. 385): ad ipsa canonica ... dimisit ... unum barril de X quinals simul cum ipso superlito de palio que habebat in ipsa sede et I feltro amarello qui ibidem erat.

1068-1071 Sanahuja, *Àger* 27, p. 348: uestimento ... optimo de oztorino amarel.

[1]Amarellus *no ha deixat rastre en el domini del català.*

[1]Amarellus *no ha dejado rastro en el dominio del catalán.*

[1]Amarellus *has left no trace in the Catalan area.*

Fig. 10. *Amarellus*, édition numérique du *GMLC* (prototype)

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

Michèle Goyens & Céline Szecel
KU Leuven

En français moderne, le vocabulaire médical est caractérisé par un type de formation lexicale appelé « composition néoclassique ». Ce procédé a été décrit entre autres par Henri Cottez (1980), Bernard Fradin (2003) et Fiammetta Namer (2007, 2009). Namer le définit comme la réunion de deux (ou plusieurs) éléments issus du grec ou du latin qui n'ont pas d'existence syntaxique autonome. En outre, les règles de la composition néoclassique veulent que l'élément sémantiquement recteur soit placé à droite, alors qu'il est placé à gauche dans la composition standard (Namer, 2009, p. 317-19). Ainsi, le terme *oculo-céphalogyre* est composé du formant *oculo*, du latin *oculus*, désignant l'« œil », d'autre part du formant *céphalo*, du grec *kephalè*, « tête » et enfin de *gyre*, l'élément sémantiquement recteur du composé, provenant du grec *guros*, signifiant « cercle » et par extension « mouvement », pour dénommer ce qui est « relatif aux mouvements de la tête liés à la vision »[1].

Si la composition dite « néoclassique » n'est mise en œuvre qu'à partir des XVII^e^ et XVIII^e^ siècles, à la suite de l'apparition de la chimie et de la physique modernes et du travail de scientifiques tels que Guyton de Morveau et Lavoisier, la formation de la terminologie médicale en français trouve déjà ses origines dans les premières traductions de traités latins en langue vernaculaire, datant du XIII^e^ et surtout du XIV^e^ siècle.

1. En ligne : www.universalis.fr.

Les conditions dans lesquelles ont travaillé les traducteurs de ces textes, ainsi que les méthodologies qu'ils ont appliquées, ont fait l'objet d'un intérêt croissant dans la dernière décennie. Citons, à titre d'exemple, le recueil *Lexiques scientifiques et techniques : Constitution et approche historique*, édité par Olivier Bertrand, Hiltrud Gerner et Béatrice Stumpf (2007) ou, pour la terminologie médicale, les travaux de Sylvie Bazin-Tacchella (2007, notamment), qui examinent les termes anatomiques relevés dans différentes traductions de la *Chirurgia Magna* de Guy de Chauliac ; l'article de Michèle Goyens et Elisabeth Dévière (2007), qui analyse la terminologie des fièvres employée dans la traduction latine et vernaculaire des *Problemata physica* pseudo-aristotéliciens, ou encore l'étude d'Isabelle Vedrenne-Fajolles (2012) sur le lexique pathologique et en particulier les maladies de la peau. Le dictionnaire en ligne *DFSM-CréaLScience* s'inscrit également dans cette optique de recherche, puisqu'il a pour objectif de recenser le vocabulaire technique et scientifique du français médiéval et met à la disposition des chercheurs des applications intéressantes permettant, par exemple, de mettre en relief les réseaux sémantiques qui se tissent entre les termes décrits[2].

Nous souhaitons présenter ici un projet de recherche initié en octobre 2014 et consacré au développement de la terminologie médicale au cours du Moyen Âge, ainsi qu'au sort des termes créés à cette époque, sort que nous désirons lier à des critères d'ordre morphologique[3]. Nous esquisserons tout d'abord les objectifs du projet, et formulerons nos hypothèses de travail avant de présenter notre corpus. Nous définirons ensuite les critères employés pour analyser les néologismes médicaux relevés au sein de notre corpus, que nous ferons suivre de quelques exemples, et nous expliquerons les principes

2. Le dictionnaire *DFSM-CréaLScience* (en ligne : www.crealscience.fr) est un projet en cours de développement, sous la direction de Joëlle Ducos et Xavier-Laurent Salvador.
3. Le projet *Latin Authority and Constructional Transparency at Work: Neologisms in the French Medical Vocabulary of the Middle Ages and their Fate* est subventionné par le Fonds de la recherche de la KU Leuven (OT/14/047). Direction : Michèle Goyens (KUL) ; co-direction : Kristel Van Goethem (UC Louvain).

du cadre théorique utilisé en vue de dégager, une fois l'analyse morphologique effectuée, des corrélations au niveau des structures morphologiques relevées. Nous terminerons en présentant au lecteur une série de perspectives.

Objectifs et hypothèses de travail

Le projet de recherche que nous présentons ici a pour objectif spécifique d'étudier les raisons pour lesquelles certains néologismes médicaux créés au cours du Moyen Âge se maintiennent jusqu'en français moderne, alors que d'autres disparaissent après un certain laps de temps.

Diverses études, entre autres celle menée par Joëlle Ducos (1998), qui concerne la terminologie météorologique en français au Moyen Âge et compare les pratiques des traducteurs Mahieu le Vilain, Evrart de Conty et Jean Corbechon, ainsi que celle proposée par Thomas Städtler (2007) et consacrée aux néologismes utilisés par Nicole Oresme, montrent que les créations « indigènes » ont plus de mal à se maintenir que des néologismes qui s'inspirent du latin. Comparons par exemple *toye qui resamble au grain du noir roisin* (Evrart de Conty, *Livre des Problemes de Aristote*, IX, 2, fol. 152r30) à *uvee* (Evrart de Conty, *id.*, XXXI, 2, fol. 201r9), emprunté au latin *uvea*. Seul ce dernier terme se maintient. Une étude préliminaire sur un corpus restreint de quatre textes avait permis de dégager les indices suivants : tout d'abord, une série de facteurs externes semble jouer un rôle important dans la lexicalisation de certains termes, comme le succès du texte, ou le prestige de l'auteur. Par ailleurs, des facteurs internes à la langue paraissent également décisifs. Ainsi, parmi les néologismes médicaux créés au cours du Moyen Âge, 77 % sont des emprunts au latin, ou au grec *via* le latin, tandis que 23 % seulement sont des créations « indigènes » (Goyens et Van Tricht, 2015).

Ces observations nous mènent à l'hypothèse de travail suivante : des critères morphologiques, et plus particulièrement la transparence constructionnelle, jouent un rôle crucial pour la préservation des néologismes. Autrement dit, 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 dérivations ou des compositions réalisées à partir de bases morphologiques françaises.

Par ailleurs, deux arguments renforcent cette hypothèse. Tout d'abord, dans le contexte de la communication scientifique médiévale, le latin est la langue de référence, à côté de l'arabe et du grec, et dès lors la langue dominante (Lusignan, 1989). Il faut de plus rappeler le caractère transparent de la morphologie lexicale latine, les morphèmes lexicaux (bases et affixes) restant stables quelles que soient les combinaisons. En guise d'illustration, l'élément *noc-* reste stable dans *noc-ere* « nuire », *noc-ibilis* « nuisible » ainsi que *noc-ivus* « nuisible, dangereux » ; la structure du lexique français est plus opaque : considérons par exemple *eau*, *aquatique*, *évier*, issus de la racine latine *aqu-(a)* « eau », bien que cette dernière ne transparaisse plus dans deux des formes citées (Goyens, 2013).

Nous développerons dans cette contribution l'analyse de quelques termes relevés dans notre corpus, ceux issus de la racine latine *flegm-*, qui parvient à se maintenir sous cette forme (dans *flegmatique* par ex.), mais pas sous sa forme *fleum-* (mfr. *fleume*, *fleumatique*, etc.) ; et nous présenterons notre base de données de termes médicaux (latins et français) et des morphèmes qui les composent (bases, affixes, latins et français).

Corpus

Notre étude, empirique, se fonde sur des néologismes[4] relevés dans un corpus de textes médicaux du Moyen Âge, datant du XIII^e^, XIV^e^ et XV^e^ siècle et composé d'une part de traductions du latin, d'autre part de textes immédiatement composés en français. Les textes du corpus sont présentés ci-dessous, chronologiquement, dans la mesure du possible, et accompagnés de la mention de leur auteur, de leur titre et de la période de leur rédaction. Pour les traductions, nous mentionnons l'auteur du texte source, le titre de la traduction et le nom du traducteur. Nous ajoutons à la

4. Sur la problématique concernant la datation de la première attestation d'un néologisme, voir Goyens (2013, p. 48-49) et Goyens et Van Tricht (2015, p. 392-393).

fin de chaque référence l'édition qui a été retenue, ou l'indication qu'il s'agit d'une transcription réalisée à partir d'un manuscrit ou d'un incunable[5].

Textes composés en français

Aldebrandin de Sienne, *Regime du corps*, 1256-1257 (éd. Landouzy-Pépin, 1911)

Anonyme, *Novele cirurgerie*, XIII^e siècle (éd. Hieatt-Jones, 1990)

Jean Pitard, *Receptaire*, *ca.* 1300 (ms.)

Traductions

Anonyme, *Médicinaire liégeois* (traducteur anonyme), XIII^e siècle (éd. Haust, 1941)

Roger Frugard, *Chirurgie* (traducteur anonyme), XIII^e siècle (éd. Hunt, 1994 ; Valls, 1995)

Albucasis, *Traitier de Cyrurgie* (traducteur anonyme), milieu du XIII^e siècle (éd. Trotter, 2005)

Nicolas de Salerne, *Antidotaire Nicolas* (traducteur anonyme), fin du XIII^e siècle (éd. Dorveaux, 1896)

(Pseudo)-Aristote, *Secré de Secrez* (traducteur : Pierre d'Abernon), peu après 1267 (éd. Beckerlegge, 1944)

(Pseudo)-Aristote, *Secret des segrez* (traducteur : Jofroi de Waterford), *ca.* 1300 (éd. Schauwecker, 2007)

Henri de Mondeville, *Chirurgie* (traducteur anonyme), 1314 (éd. Bos, 1897-1898)

Anonyme, *Consultation de la faculté de médecine de Paris de 1348* (adaptation A, traducteur anonyme), 1349-1350 (éd. Bazin-Tacchella, 2001)

Anonyme, *Consultation sur la peste* (adaptation B, traducteur anonyme), 1349-1350 (éd. Bazin-Tacchella, 2001)

5. Nous avons eu recours à des éditions existantes, mais lorsque celles-ci font défaut, nous nous basons sur des transcriptions réalisées par diverses personnes, notamment Françoise Guichard-Tesson, Ildiko Van Tricht et Sylvie Bazin-Tacchella, que nous remercions chaleureusement. Pour les références complètes des éditions, manuscrits et incunables utilisés, voir la bibliographie (1. *Corpus*).

Hippocrate, *Amphorismes Ypocras* (traducteur et commentateur : Martin de Saint-Gille), 1362-1363 (éd. du prologue Jacquart, 1997 ; section 2, V et VI Lafeuille, 1954 ; Lafeuille, 1964 ; reste du texte transcrit d'après le ms.).

Barthélemy l'Anglais, *Le Livre de propriétés des choses* (traducteur : Jean Corbechon), *ca.* 1372 (Livre V : ms. ; Livre VII : éd. Louis, 2001).

Bernard de Gordon, *Fleur de lys* (traducteur anonyme), 1377 (plusieurs versions ; mss. et incunable : Livre I, Livre II, 1-8).

(Pseudo)-Aristote, *Livre des Problemes de Aristote* (traducteur : Evrart de Conty), *ca.* 1380 (éd. sous la dir. de Françoise Guichard-Tesson et Michèle Goyens).

Anonyme, *Congnoissance des corps humains* (traducteur : Nicole Saoul), 1396 (ms.).

Anonyme, *Poeme sur la grand peste de 1348* (traducteur : Olivier de la Haye), 1426 (éd. Guigue, 1888).

Guy de Chauliac, *Grande Chirurgie* (traducteur anonyme), 3e quart / 2e tiers du XVe siècle (éd. traité I : Tittel, 2004 ; traité II, doctr. 2, chap. 5 : Bazin-Tacchella, 2001).

Arnauld de Villeneuve, *Le Regime tresutile et tresproufitable pour conserver et garder la santé du corps humain* (traducteur anonyme), 1480 (éd. Cummins, 1976).

Guillaume de Salicet, *Chirurgie* (traducteur : Nicole Prevost), 1492 (incunable).

Anonyme, *Médicinaire namurois* (traducteur anonyme), XVe siècle (éd. Haust, 1941).

Anonyme, *Anathomie mise en disputacions* (traducteur anonyme), *ca.* XVe siècle (ms.).

Bienvenu Raffe, *Le compendil pour la douleur et maladie des yeux* (traducteur anonyme), XVe siècle (ms.).

Nicolas de Salerne, *Antidotaire Nicolas* (traducteur anonyme), version du XVe siècle (éd. Dorveaux, 1896).

Nous réalisons un corpus électronique et lemmatisé de ces textes en collaboration avec l'équipe du *Dictionnaire du moyen français*, qui met ses outils à notre disposition, nous permettant d'établir des glossaires et des listes de fréquences. Ces données seront complétées à l'aide du *DMF* 2015 (ou d'une version actualisée) ainsi que des lemmes tirés du *Dictionnaire du français scientifique médiéval* de CréaLScience, du point de vue des informations sémantiques notamment.

Les critères d'analyse

Les néologismes médicaux sont relevés dans le corpus, puis analysés systématiquement selon des critères internes ou externes, d'ordre aussi bien général que morphologique, et qui forment la grille d'analyse pour une base de données électronique morphologique.

Facteurs externes

Nous analysons d'abord les facteurs externes qui peuvent influencer le maintien d'un néologisme, et sont sélectionnés dans une étude préalable réalisée par Michèle Goyens et Ildiko Van Tricht (2015). Ainsi, nous vérifions le succès de chaque texte, s'exprimant par le nombre de ses manuscrits ou de ses éditions. Par ailleurs, il faut également prendre en compte le prestige de l'auteur ou du traducteur ; en d'autres termes, une étude de la réception du texte s'impose. Nous devons, en outre, examiner si l'auteur utilise des gloses explicatives en introduisant un néologisme, et s'il est systématique dans son emploi. Enfin, nous étudions le sort connu par le néologisme jusqu'en français moderne, *a priori* à partir des instruments lexicographiques existants, afin d'observer s'il s'est lexicalisé.

Les critères internes d'ordre général

En ce qui concerne les critères internes d'ordre général, nous identifions en premier lieu l'étymon du terme concerné. Nous indiquons ensuite de quel type de néologisme il s'agit, *i.e.* d'un emprunt (formel et/ou sémantique) ou d'une création

indigène (formelle ou sémantique)[6]. Si le terme est un emprunt, nous mentionnons la langue source dont il est issu. En revanche, dans le cas où il s'agit d'une création indigène, il faut préciser si elle est le résultat d'une dérivation, d'une composition ou d'un procédé sémantique. Nous donnons également le sens du lexème du moyen français jusqu'au français moderne en consultant les instruments lexicographiques appropriés[7], ainsi que la signification de sa première attestation dans le corpus. Par ailleurs, nous indiquons le champ sémantique de la médecine médiévale auquel le terme appartient, à savoir l'anatomie, la physiologie, la pathologie ou la thérapeutique.

Les critères internes d'ordre morphologique

Un deuxième type de critères internes est d'ordre morphologique. Ces derniers ont été sélectionnés notamment d'après des études psycholinguistiques consacrées à la productivité de constructions morphologiques en français moderne (voir par ex. Dal, 2003[8]), qui en ont révélé la pertinence. Concrètement, nous décomposons chaque terme en sa base[9] et son (ou ses) affixe(s), dont nous étudions les allomorphies éventuelles[10]. Nous précisons également la taille du lexème, que nous exprimons en nombre de syllabes. En outre, nous analysons la distance du lexème (la base et l'affixe) par rapport à son étymon, exprimée en termes de phonèmes distincts. En

6. En réalité, la typologie des néologismes est plus détaillée. Nous renvoyons à la recherche qu'Ildiko Van Tricht présente dans sa thèse de doctorat (2015), *La Science en texte et contexte : la terminologie médicale française utilisée dans les* Problèmes *d'Evrart de Conty à la lumière du discours médical médiéval*, réalisée dans le cadre du projet de recherche OT/10/23 (financé par la KU Leuven).
7. Lorsque le sens est toujours attesté en français moderne, celui des époques intermédiaires n'est plus vérifié.
8. Pour une description plus détaillée, voir Goyens (2013, p. 49-52).
9. Nous ne faisons pas la distinction entre *base* et *radical* tels que les entend Denis Apothéloz (2002), pour des raisons de commodité. Pour Apothéloz, « on appelle base l'élément sur lequel opère un affixe » (2002, p. 15) et « on appelle radical le morphème lexical qui subsiste quand tous les affixes dérivationnels et flexionnels ont été enlevés » (2002, p. 16).
10. Nous ne retenons pas dans cette rubrique les morphèmes grammaticaux ou flexionnels, mais uniquement les affixes dérivationnels, sauf pour le calcul de la « distance » du lexème par rapport à l'étymon.

cinquième lieu, nous étudions la productivité du lexème, de sa base et de ses affixes, en faisant bien la distinction entre la fréquence du type (nombre d'attestations du même type, éventuellement avec une orthographe différente) et la fréquence du signe (*token* : nombre d'attestations du même signe, avec la même orthographe). Enfin, nous indiquons si le terme appartient à une famille morphologique, dont nous précisons la taille, exprimée par le nombre de lexèmes de cette famille. Nous en étudions alors la fréquence cumulée, c'est-à-dire l'ensemble des fréquences de chaque élément faisant partie de cette famille morphologique.

À l'issue de l'analyse de chaque lexème d'après ces critères d'ordre externe et interne, les résultats sont entrés dans une banque de données électronique morphologique, qui sera mise à la disposition de la communauté scientifique.

Quelques exemples

Dans ce qui suit, nous avons analysé des exemples concrets à l'aide des critères internes d'ordre général et morphologique évoqués plus haut. Nous avons effectué les recherches à partir d'un terme, *flegme* (et sa base *flegm-*), afin de sélectionner tous les vocables appartenant à sa famille morphologique tels qu'ils apparaissent dans notre corpus. Il s'agit des lexèmes *flegme*, *flegmatique*, *flegmasie* et *flegmon*.

Nous représentons ces données dans un tableau par terme. Les critères analysés sont mentionnés dans la colonne de gauche, les résultats dans la colonne de droite.

Tableau 1. Analyse du lexème *flegme*

Critères internes	*Flegme*
Étymon (selon les dictionnaires)	bas latin *phlegma/flegma*[11] (TLFi, DLD)
Type de néologisme	Emprunt
Sens au fil du temps :	
– de l'emprunt	– en moyen français : « Lymphe, flegme ; pituite » (*DMF*) – première attestation dans le corpus : « une des 4 humeurs de l'ancienne médecine » (TLFi) (ML 69r)[12] – en français moderne : « Vieux. Humeur glaireuse, liquide épais. » « Au figuré, courant, au singulier. Caractère d'une personne calme et imperturbable, qui garde son sang-froid en toutes circonstances. » (TLFi)
– de la base	– le sens de *flegm-* reste stable pour le terme *flegme* (« humeur, liquide »), mais reçoit un sens figuré supplémentaire en français moderne (« tempérament d'une personne calme »)
Domaine de la médecine médiévale	Physiologie
Analyse morphologique	Lexème simple
Allomorphie de la base Allomorphie de l'affixe	*flegm-*, *flem-*, *fleugm-*, *fleum-*[13] /[14]
Taille du lexème : Nombre de syllabes	2
« Distance » de la base / de l'affixe par rapport à son étymon : nombre de phonèmes différents	– base *flegm-* : 0 – *-e vs. -a* : 1
Fréquence du type *vs.* fréquence du signe :	
– du lexème	– du lexème : type 74 / signe 74
– de la base	– de la base : type 139 / signe 139
– de l'affixe	– de l'affixe : /
– fait partie d'une famille morphol.	– oui
– fréquence « cumulée » de la base	– base : type 139 / signe 139
– taille de la famille morphologique	– 4 lexèmes : *flegme, flegmatique, flegmasie, flegmon*

11. D'après le TLFi, il s'agirait d'une réfection de l'ancien français *fleume* sur le modèle du bas latin *phlegma* « humeur, mucus ». Nous préférons plutôt considérer *flegme* comme un emprunt.

Tableau 2. Analyse du lexème *flegmatique*

Critères internes	*Flegmatique*
Étymon (selon les dictionnaires)	bas latin *phlegmaticus* / *flegmaticus* (TLFi, DLD)
Type de néologisme	Emprunt
Sens au fil du temps :	
– de l'emprunt	– en moyen français : « Qui abonde en flegme, en lymphe, qui a les propriétés du flegme, de la lymphe, qui a trait au flegme, à la lymphe » (*DMF*) – première attestation dans le corpus : « qui abonde en flegme » (PB I, 1, 3v40)[15] – en français moderne : « Médical. Qui abonde en flegme, en lymphe » (TLFi)
– de la base	– le sens de *flegm-* reste stable pour le terme *flegmatique* (« qui abonde en flegme »)
– de l'affixe	– suffixe *-(at)ique*[16] issu du latin *-(a)-(t)-icus* « relatif à, qui est propre à », « formateur de très nombreux adjectifs épicènes parfois employés comme substantifs, appartenant notamment au vocabulaire scientifique et technique » (TLFi)[17]
Domaine de la médecine médiévale	Physiologie

12. Le sigle ML renvoie au *Médicinaire liégeois*, datant du XIII^e^ siècle (éd. Haust, 1941). Se reporter à la bibliographie (cf. *infra*, dans les références bibliographiques : *Éditions modernes*). Contexte (169 r) : « C'est a la *flegme*. Alle *flegme* assonlee el ventre, prens ravene, manjust asseis z aigue boive chaude. »
13. Chaque forme fait l'objet d'une entrée séparée dans la base de données. En outre, les formes avec la base *flem-*, *fleugm-*, *fleum-* ne sont pas des emprunts, mais des cas intermédiaires.
14. *-e* n'est pas un affixe dérivationnel, mais flexionnel.
15. Le sigle PB renvoie au *Livre des problemes de Aristote* d'Evrart de Conty (*ca*. 1380) ; voir bibliographie (cf. *infra*, dans les références bibliographiques : *Éditions modernes* pour la partie I, et *Manuscrits* pour les autres parties). Contexte : « Et selonc ce sont quatre complexions composees : …, la tierce qui excede en froidure et en moisteur, qui est appelee *flegmatique*, …, et est pour la raison des.4. humours dont je parlerai cy aprés, … »
16. Nous ne savons pas encore comment traiter certains éléments, comme par exemple -a(t)- dans *flegmatique* ou -as- dans *flegmasie*. Nous hésitons à les considérer comme faisant partie intégrante de la base. Provisoirement, nous mettons ces éléments entre parenthèses.
17. Afin d'effectuer l'analyse sémantique de cet affixe, il faudra attendre le dépouillement systématique du corpus. Provisoirement, nous avons consulté le TLFi à ce propos.

Critères internes	*Flegmatique*
Analyse morphologique	Lexème complexe Suffixation en latin : base *phlegm-* affixe *-(a)-ticus*
Allomorphie de la base Allomorphie de l'affixe	*flegm-*, *flem-*, *fleugm-*, *fleum-* Voir corpus[18]
Taille du lexème : Nombre de syllabes	4
« Distance » de la base / de l'affixe par rapport à son étymon : nombre de phonèmes différents	– base *flegm-* : 0 – affixe *-(a)-tique vs. -(a)-ticus* : 2
Fréquence du type *vs.* fréquence du signe :	
– du lexème	– du lexème : type 55 / signe *flegmatique* 26 ; signe *flegmatiques* 19
– de la base	– de la base : type 139 / signe 139
– de l'affixe	– de l'affixe : attendre le dépouillement du corpus
– fait partie d'une famille morphol.	– oui
– fréquence « cumulée » de la base	– base : type 139 / signe 139
– taille de la famille morphologique	– 4 lexèmes : *flegme, flegmatique, flegmasie, flegmon*

Tableau 3. Analyse du lexème *flegmasie*

Critères internes	*Flegmasie*
Étymon (selon les dictionnaires)	grec *phlegmasia* (TLFi)
Type de néologisme	Emprunt
Sens au fil du temps :	
– de l'emprunt	– en moyen français : « Inflammation interne » (*DMF*) – première attestation dans le corpus : « Inflammation interne » (PB I, 6, 12v40)[19] – en français moderne : « Pathologie. Vieilli. Inflammation interne. » (TLFi)

18. Afin de relever les éventuelles allomorphies de cet affixe, il faudra attendre le dépouillement systématique du corpus.
19. Contexte : « Et pource, quant les humidités et les *flegmasies* se assamblent u cors en aucuns lieus pour la disposition du tans, elles poeent lors faire apostumes et pluseurs maladies perilleuses. »

Critères internes	*Flegmasie*
– de la base	– le sens de *flegm-* reste stable pour le terme *flegmasie* (« inflammation interne »)
– de l'affixe	– suffixe *-ie* entrant dans la construction de substantifs féminins à partir de noms de personnes ou d'adjectifs substantivables. [Le dérivé désigne un état pathologique.] (TLFi)[20]
Domaine de la médecine médiévale	physiologie
Analyse morphologique	Lexème complexe Suffixation en grec : base *phlegm-* affixe *-(as)-ia*
Allomorphie de la base Allomorphie de l'affixe	*flegm-*, *flem-*, *fleugm-*, *fleum-* Attendre le dépouillement du corpus
Taille du lexème : Nombre de syllabes	4[21]
« Distance » de la base / de l'affixe par rapport à son étymon : nombre de phonèmes différents	– base *flegm-* : 0 – affixe *-(as)-ie* : 1
Fréquence du type *vs.* fréquence du signe :	
– du lexème	– du lexème : type 6 / signe *flegmasie* 1 ; signe *flegmasies* 5
– de la base	– de la base : type 139 / signe 139
– de l'affixe	– de l'affixe : attendre le dépouillement du corpus
– fait partie d'une famille morphol.	– oui
– fréquence « cumulée » de la base	– base : type 139 / signe 139
– taille de la famille morphologique	– 4 lexèmes : *flegme, flegmatique, flegmasie, flegmon*

20. Afin d'effectuer l'analyse sémantique de cet affixe, il faudra attendre le dépouillement systématique du corpus. Provisoirement, nous avons consulté ici aussi le TLFi à ce propos.

21. *Flegmasie* est sans doute composé de 4 syllabes, car le *-e* final était un *e* central se conservant sans changement au début du moyen français, mais se labialisant en *e* moyen au XV^e siècle (Joly, 1995, p. 65). Il est donc fort probable que ce *-e* final était prononcé.

Tableau 4. Analyse du lexème *flegmon*

Critères internes	*Flegmon*
Étymon (selon les dictionnaires)	latin *phlegmone, -es* (TLFi) ou latin *phlegmon, -onis* (Gaffiot)
Type de néologisme	Emprunt ou xénisme[22]
Sens au fil du temps :	
– de l'emprunt	– en moyen français : « Tuméfaction, ou plus précisément apostume sanguin » (*DMF*) – première attestation dans le corpus : « apostume de sang » (FL I, 8, 49[23]) – en français moderne : « Pathologie. Inflammation aiguë des tissus conjonctifs, pouvant évoluer vers la formation d'un abcès » ; « Synonyme de abcès » (TLFi)
– de la base	– le sens de *flegm-* reste stable pour le terme *flegmon* (« inflammation, apostume »)

22. Deroy (1956, p. 224) considère les xénismes comme des « mots sentis comme étrangers et en quelque sorte cités », caractérisés par leur forme étrangère. Pour Sablayrolles (2000, p. 234), le terme *xénisme* renvoie à des « emprunts tels quels », des « alloglottes », ressentis comme étrangers. Cet auteur distingue, à côté des xénismes, les pérégrinismes, qui sont des xénismes dont la forme a été francisée, et où l'aspect étranger est parfois même gommé. Nous nous rallions plutôt à la définition de Dubois *et al.* (2007², p. 512), qui décrivent le xénisme comme non intégré à la langue, un « mot étranger, mentionné avec référence au code linguistique d'origine et aux réalités étrangères » ; selon ce dictionnaire, le pérégrinisme renvoie alors à ce mot, mais dépourvu de marques métalinguistiques et utilisé occasionnellement dans la langue. Concrètement, dans notre corpus, il n'est pas toujours aisé de déterminer si *flegmon* est un xénisme, ne devant donc pas être qualifié de néologisme à part entière, ou un emprunt au latin, dont les finales ont été francisées et qui fait donc partie du lexique français. Dans l'exemple suivant, le lexème *flegmon* serait un xénisme, car il est clairement mentionné que c'est en médecine que l'on emploie ce terme, et le latin était la langue de référence dans ce domaine :
« Nous devons dont savoir que apostume n'est autre chose que une tumeur ou une enfleüre innaturele, sourvenant u cors humain, si comme Galiens dit, et sont ou poeent estre de quatre manieres, selonc les.4. humours, car elle se poet faire de sanc, et lors est elle appelee en medecine *flegmon*, et s'elle se fait de cole, elle est appelee herisipile, se de flegme, dismia, et s'elle se fait de melancolie, elle est appelee communement cancer. » (Evrart de Conty, *Livre des Problemes de Aristote*, I, 44, fol. 45r15)
En revanche, *flegmon* doit être considéré comme un terme emprunté et donc intégré au français dans l'extrait suivant, car il n'y est pas fait référence à une langue étrangère ou de spécialité ; le terme porte le morphème flexionnel du pluriel *–s* et est en outre accompagné de l'article défini *les* : « Caulis sont choux il sont chaulx et secz il maturent les *flegmons* et duresses et consolide et prohibe… » (Nicole Prevost, *La Cirurgie de maistre Guillaume de Salicet*, V, 10)

23. Le sigle FL renvoie à l'incunable de 1495 du *Fleur de lys* de Bernard de Gordon. Cet ouvrage date de 1377 ; voir bibliographie (*cf.* 6.1.3. *Incunables*). Contexte : « Aposteme qui sont d'humeurs se c'est de sang on les appelle *flegmon*, se c'est de cole herisipile, se de fleume zumma, se de melancolie chancre ou scliros. »

Critères internes	*Flegmon*
– de l'affixe	– voir corpus[24]
Domaine de la médecine médiévale	physiologie
Analyse morphologique	Lexème complexe Suffixation en latin (*phlegmone*) : base *phlegm-* affixe *-o-ne* Suffixation en latin (*phlegmon*) : base *phlegm-* affixe *-on*
Allomorphie de la base Allomorphie de l'affixe	*flegm-*, *flem-*, *fleugm-*, *fleum-* Attendre le dépouillement du corpus
Taille du lexème : Nombre de syllabes	2
« Distance » de la base / de l'affixe par rapport à son étymon : nombre de phonèmes différents	– base *flegm-* : 0 – affixe *-on* : 1 (par rapport à *-on* en latin) ou 2 (par rapport à *-one* en latin)
Fréquence du type *vs.* fréquence du signe :	
– du lexème	– du lexème : type 3 ou 4[24] / signe *flegmon* 2 ou 3 ; signe *flegmons* 1
– de la base	– de la base : type 139 / signe 139
– de l'affixe	– de l'affixe : voir corpus[25]
– fait partie d'une famille morphol.	– oui
– fréquence « cumulée » de la base	– base : type 139 / signe 139
– taille de la famille morphologique	– 4 lexèmes : *flegme, flegmatique, flegmasie, flegmon*

24. Afin d'effectuer l'analyse sémantique de cet affixe, il faudra attendre le dépouillement systématique du corpus. Le TLFi ne donne pas d'information à propos de cet affixe. Dans son dictionnaire du vocabulaire savant, Henri Cottez (1980, p. 284) mentionne deux suffixes *-on*, utilisés en physique pour former les noms de gaz rares ou des noms de particules élémentaires. Or, il est clair qu'il s'agit ici d'un terme médical et non d'un terme utilisé en physique.
25. La fréquence du type du lexème varie selon que nous considérons *flegmon* comme un xénisme ou un emprunt. Voir note 21 pour plus de détails.
26. Afin de connaître la fréquence de cet affixe (type et signe), il faudra attendre le dépouillement systématique du corpus.

Cette analyse devra être appliquée systématiquement à tous les néologismes médicaux de notre corpus afin que des tendances générales puissent être dégagées en ce qui concerne le maintien de certains termes (cf. *infra*, partie *Perspectives*).

Nous avons été confrontées à une difficulté dans l'analyse de la formation du lexème : une mise au point s'impose en effet concernant le statut d'emprunt ou de dérivé. Ainsi, *flegmatique* pourrait être considéré comme dérivé de la base empruntée *flegm-* à l'aide du suffixe *-(a)-tique*. Cependant, en latin classique et médiéval, le dérivé *flegmaticus* est attesté, et il nous semble dès lors plus probable, au vu de la situation médiévale, de l'analyser comme un emprunt. En d'autres termes, pour nous, un vocable français pouvant être analysé soit comme un dérivé français à partir d'une base empruntée (au latin), soit comme l'emprunt d'un dérivé (latin), doit toujours être traité comme un emprunt[27] (Goyens, 2013, p. 49 ; Goyens et Van Tricht, 2015, p. 393).

En outre, nous devons en principe nous baser sur la prononciation afin de déterminer si deux formes doivent être analysées comme distinctes, analyse que complique le caractère fluctuant de la graphie en moyen français. Il faudra donc faire la distinction entre les graphèmes rendant un même son et ceux représentant des sons vraiment différents. *Fleumasie*, par exemple, est une forme héréditaire d'après nos analyses, soit l'évolution phonétique régulière d'une forme latine. Ce terme ne correspond pas à *flegmasie*, le graphème <g> représentant le son /g/, à l'inverse du digraphe <eu> qui rend le son /ew/ ou /ø/, d'autant plus que le second terme est plus proche formellement du lexème latin *flegmasia*. Dans notre corpus ont aussi été relevés *fleugmon* et *fleugmatiser*, qu'il faut également considérer comme des formes à part entière.

27. Nous l'avons déjà expliqué ailleurs : d'après les morphologues, ce choix reste sujet à discussion, puisque l'on ne peut pas trancher. Cependant, il nous semble plus probable que le traducteur préfère emprunter un dérivé tout fait plutôt que de créer une nouvelle dérivation à partir d'un emprunt (Goyens et Van Tricht, 2015, p. 393).

L'interprétation de certaines graphies peut elle aussi poser problème. La terminaison <eus>, par exemple, représente-t-elle la terminaison latine [eus] ou plutôt la forme française [øs] ? Voici deux exemples ambigus tirés de notre corpus :

> Je dis doncques premierement que les humeurs des yeulx sont trois en nombre, desquelles la premiere est appellee *albugineus*, la seconde est appelee cristallinus, et la tierce vitreus. (Bienvenu Raffe, *Compendil pour la douleur et maladie des yeulx*, 41v.)

> Lequel humeur *albugineus* yseroit hors du pertuis du pannicule uveal si non qu'il fust couvert et ainsi a il esté necessaire de faire aultre pannicule, lequel se appelle corneus, ainsi nommé pour la resemblance qu'il a avec une corne clere et lucide, lequel est engendré du pannicule sclirotique, et lye avec le pannicule sclirotique tout l'euyl. (Nicole Prevost, *La Cirurgie de maistre Guillaume de Salicet*, IV, 1.)

Dans le premier exemple, la prononciation d'*albugineus* pourrait être la prononciation latine, puisque le terme est précédé du participe *appelee*, et qu'il fait en outre partie d'une énumération de termes dont l'un est clairement latin, à savoir *cristallinus*. Dans le second exemple en revanche, le terme est intégré dans une phrase française et ne s'accompagne d'aucun élément indiquant que son emploi serait « latin ». En l'absence de telles indications, nous considérons sa prononciation comme française.

Le cadre théorique

Afin d'étudier le sort des néologismes à partir de critères d'ordre morphologique, nous soumettrons les lexèmes analysés selon la grille décrite ci-dessus à une étude menée selon le cadre théorique de la morphologie des constructions, telle qu'elle est développée par Geert Booij (2010[28]).

Selon cette théorie les lexèmes complexes, aussi bien les dérivations que les compositions, peuvent également être considérés comme des constructions au niveau de la lexie, donc comme des constructions morphologiques. Celles-ci peuvent

28. Ce cadre théorique s'inscrit dans les travaux de la grammaire de construction (*Construction Grammar*), qui conçoit la langue comme un réseau de constructions, composé d'associations de formes et de sens à tous niveaux.

être analysées suivant un modèle hiérarchique, caractérisé par des niveaux d'abstraction de plus en plus poussés. Ces analyses mèneront à la création de familles de lexèmes et de schémas abstraits formant un réseau morphologique, d'après leur structure et leur fonction morphologique interne.

D'après nos hypothèses, ce sont les familles de lexèmes révélant des corrélations forme-sens systématiques, transparentes et productives qui contribuent au renforcement et à la survie d'une terminologie. Ces corrélations peuvent être illustrées à l'aide des arbres morphologiques suivants. Ainsi, l'arbre construit à partir de la base *flegm-* est productif et révèle un nombre relativement élevé de termes, dont la racine reste stable jusqu'en français moderne :

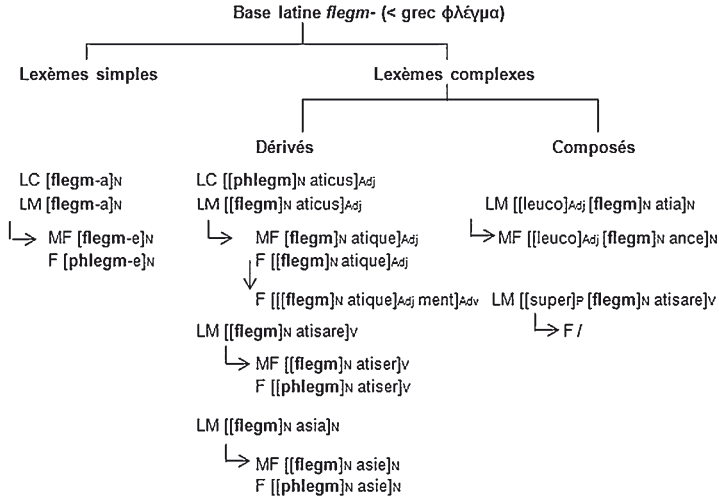


Fig. 1. Arbre morphologique à partir de la base *flegm-*[29]

29. Nous trouvons tantôt l'initiale *f-*, tantôt l'initiale *ph-* en latin et en français. Nous optons pour l'orthographe en *f-*, sauf si l'orthographe en *ph-* est la seule attestée. Les abréviations utilisées sont les suivantes : LC (latin classique), LM (latin médiéval), MF (moyen français), F (français moderne) ; N (nom), Adj (adjectif), V (verbe), P (préposition).

Après qu'un nombre maximal de termes médicaux auront été examinés, des analyses ultérieures permettront de dégager des relations et des schémas plus abstraits.

En revanche, l'arbre formé à partir de la racine *fleum-* reste stérile et n'aboutit pas à un réseau, probablement parce qu'il s'éloigne du latin, bien que cette hypothèse reste à confirmer :

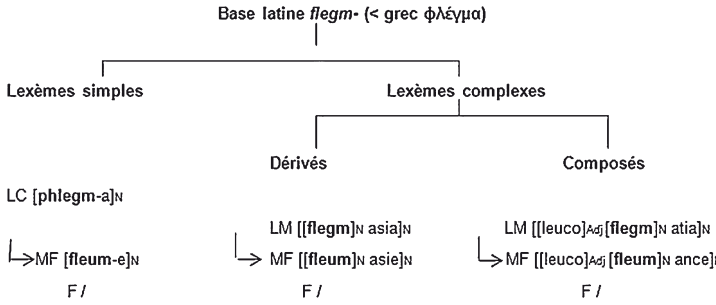


Fig. 2. Arbre morphologique à partir de la base *flegm-* évoluant en *fleum-*

Perspectives

Une fois la banque de données complétée pour les différents critères de l'analyse, les résultats seront soumis à des analyses statistiques multivariées, afin de révéler si certains facteurs (ou certaines combinaisons de facteurs) sont responsables de la longévité du terme, confirmant peut-être notre hypothèse de départ, c'est-à-dire que les termes présentant une relation formelle proche de l'élément latin dont ils sont issus se maintiennent mieux que des créations françaises indigènes.

Nous ne pouvons donc malheureusement pas encore, à ce stade du projet, présenter des résultats concrets issus de notre étude. Cependant, nous avons voulu montrer en quoi consiste concrètement ce programme de recherche, qui est véritablement novateur. Il l'est d'une part grâce au corpus de textes médicaux informatisé sur lequel il s'appuie, unique en son genre, et qui peut inspirer d'autres chercheurs dans la perspective de la constitution de corpus semblables. D'autre part, la réalisation

d'une banque de données morphologique de termes médicaux français et de leurs contreparties latines (ou grecques) permettra des recherches diverses sur des formes latines, françaises, des bases, des affixes, etc. Cette banque de données sera accessible aux chercheurs et est susceptible d'avoir un impact important sur la description du développement de la terminologie scientifique française. Par ailleurs, le recours au cadre théorique de la morphologie des constructions offre de nouvelles perspectives, puisqu'il s'agit d'un modèle rarement utilisé pour des études diachroniques, et jamais pour le Moyen Âge. Ainsi, le projet permettra peut-être d'apporter des compléments à la théorie elle-même.

Enfin nous espérons que cette analyse, décomposant les termes médicaux forgés au cours du Moyen Âge et comparant systématiquement les morphèmes ainsi dégagés avec les éléments latins correspondants, permettra de confirmer nos hypothèses concernant la lexicalisation des néologismes et les premières traces du mécanisme de la composition néoclassique caractérisant le vocabulaire moderne des sciences.

Références bibliographiques

Corpus

Éditions modernes

BECKERLEGGE, Oliver A., *« Le Secré de Secrez » by Pierre d'Abernun of Fetcham*, Oxford, Basil Blackwell, 1944.

BOS, Alphonse, « *La Chirurgie de maître Henri de Mondeville* ». *Traduction contemporaine de l'auteur, publiée d'après le ms. unique de la Bibliothèque nationale*, 2 vol., Paris, Firmin Didot, 1897-1898.

BAZIN-TACCHELLA, Sylvie, « Rupture et continuité du discours médical à travers les écrits sur la peste de 1348. Le *Compendium de epidemia* (1348) et ses adaptations françaises. La relation de peste contenue dans la *Chirurgia Magna* de Guy de Chauliac (1363) », dans BAZIN-TACCHELLA, Sylvie *et al.* (dir.), *Airs, miasmes et contagion. Les épidémies dans l'Antiquité et au Moyen Âge*, Langres, Dominique Guéniot, 2001, p. 105-156 [éd. de la *Consultation de la faculté de médecine de Paris de 1348 (adaptation A)* aux p. 132-145 ; éd. de la *Consultation sur la peste (adaptation B)* aux p. 146-153].

CUMMINS, Patricia, *A Critical Edition of « Le regime tresutile et tresproufitable pour conserver et garder la santé du corps humain »: With the Commentary of Arnoul de Villeneuve, Corrected by the « Docteurs Regens » of Montpellier: 1480, Lyon: 1491*, Chapel Hill, U.N.C. Department of Romance Languages, 1976.

DORVEAUX, Paul, *L'Antidotaire Nicolas*, Paris, H. Welter, 1896.

GUICHARD-TESSON, Françoise et GOYENS, Michèle, avec la collaboration de DUMAS, Geneviève (dir.), *Le « Livre des Problemes de Aristote » par Evrart de Conty, Livre I*, Paris, Honoré Champion, à paraître.

GUIGUE, Georges, *Olivier de la Haye, « Poëme sur la Grande Peste de 1348 », publié d'après le ms. de la bibliothèque du Palais Saint-Pierre*, Lyon, Georg, 1888.

HAUST, Jean, *Médicinaire liégeois du XIII^e siècle et médicinaire namurois du XV^e (Mss. 815 et 2769 de Darmstadt)*, Bruxelles, Palais des Académies, 1941.

HIEATT, Constance Bartlett et JONES, Robin F. (éd.), *La Novele cirurgerie*, London, Anglo-Norman Text Society, 1990.

HUNT, Tony, *Anglo-Norman Medicine*, t. I, *Roger Frugard's « Chirurgia » and the « Practica Brevis » of Platearius*, Cambridge, D. S. Brewer, 1994 [éd. du ms. Cambridge, Trinity College 0.1.20, fol. 24va-30rb, qui est un fragment ou un abrégé d'une version en ancien français].

JACQUART, Danielle, « Hippocrate en français. Le Livre des amphorismes de Martin de Saint-Gille (1362-1363) », dans JACQUART, Danielle (dir.), *Les Voies de la science grecque. Études sur la transmission des textes de l'Antiquité au XIX^e siècle*, Genève, Droz, 1997, p. 241-327 [éd. du prologue des *Amphorismes Ypocras* de Martin de Saint-Gille aux p. 293-327].

LAFEUILLE, Germaine (éd.), *Les « Amphorismes Ypocras » de Martin de Saint-Gille*, Genève, Droz, 1954.

LAFEUILLE, Germaine (éd.), *Les Commentaires de Martin de Saint-Gille sur les « Amphorismes Ypocras »*, Genève, Droz, 1964.

LANDOUZY, Louis et PÉPIN, Roger, *Le Régime du corps de maître Aldebrandin de Sienne*, Paris, Honoré Champion, 1911.

LOUIS, Sylvain, *Édition critique du livre VII de la traduction par Jean Corbechon du « De proprietatibus rerum » de Barthélémi l'Anglais*, thèse de doctorat, Rouen, Université de Rouen, 2001.

TITTEL, Sabine, *Die « Anathomie » in der « Grande Chirurgie » des Gui de Chauliac. Wort- und sachgeschichtliche Untersuchungen und Edition*, Tübingen, Niemeyer, 2004.

SCHAUWECKER, Yela, *Die Diätetik nach dem « Secretum Secretorum » in der altfranzösischen Version von Jofroi de Waterford. Teiledition und lexikalische Untersuchung*, Würzburg, Königshausen und Neumann, 2007.

TROTTER, David, *Albucasis : « Traitier de Cyrurgie ». Édition de la traduction en ancien français de la Chirurgie d'Abū'l Qāsim*

Halaf Ibn 'Abbās al-Zahrāwī du ms. BnF, fr. 1318, Tübingen, Niemeyer, 2005.

VALLS, Helen Elizabeth, *Studies on Roger Frugardi's Chirurgia*, thèse de doctorat, Toronto, Centre for medieval studies, 1995 [éd. de la version complète du texte du ms. London, BL, Sloane 1977, 10ra-46ra].

Manuscrits

Anonyme, *Anathomie mise en disputacions*, trad. anonyme, *ca*. XV^e^ siècle. Paris, BnF, fr. 19994, fol. 39r-57v., transcription réalisée par I. Van Tricht.

JEAN PITARD, *Receptaire, ca*. 1300. Paris, BnF, fr. 12323, transcription réalisée par S. Bazin-Tacchella.

(Pseudo)-ARISTOTE, *Livre des Problemes de Aristote,* Livres II-XXXVIII (trad. Evrart de Conty), *ca*. 1380. Paris, BnF, fr. 24281 (parties I-XV1) et fr. 24282 (parties XV2-XXXVIII), transcriptions sous la direction de F. Guichard-Tesson.

Anonyme, *Congnoissance des corps humains* (trad. Nicole Saoul), 1396. Paris, BnF, fr. 1317, fol. 5ra-5ova, transcription réalisée par I. Van Tricht.

HIPPOCRATE, *Amphorismes Ypocras* (trad. et comm. Martin de Saint-Gille), 1362-1363. Paris, BnF, fr. 24246, fol. 1or-fin, transcription réalisée par I. Van Tricht.

BARTHÉLEMY L'ANGLAIS, *Le Livre de propriétés des choses*, Livre V (trad. Jean Corbechon), *ca.* 1372. Paris, BnF, fr. 16993, fol. 39 vb.-73vb, transcription réalisée par I. Van Tricht.

BERNARD DE GORDON, *Fleur de lys* (trad. anonyme), 1377. Paris, BnF, fr. 1288, fol. 136rb-140vb (*Veci ung petit tractié qui parle dez passions… selon maistre Bernard de Gourdon*) et fr. 1327, fol. 1r-37r (*Compillacion faicte par Bernard de Gourdon*), transcriptions réalisées par I. Van Tricht.

BIENVENU RAFFE, *Le compendil pour la douleur et maladie des yeux*, trad. anonyme, XV^e^ siècle. Paris, Bnf, fr. 1327, fol. 38r-60v, transcription réalisée par I. Van Tricht.

Incunables

Bernard de Gordon, *Fleur de lys en medecine*, Livre I, Livre II, 1-8, Lyon, 1495, transcription réalisée par I. Van Tricht.

Nicole Prevost, *La Cirurgie de maistre Guillaume de Salicet traduite du latin par Nicole Prevost*, Lyon, Matthias Huss, 1492, transcription réalisée par I. Van Tricht.

Bibliographie générale

Apothéloz, Denis, *La Construction du lexique français. Principes de morphologie dérivationnelle*, Paris, Ophrys, 2002.

Bazin-Tacchella, Sylvie, « Constitution d'un lexique anatomique en français aux XV^e^ et XVI^e^ siècles : L'exemple des noms des intestins et des os dans les traductions françaises de la *Chirurgia Magna* de Guy de Chauliac », dans Bertrand, Olivier *et al.*, *Lexiques scientifiques et techniques. Constitution et approche historique*, Palaiseau, Éditions de l'École polytechnique, 2007, p. 65-80.

Bertrand, Olivier, Gerner, Hiltrud et Stumpf, Béatrice (dir.), *Lexiques scientifiques et techniques. Constitution et approche historique*, Palaiseau, Éditions de l'École polytechnique, 2007.

Booij, Geert, *Construction Morphology,* Oxford, Oxford University Press, 2010.

Cottez, Henri, *Dictionnaire des structures du vocabulaire savant. Éléments et modèles de formation*, Paris, Le Robert, 1980.

Dal, Georgette (dir.), *La Productivité morphologique en questions et en expérimentations*, Paris, Larousse, 2003.

DMF 2015 = *Dictionnaire du moyen français*, version 2015. ATILF/CNRS - Université de Lorraine. En ligne : www.atilf.fr/dmf [consulté le 30 avril 2015].

DLD 01/08/2014 = *Database of Latin Dictionaries*, Turnhout, Brepols. En ligne : clt.brepolis.net/dld/ [consulté le 30 avril 2015].

Deroy, Louis, *L'emprunt linguistique*, Paris, Les Belles Lettres, 1956.

Dubois, Jacques *et al.*, *Grand dictionnaire, Linguistique & Sciences du langage* [1994], Paris, Larousse, 2007.

DUCOS, Joëlle, *La Météorologie en français au Moyen Âge (XIII^e-XIV^e siècle)*, Paris, Honoré Champion, 1998.

Encyclopaedia Universalis. En ligne : www.universalis.fr [consulté le 30 avril 2015].

FRADIN, Bernard, *Nouvelles approches en morphologie*, Paris, PUF, 2003.

GAFFIOT, Félix, *Le Grand Gaffiot. Dictionnaire latin-français*, nouvelle édition revue et augmentée sous la dir. de Pierre Flobert, Paris, Hachette, 2000.

GOYENS, Michèle, « Le sort des néologismes dans la langue des sciences au Moyen Âge : une question de morphologie ? », *Neologica*, n° 7, 2013, p. 41-56.

GOYENS, Michèle et DÉVIÈRE, Elizabeth, « Le développement du vocabulaire médical en latin et moyen français dans les traductions médiévales des *Problemata* d'Aristote », dans GALDERISI, Claudio et PIGNATELLI, Cinzia (dir.), *The Medieval Translator. Traduire au Moyen Âge*, 11, *La Traduction vers le moyen français*, Turnhout, Brepols, 2007, p. 259-281.

GOYENS, Michèle et VAN TRICHT, Ildiko, « Albathe face à pustule. Disparition *versus* lexicalisation des néologismes en français médiéval », dans BADIOU-MONFERRAN, Claire et VERJANS, Thomas (dir.), *Disparitions. Contributions à l'étude du changement linguistique*, Paris, Honoré Champion, 2015, p. 389-405.

JOLY, Geneviève, *Précis de phonétique historique du français*, Paris, Armand Colin, 1995.

LUSIGNAN, Serge, « La topique de la *translatio studii* et les traductions françaises de textes savants au XIV^e siècle », dans CONTAMINE, Geneviève (dir.), *Traduction et traducteurs au Moyen Âge*, Paris, Éditions du CNRS, 1989, p. 303-315.

NAMER, Fiammetta, « Composition néoclassique : est-on dans l'"hétéromorphosémie" ? », dans HATHOUT, Nabil et MONTERMINI, Fabio (dir.), *Morphologie à Toulouse. Actes du colloque international de Morphologie 4^es Décembrettes*, München, Lincom Europa, 2007, p. 185-206.

Namer, Fiammetta, *Morphologie, lexique et traitement automatique des langues*. *L'analyseur DériF*, Paris, Lavoisier, 2009.

Sablayrolles, Jean-François, *La néologie en français contemporain. Examen du concept et analyse de productions néologiques récentes*, Paris, Champion, 2000.

Städtler, Thomas, « Le traducteur, créateur de néologismes : le cas de Nicole Oresme », dans Bertrand, Olivier *et al.* (dir.), *Lexiques scientifiques et techniques. Constitution et approche historique*, Palaiseau, Éditions de l'École polytechnique, 2007, p. 47-61.

TLFi : *Trésor de la langue française informatisé*. En ligne : atilf.atilf.fr [consulté le 30 avril 2015].

Vedrenne-Fajolles, Isabelle, « Les Pratiques linguistiques des médecins, auteurs, traducteurs ou copistes de traités médicaux. L'exemple des maladies de peau (XII^e^-XV^e^ siècles) », dans Ducos, Joëlle (dir.), *Sciences et langues au Moyen Âge*, Heidelberg, Universitätsverlag Winter, 2012, p. 173-244.

Base de français médiéval : une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique[1]

Céline Guillot, Serge Heiden & Alexei Lavrentiev
UMR ICAR – ENS de Lyon / Université de Lyon /
CNRS – LabEx ASLAN

Les mutations induites par le développement du numérique dans le champ des sciences du langage ont eu une répercussion très directe ces dernières années sur la linguistique diachronique, tout spécialement dans le domaine français. Par son objet d'étude – les états de langue passés pour lesquels nous ne disposons pas de locuteurs ni de compétence linguistique –, la linguistique diachronique s'appuie depuis toujours sur des corpus de données attestées. Mais l'essor récent des ressources numériques a considérablement renouvelé les méthodologies d'analyse, les résultats produits par la recherche et parfois aussi les phénomènes étudiés. Ces évolutions en cours tendent à renforcer l'attitude réflexive du diachronicien, nécessairement confronté à l'altérité des données qu'il décrit. Et par certains côtés, les questions nouvellement posées par l'essor du numérique rejoignent ce qui était au centre de l'approche philologique traditionnelle.

Nous illustrerons quelques aspects de ces mutations récentes en nous appuyant sur un corpus numérique à l'usage

1. Les auteurs remercient le LabEx ASLAN (ANR-10-LABX-0081) de l'université de Lyon pour son soutien financier dans le cadre du programme « Investissements d'avenir » (ANR-11-IDEX-0007) de l'État français, géré par l'Agence nationale de la recherche (ANR).

des linguistes médiévistes, la Base de français médiéval (BFM[2]). Il s'agit d'un corpus numérique déjà relativement ancien (initié en 1989) s'appuyant sur une pratique numérique en évolution constante, composé d'éditions de référence (éditions originales et éditions imprimées numérisées) encodées au format XML-TEI et enrichies à de multiples niveaux (métadonnées textuelles, codage interne, segmentation en mots et annotation linguistique). Nous donnerons un aperçu des possibilités nouvelles offertes à l'analyse par ce corpus outillé (section 1), qui motivent la mise en place d'une double chaîne, philologique pour la constitution et la préparation des données textuelles (section 2) et analytique pour leur exploitation outillée (section 3). À travers l'exemple de la BFM, nous tenterons de dégager les contraintes et apports d'un tel cadre méthodologique dans une perspective plus large et plus communautaire.

Nouvelles avancées méthodologiques dans le domaine de la linguistique diachronique de corpus

Depuis sa création, la Base de français médiéval a été conçue comme un outil dédié à l'étude linguistique historique et diachronique du français. Actuellement exploitée par une communauté internationale de 400 utilisateurs environ, elle a depuis ses origines été le support de nombreuses thèses et travaux de recherche portant sur la langue médiévale. Elle est également utilisée de manière constante par l'équipe en charge de son développement au sein du laboratoire ICAR et de l'ENS de Lyon. Les travaux de recherche menés dans ce cadre alimentent et infléchissent les évolutions de la base. Bien qu'elles portent sur des sujets très variés (de l'évolution de la ponctuation médiévale, de la sémantique des démonstratifs, de l'oral représenté ou des incises en français, pour ne citer que les plus récentes), ces recherches ont pour caractéristique commune de s'appuyer toujours sur les méthodologies définies dans le cadre de la linguistique de corpus. Elles motivent

2. Le site internet du projet BFM (en ligne : http://bfm.ens-lyon.fr) présente la base dans son état actuel et ses objectifs de recherche.

et dirigent l'implémentation dans la base de ressources textuelles et logicielles dont le développement s'effectue de manière parallèle et très étroitement inter-reliée (définition de métadonnées textuelles qui s'articulent aux fonctionnalités de création de corpus/sous-corpus et de contrastes, modèles/outils d'annotation et textes annotés, etc.).

Les ressources numériques ainsi produites permettent de développer des analyses basées sur une démarche empirique, fondée sur des données authentiques et quantifiables, dont les résultats sont reproductibles et vérifiables. Les outils utilisés par l'équipe, qui relèvent de l'approche dite « textométrique » (Lebart et Salem, 1994, en ligne : http://textometrie.ens-lyon.fr), permettent l'analyse quantitative des phénomènes étudiés sans jamais disjoindre les données de leur contexte d'occurrence et des éléments nécessaires à l'interprétation qualitative des résultats.

Une étude récente (Guillot *et al.*, 2015) portant sur les caractéristiques de l'oral représenté a permis, par exemple, d'utiliser le calcul statistique de l'analyse factorielle des correspondances (AFC) pour mettre en évidence les spécificités très fortes, stables et durables, qui caractérisent le discours direct au Moyen Âge.

Pour cette étude, le balisage numérique des segments au discours direct dans tous les textes de la base a permis de réaliser un calcul d'AFC comparant les fréquences des étiquettes morphosyntaxiques associées aux mots du discours direct à celles des autres parties de chaque texte pour produire une visualisation graphique à deux dimensions positionnant chaque plan de texte (discours direct/parties narratives de chaque texte) en fonction des différences observées. Le plan factoriel montre clairement que l'opposition discours direct/parties narratives constitue un contraste dominant à l'intérieur des textes de la base, puisque les cercles (parties au discours direct) et les triangles (parties narratives) se positionnent d'eux-mêmes de part et d'autre de l'espace correspondant à cette opposition quels que soient les textes. Nous avons tracé une diagonale

séparatrice pour mettre en évidence cette distribution. Le retour aux données qui sont à l'origine de la construction du graphique permet d'interpréter la position originale des parties au discours direct du *Comput* de Philippe de Thaon (représenté par le cercle situé en haut à droite de la **fig. 1**). Cette position est liée à l'usage très particulier des guillemets dans ce texte : ils n'indiquent pas les segments au discours direct mais servent à citer des mots isolés. Les guillemets étant les marques formelles sur lesquelles a reposé l'encodage du discours direct et sa dissociation des autres parties de textes dans la base, l'usage déviant de ces marques dans le *Comput* explique son positionnement excentrique dans le graphique[3].

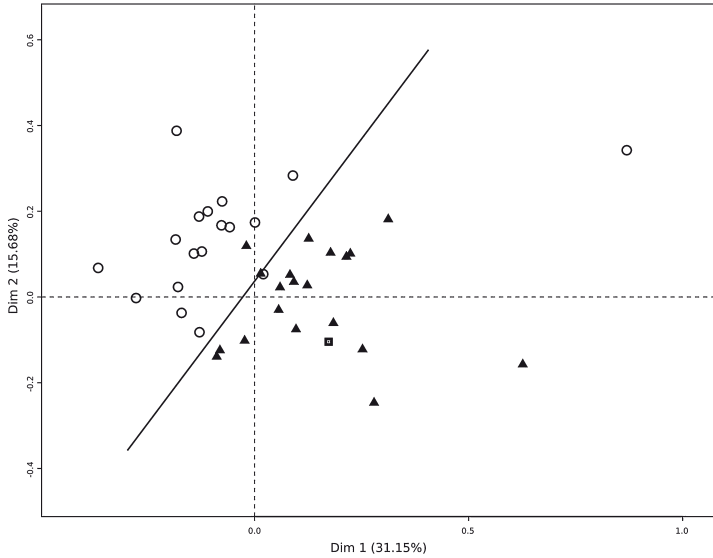


Fig. 1. AFC des textes du corpus, en distinguant les parties narratives *vs.* au discours direct. Pour le calcul, les textes sont modélisés par la fréquence des catégories morphosyntaxiques qu'ils utilisent. Les parties au discours direct sont représentées par des cercles, et les parties narratives par des triangles.

3. Le balisage du discours direct a été réalisé de manière semi-automatique dans tous les textes de la base. Il s'appuie sur les marques graphiques (guillemets ouvrants et fermants) insérées par les éditeurs modernes des textes médiévaux. Les limites de ce balisage automatique sont évidentes. Il permet néanmoins de dégager des tendances générales grâce à l'analyse d'un grand nombre de textes.

L'analyse des étiquettes morphosyntaxiques qui fondent la position relative de chaque point selon le premier axe[4] permet de se faire une idée assez précise des éléments les plus spécifiques au discours direct ou aux autres parties de textes. La fréquence élevée des pronoms personnels, conjonctions de subordination, négations, pronoms impersonnels, interjections et adverbes caractérise le discours direct, celle des noms propres, articles définis, contractions de l'article et des prépositions (*du*, *au*, etc.), noms communs, déterminants cardinaux et participes présents distingue tout ce qui lui est extérieur.

Les études qui sont menées dans un tel cadre d'analyse reposent sur des ressources riches et adaptées. Elles supposent la possibilité d'exploiter les données textuelles avec des outils numériques de synthèse et de recherche. Elles impliquent le traitement d'un volume de données suffisant et d'une diversité assez représentative pour qu'on parvienne à des résultats stables et généraux. Elles imposent aussi un équipement numérique relativement poussé des textes : encodage du discours direct, étiquetage morphosyntaxique de tous les mots, description des unités textuelles grâce à un système de métadonnées permettant l'interprétation des résultats et l'étude de la variation.

Une telle méthode de recherche, qui s'élabore dans un cadre de plus en plus expérimental, vise également à permettre à la communauté scientifique de reproduire les mêmes analyses, grâce à la diffusion, la documentation et la pérennisation des ressources. Cette méthode favorise l'enrichissement continu des textes numériques au fil des analyses linguistiques, les informations rendues disponibles par l'analyse ayant vocation à être associées aux données elles-mêmes pour pouvoir être réutilisées lors de recherches ultérieures. Dans le cas de l'étude citée ci-dessus par exemple, l'analyse des spécificités de l'oral représenté dans la Base de français médiéval amène à réinterpréter la fonction des guillemets dans le texte du *Comput*

4. Pour des raisons de lisibilité nous avons choisi de ne pas faire figurer ces étiquettes sur le graphique, mais l'outil donne directement accès aux informations qui sous-tendent la position des points du graphique.

et à revoir le balisage des séquences au discours direct dans ce texte.

L'amélioration continue des données et le coût inhérent à la préparation et à l'équipement numérique des textes rendent par ailleurs de plus en plus nécessaire le développement partagé et communautaire des ressources. Aux plans juridique et pratique, il est devenu indispensable de permettre la libre circulation et la rediffusion responsable de ces ressources parmi l'ensemble des partenaires qui participent, pour une part variable et à différents niveaux, à leur production et leur exploitation. Le respect de normes et standards internationaux de représentation numérique des textes, l'utilisation de licences de (re)diffusion ouvertes compatibles avec les juridictions et la jurisprudence internationales sont les principaux instruments de cette politique d'échanges communautaires. Nous essaierons de montrer, dans la suite de cet article, les implications très concrètes de cette méthode de travail concernant les aspects philologiques des textes (section 2) comme les outils d'analyse qui permettent de les exploiter (section 3).

Chaîne philologique ouverte pour l'établissement et l'annotation des textes de la BFM

Principes méthodologiques

Les recherches qui sont menées dans le cadre de la linguistique de corpus et que nous avons illustrées ci-dessus par l'étude des spécificités du discours direct en français médiéval exploitent le plus souvent un volume important de données textuelles. Elles permettent surtout d'appliquer les outils informatiques à l'analyse de données de type et de niveau très variés. Certaines de ces informations concernent les unités textuelles dans leur ensemble (métadonnées textuelles), d'autres sont internes aux textes ou à des parties de textes (structures textuelles, comme les passages au discours direct, les groupes de mots correspondant à des unités inférieures, les mots du texte, les caractères, etc.).

Le standard de balisage XML-TEI[5] permet d'encoder toutes ces informations aux niveaux qui leur correspondent (en-têtes pour les métadonnées textuelles, corps du texte pour tout le reste).

La méthodologie de corpus implique par conséquent que l'on identifie et traite séparément au moins trois types d'information : (i) les métadonnées textuelles, qui servent à caractériser les textes, à les regrouper ou à les dissocier, à interpréter les variations observées grâce à l'approche contrastive ; (ii) les structures internes ou unités linguistiques sur lesquelles portent les analyses et qui demandent à être clairement délimitées et balisées (dans l'exemple cité, les segments au discours direct, les limites de mots) ; (iii) les propriétés associées à ces unités linguistiques destinées à être mobilisées lors de l'analyse (étiquettes morphosyntaxiques, lemmes, annotations syntaxiques, etc.). C'est de la combinaison de ces informations multiples que dépendent l'interprétation des résultats et la richesse des analyses.

Lorsqu'il s'agit de corpus de textes médiévaux (ou plus généralement de ceux dont l'édition demande un travail philologique important), des questions méthodologiques supplémentaires doivent par ailleurs être résolues ou en tout cas prises en compte. Il convient en premier lieu de distinguer les sources primaires (les manuscrits, pour l'époque médiévale) des sources secondaires (éditions scientifiques). Il n'est pas envisageable de constituer de grands corpus de textes médiévaux directement à partir des sources primaires, car une bonne transcription de manuscrit est un travail philologique très laborieux qui comprend notamment l'étude de la tradition manuscrite et le choix d'un manuscrit de base, l'identification des erreurs scribales éventuelles, la résolution des nombreuses ambiguïtés des graphies médiévales (telles que les séries de jambages, agglutinations ou abréviations non univoques) et éventuellement la consultation d'autres manuscrits de la même œuvre afin d'éclaircir les passages difficiles. Un tel investissement

5. En ligne : http://www.tei-c.org.

est discutable si une bonne édition scientifique existe déjà pour un texte. Mais l'utilisation des éditions scientifiques comme source de données pour les corpus numériques pose, d'un autre côté, des problèmes importants.

Ré-ingénierie numérique d'éditions scientifiques existantes

On observe d'abord que les pratiques d'établissement du texte varient considérablement d'une tradition philologique à l'autre ; elles évoluent avec le temps, et dépendent dans une mesure non négligeable des choix personnels de l'éditeur. Même si, dans le domaine de l'édition de textes en français médiéval, la tradition « bédiériste[6] » (qui consiste à respecter autant que possible le manuscrit de base) est largement dominante, le degré de « liberté » que les philologues se donnent dans la correction du manuscrit originel est très variable. Ainsi, May Plouzeau (1994) a démontré que la dernière version de l'édition de *La Mort Artu* par Jean Frappier (1964) ne constituait plus une source de données linguistiques fiable.

Certains aspects de l'établissement de texte sont laissés entièrement à l'appréciation de l'éditeur scientifique. Il s'agit en particulier de la ponctuation et de la segmentation des locutions qui se sont figées et sont devenues des lexies uniques au gré de l'évolution de la langue (par ex. la locution prépositionnelle *par mi*, le groupe adverbial *ja mais* ou le syntagme nominal *bon heur*[7]). Cette hétérogénéité des pratiques pose des problèmes évidents pour l'annotation morphosyntaxique et la lemmatisation du corpus, ainsi que pour les recherches et l'analyse des données textuelles.

Une solution partielle aux problèmes posés par la diversité des pratiques philologiques repose sur la normalisation de l'encodage des textes grâce à l'application des recommandations du consortium TEI (*Text Encoding Initiative*). On peut ainsi neutraliser les différentes manières d'indiquer les mêmes types

6. Nommée ainsi en l'honneur de Joseph Bédier, qui en a formulé les principes dans son étude de la tradition manuscrite du *Lai de l'ombre* (1928).
7. Le processus inverse est possible, mais beaucoup plus rare : par exemple le préfixe *tres-* (du latin *trans-*) devenu l'adverbe *très*.

d'interventions éditoriales. Par exemple, les fragments restitués par l'éditeur scientifique à la place des lacunes peuvent être signalés, selon les éditions, par des crochets ou par des chevrons (plus rarement). La balise TEI <supplied> peut être utilisée dans les deux cas. Un autre exemple concerne l'indication des passages au discours direct. C'est toujours l'éditeur scientifique qui place les guillemets, car les manuscrits médiévaux n'utilisaient pas cette marque graphique dans cette fonction. En revanche, selon les traditions philologiques et les règles typographiques adoptées dans différents pays, les guillemets peuvent être français (« ») ou anglais (“ ”), être ou ne pas être fermés devant les incises ou entre les prises de parole dans les dialogues. La balise TEI <q> permet d'harmoniser toutes ces pratiques hétérogènes. Enfin, le balisage des mots du texte peut permettre de dissocier la segmentation visuelle réalisée à l'aide des blancs typographiques de la segmentation analytique utilisée dans l'annotation linguistique et dans les requêtes appliquées au corpus. On peut ainsi procéder à une normalisation massive tout en respectant les choix de l'éditeur dans la présentation graphique. L'harmonisation de la segmentation graphique étant cependant une tâche particulièrement lourde, elle n'a pas encore été réalisée dans le corpus de la BFM[8].

Une seconde source de difficulté est liée au fait que l'état de la propriété intellectuelle de nombreuses éditions n'est pas clair. En France (à la différence de l'Allemagne et de l'Italie, par exemple), il n'existe pas de texte législatif spécifique concernant les éditions critiques (Margoni et Perry, 2011). Si on considère ces éditions comme des œuvres originales créées par les éditeurs scientifiques, les droits patrimoniaux restent protégés pendant soixante-dix ans après la mort de l'éditeur. Certaines maisons d'édition prétendent détenir les droits de diffusion numérique des textes dont les éditeurs scientifiques sont décédés depuis plusieurs décennies. La recherche d'éventuels ayants droit

8. En effet, les annotations syntaxiques réalisées sur certains textes de la BFM dans le cadre du projet SRCMF (Stein et Prévost, 2013) reposent sur la segmentation actuelle des mots. Or, toute modification des choix de segmentation lexicale suppose de réaligner ces annotations sur les nouvelles unités lexicales qui pourraient être créées.

de ces éditions s'avère souvent très longue et complexe. Les contrats d'édition récents prévoient généralement la cession exclusive des droits de diffusion numérique de toutes sortes à la maison d'édition, ce qui les rend inutilisables dans des corpus numériques, pour lesquels la libre diffusion des données est vitale (Guerreau, 2015). Même si la maison d'édition donne son accord pour l'intégration de « son » texte dans un corpus, elle peut le retirer à tout moment, ce qui risque de nuire à la reproductibilité et à la continuité des recherches basées sur ces données. Pour ne plus faire courir ce risque à ses utilisateurs, la Base de français médiéval a été contrainte de retirer un certain nombre de textes (plus d'un million d'occurrences mots au total) en août 2014 à la suite de la rupture d'une convention avec une maison d'édition.

Selon un autre point de vue, défendu récemment par l'une des parties dans un procès opposant deux maisons d'éditions, le « corps » du texte d'une édition scientifique (à l'exclusion de l'introduction, des notes, des variantes et des annexes de toutes sortes) n'est pas une création de l'éditeur scientifique au sens où l'entend le Code de la propriété intellectuelle, et n'est donc pas protégeable. Un jugement de première instance a confirmé cette position, mais la controverse est loin d'être close dans ce débat juridique. Par ailleurs, les notes du texte peuvent comporter des informations très importantes et nécessaires à son analyse (comme l'indication de variantes ou la justification d'une correction).

L'objectif de la Base de français médiéval étant d'offrir à la communauté des chercheurs la ressource la plus riche et la plus fiable possible pour étudier la langue française des premiers textes à la fin du XV^e^ siècle, de multiples facteurs sont pris en compte lors de la sélection des textes à intégrer au corpus. Les œuvres sont d'abord sélectionnées en fonction de leur intérêt linguistique (on cherche à équilibrer le corpus sur le plan diachronique en tenant compte des genres et domaines textuels). La qualité philologique des éditions et leur statut juridique (qui peuvent être facteurs d'exclusion) sont ensuite

évalués. Dans le cas d'éditions récentes dont les fichiers de saisie sous un logiciel de traitement de texte sont disponibles et dont les auteurs n'ont pas cédé l'exclusivité des droits à une maison d'édition, la BFM négocie directement avec les éditeurs scientifiques pour obtenir ces fichiers sources et pouvoir les diffuser sous une licence libre. Des éditions plus anciennes sont numérisées aux frais de l'équipe de la BFM, avec l'accord des ayants droit, lorsqu'on les trouve.

Création d'éditions numériques originales

Tous les problèmes liés à la réutilisation d'éditions traditionnelles peuvent être résolus dans des éditions « nativement numériques ». Il est possible, notamment, de fournir plusieurs niveaux de transcription dont chacun est adapté à des usages et à des catégories de lecteurs différents (Guillot *et al.*, 2017 ; Marchello-Nizia *et al.*, 2015).

Dans la pratique, il nous semble qu'une représentation à deux niveaux, qu'on peut qualifier de « normalisée » et « diplomatique », peut satisfaire la grande majorité des utilisateurs. Le niveau normalisé se rapproche de la tradition de l'édition des textes littéraires, avec toutefois l'application de règles plus explicites concernant notamment la ponctuation, la segmentation graphique des mots et la résolution des abréviations. Le niveau diplomatique se rapproche davantage du système graphique du document source : les lettres restituées à la place des abréviations sont signalées par des italiques, les distinctions « ramistes » (phonétiques) des lettres *i/ĩ* et *u/ṽ* ne sont pas introduites, les diacritiques modernes ne sont pas ajoutés et aucune marque de ponctuation n'est utilisée, lorsqu'il n'y en a pas dans le document transcrit. La segmentation graphique correspond dans la mesure du possible à celle du document source[9]. Ce type de transcription peut être indispensable pour certains types de recherche linguistique, en particulier dans le domaine de la morphologie (Schøsler, 2004,

9. Dans certains cas, faute d'instrument de mesure précis, la lecture et la décision de transcrire un blanc entre deux mots du manuscrit restent à l'appréciation de l'éditeur.

p. 463). Pour réaliser une transcription « à deux niveaux », il n'est pas nécessaire de transcrire deux fois le texte source. Il suffit d'utiliser un petit nombre de raccourcis typographiques, dans le cadre d'une convention de transcription utilisant un mécanisme de caractères spéciaux, qui permettent de générer automatiquement les deux types de transcription à partir d'un fichier unique. Par exemple, le caractère dièse permet de signaler dans *#Dieu* que la majuscule du nom propre est due à la normalisation éditoriale et que la graphie du document source comporte une minuscule.

Les principes de la segmentation lexicale pour les outils d'annotation linguistique et pour le moteur de recherche peuvent être clairement définis et appliqués dans le cadre de grandes collections d'éditions numériques et, idéalement, partagés par la communauté internationale des philologues. Il convient de souligner que la normalisation de la segmentation au niveau du codage informatique n'empêche pas l'éditeur scientifique d'appliquer ses propres choix de segmentation visuelle dans l'édition à l'écran ou imprimée. En règle générale, le codage de la segmentation la plus fine est préférable, car il est plus simple de regrouper que de découper des unités *a posteriori*.

Le modèle économique des éditions numériques diffère considérablement des éditions imprimées. Le coût de la fabrication et de la diffusion du livre est important et peut justifier la cession des droits à l'éditeur commercial. Pour les éditions numériques basées sur une chaîne de production bien réglée et disposant d'une plateforme de diffusion adaptée, c'est le travail philologique de l'éditeur scientifique qui représente l'investissement principal. Le coût d'hébergement d'une ressource sur la toile est relativement faible, et des services à forte valeur ajoutée (impression à la demande, export dans un format particulier) peuvent être proposés aux lecteurs. Ceci rend tout à fait possible la diffusion des éditions numériques sous une licence libre de type *Creative Commons* ou similaire. Cela est important non seulement pour faciliter l'accès à la lecture de

ces éditions par les membres de la communauté académique et un public plus large, mais aussi et surtout pour permettre leur intégration dans des archives ouvertes, dans des corpus divers et variés, ainsi que dans la toile de données. La possibilité d'accéder aux données primaires des travaux de recherche pour reproduire leurs résultats est un élément important de leur scientificité. Enfin, plus une ressource numérique est utilisée et reproduite, plus il y a de chances qu'elle puisse s'adapter aux évolutions technologiques constantes.

La diffusion ouverte des données implique l'utilisation de formats de représentation ouverts, le respect des normes et standards d'encodage et la documentation des pratiques particulières à une équipe. Pour ce qui concerne l'encodage d'éditions scientifiques numériques, le cadre proposé par le consortium TEI (déjà évoqué plus haut) semble à ce jour le mieux adapté. Les avantages de la TEI sont sa riche expérience (plus de vingt-cinq ans d'existence), la variété des types de textes et d'éditions pris en charge, la souplesse des schémas de balisage proposés, sa documentation extensive et sa communauté active. Certains des points forts de la TEI peuvent également devenir ses faiblesses. Le très grand nombre de balises disponibles pour l'encodage et le fait qu'il existe toujours plusieurs façons de faire pour encoder un même phénomène rend difficile la mise au point d'outils d'analyse. La documentation fournie par la TEI ne suffit pas toujours pour expliciter les choix faits au niveau d'un projet de recherche particulier. Pour cette raison, la TEI recommande de personnaliser le schéma de balisage utilisé par un projet ou par une communauté et fournit un mécanisme facilitant cette personnalisation et sa documentation. La BFM utilise le balisage TEI pour ses éditions numérisées depuis le début des années 2000 et documente ses pratiques de manière précise (Bertrand *et al.*, 2014). Les éditions nativement numériques appliquent le même schéma de base que le reste de la BFM, mais utilisent un certain nombre de balises supplémentaires, notamment pour la représentation des transcriptions multi-niveaux.

Chaîne analytique ouverte pour l'exploitation textométrique des textes de la BFM

L'application d'un sous-ensemble précis des recommandations de la TEI documenté dans le cadre de la BFM a non seulement rendu possible la mise en place un réseau d'échanges de textes entre partenaires partageant les mêmes pratiques philologiques, mais elle a également permis à ce corpus de textes d'être intégré dans la plateforme TXM pour son analyse et sa diffusion.

Développée initialement dans le cadre d'un projet financé par l'ANR en 2007-2010, cette plateforme a pour objectif de pérenniser et de mutualiser les développements informatiques d'outils textométriques comme *Hyperbase, Lexico 3, Le Trameur, DTM* et *Weblex*. La textométrie est une méthode d'analyse de corpus textuels développée depuis les années 1980[10] combinant des outils statistiques appliqués aux différentes unités des textes (analyse factorielle, calcul de spécificités, classification, analyse de co-occurrents) et des outils documentaires (listes de mots, recherche plein texte de patrons de mots pour l'établissement de concordances, lecture des éditions de textes du corpus). Son implémentation dans la plateforme TXM a été l'occasion d'élargir la méthode aux corpus textuels richement encodés en XML-TEI et annotés par différents outils de traitement automatique de la langue (comme le lemmatiseur *TreeTagger*) et de produire une version pour poste Windows, Mac OS X ou Linux (appelée « logiciel TXM ») ainsi qu'une version serveur pour l'accès par internet (appelée « portail TXM »), les deux versions partageant la même plateforme de base pour l'exploitation des corpus.

La mutualisation de la construction et de la maintenance de la plateforme est obtenue par un mode de développement ouvert appelé « *open-source* », bien établi dans les projets de recherche en informatique depuis vingt ans, qui fonctionne sur deux plans. D'une part, tout partenaire peut accéder aux sources du logiciel pour l'adapter ou l'améliorer en respectant les termes de sa licence

10. En ligne : http://textometrie.ens-lyon.fr/spip.php?rubrique80.

de diffusion[11]. Et d'autre part, la plateforme intègre elle-même de nombreux composants logiciels développés par d'autres projets *open-source*, en particulier l'environnement de calcul statistique R[12], le moteur de recherche CQP[13] et la plateforme Eclipse[14] pour la version pour poste de TXM. Le fait de pouvoir accéder aux sources du logiciel TXM est par ailleurs un gage de scientificité des travaux réalisés grâce à cet outil, parce qu'il ne fonctionne pas comme une boîte noire. Tous ses calculs sont décomposables et vérifiables à partir de ses sources. Le fait de déléguer certains calculs à d'autres composants *open-source* permet de profiter de leurs performances et de leurs améliorations constantes par leur communauté de développement. Mais il faut s'assurer que chaque composant soit bien maintenu par une communauté de développeurs dynamique, par des institutions ou des entreprises au risque qu'il ne soit un jour plus développé et ne puisse plus suivre les évolutions technologiques et continuer à fonctionner. Auquel cas on doit soit le remplacer par un composant *open-source* équivalent, soit le maintenir soi-même.

La pérennisation du développement repose sur deux plans : d'une part l'utilisation d'un langage de programmation correspondant à un standard industriel reconnu et développé selon un mode communautaire ouvert (Java[15]) et une architecture logicielle standard (OSGi[16]), d'autre part l'utilisation d'une plateforme de versionnage des sources du logiciel, qui permet la traçabilité de l'attribution et de la datation de toute modification apportée aux sources et offre la possibilité de revenir à une version antérieure, quelle que soit sa date.

Conçue dès l'origine comme devant être capable d'exploiter des corpus textuels richement encodés en XML-TEI et annotés finement au niveau des mots, la plateforme TXM a pu utiliser la BFM comme corpus de validation de ses capacités d'intégration

11. La licence GNU GPL V3. Voir en ligne : http://www.gnu.org/licenses/gpl-3.0.fr.html.
12. Voir en ligne : http://www.r-project.org.
13. Voir en ligne : http://cwb.sourceforge.net.
14. Voir en ligne : https://eclipse.org.
15. Voir en ligne : https://www.jcp.org.
16. Voir en ligne : http://www.osgi.org.

et d'exploitation de corpus textuels riches en encodage et annotations.

La chaîne analytique de la BFM commence par un processus d'importation des fichiers sources encodés en XML-TEI dans la plateforme TXM à l'aide d'un module d'importation de sources appelé « XML-TEI BFM ». Ce module a été développé spécialement pour ce corpus à partir de la documentation des pratiques d'encodage XML-TEI des textes de la BFM telle qu'elle est publiée sur le site du projet de la Base. Il est chargé d'interpréter les fichiers source de façon à construire le « modèle de corpus » exploité par TXM. Les métadonnées nécessaires et utiles à l'analyse des textes sont extraites des en-têtes TEI, les éléments TEI pertinents pour l'analyse (comme par exemple les éléments <q> contenant le discours direct) sont indexés et certaines informations sont projetées au niveau des unités lexicales afin de simplifier les requêtes de recherche. D'autres éléments (comme les notes éditoriales) sont exclus de la surface du texte afin de ne pas être mélangés avec le corps du texte dans les recherches et les décomptes, mais sont intégrés aux éditions pour aider à la lecture des textes. Les éditions sont paginées en fonction des sauts de page encodés dans les sources numérisées. Une fois importée dans TXM à l'aide de ce module d'importation, la BFM bénéficie de tous les services d'analyse offerts par la plateforme dans sa version pour poste ou dans sa version portail. Le portail BFM[17] est un portail TXM hébergeant le corpus BFM. Il offre des services supplémentaires par rapport à la version pour poste de personnalisation de pages d'accueil ou de documentation, de création de comptes utilisateurs et de contrôle d'accès texte par texte.

Conclusion : une synergie entre les chaînes philologique et analytique pour une ressource libre

Aujourd'hui la BFM est consultée et analysée au moyen d'un logiciel libre (la plateforme TXM) et offre un accès libre aux sources de ses textes. Ces sources sont établies par une chaîne philologique complète et ouverte, de façon analogue à la chaîne

17. Voir en ligne : http://txm.bfm-corpus.org.

d'analyse qui repose sur le logiciel TXM, lui-même développé en *open-source*. L'emboîtement entre ces deux chaînes est rendu possible par un usage précis du standard de représentation des textes XML-TEI. Développée à l'origine pour l'échange de représentations numériques de textes entre partenaires, la TEI commence donc à mettre en relation des projets d'établissement de corpus de textes avec des projets de développement d'outils d'analyse et d'exploitation qui relèvent pourtant souvent de communautés de recherche très différentes en termes d'objectifs et de mode de fonctionnement. L'adoption parallèle d'un mode de fonctionnement ouvert par les deux chaînes pour faciliter la mutualisation et la traçabilité des développements (établissement de texte d'un côté, implémentation de méthode de l'autre) nous semble être un gage de pérennité et de scientificité des travaux pouvant être réalisés à l'aide de la BFM.

Références bibliographiques

Bédier, Joseph, « La tradition manuscrite du *Lai de l'Ombre*, réflexions sur l'art d'éditer les anciens textes », *Romania*, n° 54, 1928, p. 161-196 ; p. 236-356.

Bertrand, Lauranne, Lavrentiev, Alexei, Pincemin, Bénédicte, Guillot, Céline, Heiden, Serge et Lascar, Justine, *Tutoriel TXM pour la BFM*, Version 2.0, Lyon, ENS de Lyon, 2014. En ligne : http://txm.bfm-corpus.org/files/Tutoriel_TXM_BFM_V1.pdf.

Frappier, Jean (éd.), *La Mort Artu*, Genève/Paris, Droz/Minard, 1964.

Guerreau, Alain, *L'Avenir de la philologie. Textes anciens et domaine public*. En ligne : halshs-01112213, 2015.

Guillot, Céline, Lavrentiev, Alexei, Rainsford, Thomas, Marchello-Nizia, Christiane et Heiden, Serge, « La "philologie numérique" : tentative de définition d'un nouvel objet éditorial », dans Trachsler, Richard, Duval, Frédéric et Leonardi, Lino (dir.), *Actes du XXVII^e Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013). Section 13 : Philologie textuelle et éditoriale*, 2017.
En ligne : http://www.atilf.fr/cilpr2013/actes/section_13/Guillot_Heiden_Lavrentiev_Marchello-Nizia_Rainsford.pdf.

Guillot, Céline, Heiden, Serge, Lavrentiev, Alexei et Pincemin, Bénédicte, « L'oral représenté dans un corpus de français médiéval (IXe-XVe) : approche contrastive et outillée de la variation diasystémique », dans Jeppesen Kragh, Kirsten et Lindschouw, Jan (dir.), *Les Variations diasystémiques et leurs interdépendances dans les langues romanes. Actes du colloque DIA II à Copenhague (19-21 novembre 2012)*, Strasbourg, Éditions de linguistique et de philologie, 2015, p. 15-28.
En ligne : halshs-00760647.

Lebart, Ludovic et Salem, André, *Statistique textuelle*, Paris, Dunod, 1994.

Marchello-Nizia, Christiane, Lavrentiev, Alexei et Guillot-Barbance, Céline, « Édition électronique de la *Queste del saint Graal* », dans Trotter, David (dir.), *Manuel de la philologie de l'édition*, Berlin/Boston, De Gruyter, 2015.

Margoni, Thomas et Perry, Mark, « Scientific and Critical Editions of Public Domain Works: An Example of European Copyright Law (Dis)Harmonization », *Canadian Intellectual Property Review*, n° 27, 2011, p. 157. En ligne : http://ssrn.com/abstract=1961535.

Plouzeau, May, « À propos de *La Mort Artu* de Jean Frappier », *Travaux de linguistique et de philologie*, n° 32, 1994, p. 207-221.

Schøsler, Lene, « Historical corpora. Problems and Methods », dans Bozzi, Andrea, Cignoli, Laura et Lebrave, Jean-Louis (dir.), *Digital technology and philological disciplines*. *Linguistica computazionale*, t. XX-XXI, Pisa/Roma, Istituti editoriali e poligrafici internazionali, 2004, p. 455-472.

Stein, Achim et Prévost, Sophie, « Syntactic Annotation of Medieval Texts: the Syntactic Reference Corpus of Medieval French (SRCMF) », dans Bennett, Paul, Durrell, Martin, Scheible, Silke et Whitt, Richard (dir.), *Corpus Linguistics and Interdisciplinary Perspectives on Language (CLIP)*, n° 3, « New Methods in Historical Corpora », 2013, p. 275-282.

Terminographie diachronique : le cas de la terminologie médiévale française

Gérard Petit
EA 4509, Sens Texte Informatique Histoire
Université de Paris Nanterre

Problématiques

Une approche diachronique du lexique se heurte à au moins deux problèmes cruciaux :

– celui de la définition de la diachronie en question (T-1). L'intervalle couvert conditionnera la complexité des résultats obtenus ;
– celui du réglage théorique de l'approche en elle-même. À partir de quel modèle de représentation (morphologique et sémantique notamment) l'investigation est-elle menée ? Celui qui est contemporain de la recherche elle-même ou bien un autre, qui serait spécifique à la diachronie envisagée ?

Le projet CréaLScience (ANR-10-CREA-0007) a été l'occasion de remettre à plat un certain nombre d'acquis théoriques et méthodologiques, que leur évidence impose au chercheur, mais dont l'influence sur la nature des résultats ne doit pas être sous-estimée. D'abord en situant son objet, la langue médiévale, sur une synchronie isolée, passée (T-1), mais qui couvre à elle seule une véritable diachronie. À cet effet, il privilégie la cartographie en T-1 plutôt que l'évolution vers To, laquelle se déduira de la distance entre les deux configurations. Ensuite, en tentant de restituer autant que faire se peut l'état des représentations conceptuelles en T-1, telles qu'on peut supposer qu'elles étaient configurées et exprimées alors. De ce fait il renonce à les exposer à travers le filtre des connaissances

en To. Ainsi, il se démarque fondamentalement des perspectives diachroniques adoptées usuellement dans les recherches linguistiques. Mais il intègre en même temps l'incertitude et l'insécurité comme données heuristiques primordiales. Enfin, le projet CréaLScience vise la production d'un dictionnaire électronique représentant les données de l'investigation. Ce dictionnaire, le *Dictionnaire du français scientifique médiéval* (*DFSM*) est l'enjeu d'un développement d'outils (théoriques, méthodologiques, informatiques) dédiés à une investigation terminologique diachronique.

Le *DFSM* est un ouvrage inédit à plusieurs titres :

– il se dote d'un objet explicitement terminologique portant sur une synchronie non explorée jusqu'ici par la terminologie générale ;
– il se démarque des principaux dictionnaires consacrés à la langue du Moyen Âge sur plusieurs points : (i) il ne constitue pas un ouvrage de traduction intralinguale du français médiéval vers le français moderne ; (ii) il ne vise pas un objectif anecdotique, qui reposerait sur des dépouillements partiels et non explicites et qui chercherait à présenter une forme d'exotisme culturel, caractéristique d'un passé[1] ; (iii) il ne cherche pas à décrire des emplois non reconnus au sein de discours spécialisés[2] ; (iv) il vise une représentation des concepts aussi proche que possible de leur état dans la conscience supposée des locuteurs de T-1 ;
– ses inventaires reposent sur le dépouillement systématique de textes spécialisés[3].

1. Sont concernés ici des ouvrages destinés au grand public qui, en tant que tels, ont leur validité, mais qui n'entrent pas dans un projet scientifique. Citons entre autres : http://www.medieval-moyen-age.net/categorie-125346.html ; http://www.castlemaniac.com/lexique-medieval/lexique-medieval.php ; http://www.provins-banquet-des-troubadours.fr/vocabulaire-medieval.
2. Pour cette raison, ses inventaires se distinguent pour partie de ceux du *Dictionnaire du moyen français* (en ligne : http://www.atilf.fr/dmf).
3. Voir la bibliographie générale du projet.

Compte tenu de ces paramètres, deux orientations lexicographiques majeures ont été adoptées :

– **Dictionnaire « de langue »** *vs.* **dictionnaire de corpus**. Opérant sur un état de langue pour lequel aucune compétence vive, nourrie de l'interlocution, n'est disponible, le dictionnaire ne peut puiser ses informations que dans des corpus écrits, et reste donc tributaire de ceux-ci. Deux attitudes sont possibles : (i) soit il cherche à se dégager des corpus pour viser une représentation abstraite, et donc se conformer à son objectif initial (approcher les concepts dans la culture de T-1). Cet objectif n'est tenable qu'en opérant une moyenne sur les informations fournies par les corpus ; (ii) soit il adhère aux corpus, et s'achemine alors vers une forme de philologie. Aucune des deux directions ne peut être tenue pleinement, et ce pour des raisons différentes : l'objectif philologique ne présente aucun intérêt pour une démarche terminologique générale ; l'objectif linguistique, fondé sur des moyennes, ne peut être mené à son terme en toutes circonstances, compte tenu de l'état de langue investigué ;
– **Dictionnaire « de langue »** *vs.* **dictionnaire encyclopédique**. Le projet CréaLScience s'inscrit dans la tradition des dictionnaires encyclopédiques, puisque son objet est le vocabulaire technique et scientifique médiéval. Toutefois, visant la description des concepts en T-1, il s'éloigne de l'objectif de principe d'un dictionnaire encyclopédique traditionnel : son but n'est pas la connaissance du monde en lui-même (le réel), mais du monde pensé, représenté à un moment donné de l'histoire (les conceptologies). Par ailleurs, les emplois syntaxiques particuliers (notamment pour les verbes, le passif, l'emploi pronominal, le participe passé[4]) sont recensés et font l'objet d'un traitement dans des divisions spécifiques. Le dictionnaire envisagé est un ouvrage foncièrement terminologique, et s'oriente donc vers l'encyclopédie, tout en reposant sur un socle dont

4. Par ex. *aiguiser*, *apostumer*.

les fondements théoriques et épistémologiques sont explicitement linguistiques.

Cette étude consacrera un premier développement aux problèmes posés par l'établissement de la nomenclature, compte tenu notamment de l'extrême variabilité rencontrée s'agissant de la graphie du lexique dans la langue médiévale. Dans un second temps, nous aborderons la microstructure en concentrant notre attention sur les contraintes qui obèrent la représentation sémantique, l'objectif du *DFSM* étant, rappelons-le, d'approcher au plus près la configuration du concept, tel qu'il pourrait être formulé dans la synchronie visée.

La métalangue de description est le code dans lequel sont formalisés l'ensemble des descripteurs, mais aussi les règles qui président à la description elle-même. Au titre des descripteurs, on retiendra :

– la forme graphique de l'entrée ;
– l'indication de catégorie grammaticale ;
– la définition ;
– l'exemple et sa référence bibliographique ;
– les notes diverses, linguistiques et encyclopédiques, qui accompagnent de droit la définition.

L'ensemble de ces descripteurs est usuellement réparti entre la macrostructure (les entrées d'un dictionnaire) et la microstructure (les constituants de l'article de dictionnaire).

La macrostructure : gérer la variation, frein de la lexicalité

La macrostructure est la liste des entrées d'un dictionnaire. Elle constitue le premier matériau réuni en vue de son élaboration. Fonctionnellement, elle renseigne sur les choix et prélèvements effectués au sein des corpus, le degré de profondeur ou d'étendue de l'investigation. L'établissement de la nomenclature est également l'enjeu d'un réglage morphosyntaxique[5], capital quand il s'agit de la langue médiévale, concernant :

5. Il touche par ex. le classement alphabétique des caractères accentués quand ils ont valeur discriminante *(tache*/*tâche* ; *coupe*/*coupé*).

(1) le traitement à appliquer à la variation graphique et à la difficulté, parfois très importante, à dégager un lemme ;

(2) le mode de classement des séquences polylexicales (noms composés, locutions adjectivales ou verbales) ;

(3) le traitement à apporter aux emplois morphosyntaxiquement contraints (passif, participe passé, pluriel, emplois substantivés, adjectivés) au sein du dictionnaire : disposent-ils d'une entrée ou bien sont-ils traités comme subdivisions d'un article ? Sur la base de quels critères la décision est-elle prise ?

Le traitement de la variation graphique et/ou phonologique

La langue médiévale se caractérise par une forte variation (notamment graphique et morphologique), laquelle hypothèque parfois l'identification d'un lemme commun à plusieurs occurrences et susceptible de synthétiser l'unité psycholinguistique du signe. Une investigation ciblée portant sur les lettres A et C révèle les configurations suivantes :

Existence d'un terme en ancien français (AF), mais pas en français moderne (FM)

Sont concernés les termes renvoyant soit à un concept sans équivalent en FM (par ex. : *caladre*), soit à un concept disponible en FM mais associé à un autre signifiant (par ex. : *amoustir*, *apostume*, *eaueux*, *aigos*, *aidement*). La variation affecte la structure morphologique de l'unité.

Tableau 1. Variations terminologiques entre FM et AF

Français moderne	Ancien français
humidifier	*amoustir*
abcès, tumeur	*apostume*
« oiseau blanc qui a le pouvoir de prédire la mort… »	*caladre*
aqueux	*eaueux, aigos*
aide	*aidement*

Existence d'un terme en AF avec des graphies multiples, dont éventuellement celle attestée en FM (nous soulignons)

Ces variantes peuvent résulter de différences phonologiques affectant plus ou moins largement la structure du signifiant.

Tableau 2. Variations graphiques en AF

Français moderne	Ancien français
courage	*corage, coraige,* ***courage***
concombre	*cucumère, concourde,* ***concombre****, coucombre, cucumerus*
aloès	*alee, aloe, aloee, aloem, aloen, aloü, aloüs*

Existence d'un terme en AF avec des graphies multiples, mais non celle attestée en FM

La variation peut affecter trois plans différents de l'unité : graphique *(cyprès*/*cypré* ; *cyclamen*/*ciclamen*), phonologique *(cyprès*/*cipre),* morphologique *(camomille*/*chermière*), de manière indépendante ou cumulée[6].

Tableau 3. Graphies spécifiques à l'AF

Français moderne	Ancien français
croûte	*crouste, croste*
cyclamen	*ciclamen, ciclam, malum terre, panis porcinus, pain à porc, pain de pourceau*
camomille	*chermière, camomil*
cyprès	*cyprus, cypré, cipré, cypres, cippre, cyperi(s)*
châtaignier	*chastignier, chasteignier, chastaignier*
aiguiser	*aguiser*
haleine	*alaine, alainne*
ammoniac	*amoniac, armoniac, amoniat*

La difficulté consiste donc à déterminer la forme de l'entrée. Dans la configuration (1), il n'existe pas d'autre alternative que d'entrer la forme médiévale en vedette, en sélectionnant celle qui sera la moins marquée dialectalement en ancien français, ou la mieux attestée en cas de variantes graphiques. Pour (2) et (3), le choix pourrait se porter sur le FM ou l'AF. Toutefois, partant du principe que le *DFSM* s'adresse à un public médiéviste, mais

6. Les dimensions de cette étude ne nous permettent pas d'entrer ici dans le détail.

aussi à des lecteurs venant de tous horizons, le choix de la forme en FM sera privilégié, ce pour plusieurs raisons :

– le FM peut être valablement tenu comme la langue de consultation des lecteurs non médiévistes ;
– le FM est la langue dans laquelle s'effectue la description dans la microstructure ;
– l'ensemble des variantes recensées dans le *DFSM* ne sont pas connues de tous les médiévistes.

Les articles consacrés à des termes présentant des variantes sont donc traités de manière hiérarchique :

– le lemme de l'entrée est exprimé en FM et institue ainsi un article tutélaire : *croûte*, *cyclamen*, *camomille*, *cyprès*, *châtaigner*, *aiguiser*, *ammoniac*, *haleine* ;
– un renvoi par lien hypertexte à la (ou aux) variante(s) est ajouté sous l'entrée ;
– la (ou les) variante(s) présente(nt) un article vide, constitué uniquement d'un renvoi à l'article tutélaire : *crouste, croste* vers *croûte* ; *amoniac, armoniac, amoniat* vers *ammoniac*, etc.
– lorsque la variation affecte la structure morphologique de l'unité, les différentes formes sont traitées indépendamment, chacune comme vedette d'un article plein avec définition et renvoi synonymique réciproque : *cyclamen* vs. *pain à porc, pain de pourceau*.

Variante et nomen

La variation graphique, morphologique et la synonymie affectent des unités exprimées en AF, mais aussi des formants latins apparaissant dans des énoncés produits en AF. Certains de ces formants correspondent à des unités disposant par ailleurs d'une dénomination en AF :

– lat. *malum terre*[7], *panis porcinus* / fr. *ciclamen* (FM : *cyclamen*) ;
– lat. *cacochia* / fr. *cacochie*[8] ;

7. Littéralement, « pomme de terre » (sic !).
8. « Dégradation de l'état de santé due à un déséquilibre des humeurs* ».

– lat. *calamus aromaticus* / fr. *calame aromatique*[9] ;
– lat. *calcatrix* / fr. *hydre* ;
– lat. *capparis* / fr. *capre* (FM : *câpre*).

Ces formants latins sont la trace d'une dénomination antérieure, laquelle perdure alors qu'une concurrence avec le français s'est installée. La relation du latin au français médiéval peut être de l'ordre de l'adaptation morphosyntaxique (*cacochia*/*cacochie*, *calamus aromaticus*/*calame aromatique*) ou de la traduction (éventuellement à partir d'un autre modèle) : *panis porcinus*/*pain à porc* ; *calcatrix*/*hydre*[10]. Ces formants fonctionnent dans le discours scientifique de manière analogue aux nomens (étiquettes latines que les classifications ultérieures utiliseront, aux XVII^e^ et XVIII^e^ siècles, pour systématiser les taxinomies).

Pour tout binôme associant un terme français à un nomen, que des variantes graphiques existent ou non pour l'un et/ou l'autre, les deux unités disposent chacune d'un article plein, avec définition(s) et exemple(s), du fait qu'elles sont morphologiquement distinctes. Certaines variantes présentent un caractère hybride, car elles mêlent le latin et le français. Tel est le cas de *bol armenicum*, variante de *bol armenic* (« Argile que l'on fait venir d'Arménie, et qui se caractérise par sa couleur rouge et sa grande viscosité »). Du fait de leur divergence morphologique, les deux formes sont assimilées à des synonymes car elles ne constituent ni des nomens à part entière, ni de simples variantes graphiques. Leur existence met néanmoins en évidence un phénomène dont la terminologie médiévale se fait le témoin : la distinction entre statuts sémiotiques (dénomination, variante, nomen, synonyme) n'est pas étanche, mais se négocie sur la base d'une continuité, d'où l'existence d'unités qui n'appartiennent pleinement à aucune catégorie faute d'en présenter toutes les caractéristiques.

9. « Petit arbre* qui pousse en Inde, dans des sols humides et qui a une tige creuse, ses feuilles ressemblant à celles de l'iris, roseau odorant ».
10. La place nous manque pour traiter le cas de *malum terre*, littéralement « pomme de terre ».

D'autres formants latins ne disposent pas d'équivalent français : par ex. *carpobalsamum*. Leur effectif est moins nombreux que le précédent, les auteurs médiévaux cherchant à franciser les dénominations afin d'homogénéiser l'expression de leur discours. Ces termes latins « orphelins » sont traités comme des dénominations à part entière : ils figurent en entrée d'article, sont définis et illustrés d'exemples. Ces unités présentent toutefois un régime qui rend délicate l'assignation d'un statut sémiotique fixe : dénomination, variante ou nomen. Ainsi, pour poursuivre avec le terme *carpobalsamum* :

> Fruit d'un arbrisseau, le baumier, charnu et à un seul noyau blanc osseux, d'un gris rougeâtre, d'une saveur légèrement amère et aromatique.

Cette unité fonctionne dans les discours comme nomen associé à *baume*. Or, sa définition précise qu'il identifie un référent spécifique à une sous-catégorie de celui-ci. *Carpobalsamum* fonctionne donc comme une dénomination à part entière et se positionne comme hyponyme de *baume*. Ce type de difficulté est révélateur de l'ambivalence sémiotique qui affecte certains termes de la langue médiévale.

Entrée et sous-entrée

Une tradition lexicographique empiriquement instituée depuis le XVII^e^ siècle affecte l'entrée aux dénominations correspondant au gabarit du mot graphique[11]. Les unités polylexicales et autres locutions sont traitées comme des sous-entrées : elles figurent à l'intérieur de l'article, dans une division spécifique. Ces sous-entrées disposent de leur propre lemme, sont définies et illustrées d'exemples. Bref, elles reçoivent (presque) le même traitement que les entrées à part entière.

Le *DFSM* comporte des sous-entrées, qui reçoivent le même traitement que celui appliqué dans les autres dictionnaires. Ainsi, sous l'entrée *ail*, figure le nom composé *ail sauvage* :

11. Le mot graphique peut se définir comme une chaîne de caractères séparée par des espaces. Le trait d'union vient abolir l'espace et réduire au format du mot graphique des suites polylexicales (par ex. *château-fort*). Les mots à trait d'union figurent traditionnellement en entrée d'article dans les dictionnaires.

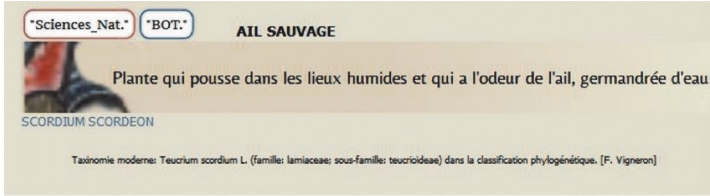


Fig. 1. Ail sauvage (*DFSM* - capture d'écran)

qui dispose de ses propres nomens, différents de celui d'*ail* (*allium*).

Le champ réservé traditionnellement à l'expression de la sous-entrée est également affecté aux emplois syntaxiques particuliers de l'entrée qui justifient une définition à part entière, associée à une relation référentielle spécifique. Ainsi le verbe *consolider* connaît-il trois emplois dans le domaine médical :

– transitif : « Assurer la cicatrisation d'une plaie, la consolidation d'une fracture » ;
– intransitif : « Assurer sa propre cicatrisation en parlant d'une plaie, sa propre réparation en parlant d'une fracture » ;
– pronominal : « Se fermer, *en parlant d'une plaie ou d'un ulcère* », ce dernier disposant d'un périmètre sémantique différent (nous soulignons).

Pour prendre un autre exemple, le verbe *aggraver* possède en AF une acception médicale spécifique, au passif : « Être attaqué, mis en danger, en parlant du corps et en particulier des vertus* qui l'animent », qui est renseignée en sous-entrée.

Fonctionnellement, le champ de la sous-entrée accueille du matériau qui marque un écart relativement au régime prédictible ou lexicalisé de l'entrée, soit au plan morphologique, soit au plan syntagmatique. Il constitue le site d'accueil des formes ne présentant pas la sémiotique intègre de l'entrée, mais ne justifiant pas pour autant leur éviction du champ de l'unité psycholinguistique du signe.

Les signes diacritiques

La nomenclature est la liste des entrées d'un dictionnaire. C'est aussi un ensemble de règles qui président à son organisation. À ce titre se pose régulièrement la question du traitement des lettres portant des signes diacritiques (accents, tréma, cédille) et notamment de l'ordre de leur classement relativement aux caractères « simples » dont ils s'écartent. Les dictionnaires choisissent usuellement de les classer soit avant, soit après ceux-ci dans l'ordre alphabétique de la nomenclature.

La graphie des entrées privilégiant le FM pour favoriser l'ergonomie de la consultation, la distribution et le classement des caractères accentués ne répond pas à une propriété de la langue médiévale, qui s'écrit sans accent : l'accent aigu attesté sur le « e » en finale de mot est une commodité moderne permettant de distinguer des participes de formes homonymes ou d'éviter des interprétations fautives. Si le « ç » est attesté, il dispose d'un régime différent de celui qui est le sien en FM :

(1) il peut se rencontrer devant une voyelle autre que « e » et « i » à l'intérieur d'un mot, comme en FM : *commençoit* (FM « *commençait* ») ;

(2) toujours devant une de ces voyelles, il peut apparaître comme variante de « s » ou de « ss », par ailleurs attestés en AF *(çucre,* pour *sucre* ; *cuiçon*, pour *cuisson*) ;

(3) il peut doubler un « s » (*sçavoir,* pour *savoir*) ;

(4) comme on le voit en (2), le « ç » peut figurer à l'initiale d'un mot, position qui n'est plus la sienne en FM. La nomenclature du *DFSM* associe les deux logiques diacritiques, celles du FM et de l'AF. Ainsi, pour reprendre un exemple cité plus haut, le dictionnaire comprendra deux entrées : *sucre* et *çucre,* la seconde n'étant qu'une variante de la première, ou encore une entrée *çoire* (synonyme de *pois chiche*). Les lemmes comportant des caractères accentués relèvent donc uniquement d'une intervention en FM. Ajoutons que, dans le but d'éviter une confusion dans la prononciation, certains lemmes de termes spécifiquement médiévaux ont

été accentués *(atempréement* à la nomenclature devient *atempreement* dans les citations).

Dans le déroulé alphabétique de la nomenclature du *DFSM* les caractères portant des signes diacritiques n'entrent pas en concurrence avec leurs équivalents nus. Pour cette raison, ils sont traités à l'identique de ceux-ci (nous soulignons) :

cedewale > ***cèdre*** > ***ciguë*** > *ceindre*
cane > *canel* > *canele* > ***çaneson*** > *canette*
capel > ***céphalée*** > *cephalica*
codling > ***cöe*** > ***coéqualité*** > ***coéquation*** > *coer* > *cœur* > *coffin*
coingnier > ***çoire*** > *cois*

Actuellement, compte tenu de l'état de la nomenclature, aucune concurrence n'apparaît entre caractères accentués et non accentués justifiant une décision relative à l'ordre d'apparition des lemmes dans la succession alphabétique. L'orientation terminologique, donc domainière du dictionnaire n'est probablement pas étrangère au fait que la question de la précellence d'une lettre sur une autre ne se pose pas dans la synchronie considérée.

La microstructure

La microstructure d'un dictionnaire comprend l'ensemble des informations ayant en charge la description de l'entrée. Nous examinerons ici l'une de ses composantes principales : la définition.

La définition : la diachronie contre l'anachronisme

La définition est à la fois une opération sémantique et la séquence linguistique qui la concrétise. En terminographie, elle possède également une fonction signalétique, car elle justifie le rattachement d'un terme à un domaine.

Nous ne reviendrons pas sur la difficulté de principe liée à l'activité de définition, *a fortiori* lorsqu'elle est terminologique, et qui repose sur le postulat (indémontrable !) que le sens est une entité close, objectivable dans une périphrase, que l'on peut segmenter en éléments simples sur le même modèle que

les sons d'une langue[12]. La question devient plus aiguë dès que l'on aborde le lexique en diachronie et que l'on cherche à représenter le concept dans sa configuration T-1. La définition lexicale ou terminographique n'est plus alors qu'une hypothèse, dont les contours sont au moins partiellement informés par les représentations de son producteur en To[13].

Approcher le concept au plus près de son état dans la synchronie T-1 implique de reproduire l'état de la culture qui pensait le réel, tel qu'elle le pensait. C'est également projeter une perspective sur la terminologie et admettre que les concepts ne reflètent pas un réel en soi (tel que le postuleraient une onomasiologie ou une ontologie modernes), intemporel et insensible aux variations de l'espace, mais sont des représentations ancrées dans la culture de leur temps, de leur espace géographique et surtout dans les mots qui les véhiculent. Pour cette raison, la terminologie est abordée sous un angle sémasiologique. Elle repose sur le principe que la définition d'un concept doit passer par des termes en accord avec la conscience de l'époque et du lieu.

Le risque rencontré par la terminographie est alors celui de l'anachronisme. Celui-ci peut avoir deux sources : la configuration du concept lui-même ou bien le choix des définissants. Le *DFSM* tient compte de ces deux ordres de contraintes, bien qu'ils répondent à des logiques différentes. La structuration du concept, envisagé comme chaînage de traits structuré sur un modèle logique par incluants et différences spécifiques, peut être attestée dans le corpus par des énoncés définitoires (nous soulignons) :

> Boterel **est un ver envenimé qui** habite en terre et en lieu moiste si come dit Plinius ou XXIIe chapitre de son XVIIIe livre. (Jean Corbechon, *Livre des propriétés des choses*, XVIII, 15, fol. 299 rb.)

12. Nous préciserons seulement que la définition, qu'elle soit lexicographique ou non, reflète davantage la construction du modèle de représentation sémantique dont elle est l'application que les données du réel extralinguistique.
13. Or, approcher le concept en T-1 impose que l'on fasse abstraction de ses propres représentations sémantiques en To.

> Apotaine[14] **est un poisson qui** est apellez chevalfl[uvi]el, por ce que il naist ou fluve dou Nil ; et son dos et ses crins et sa voiz est come dechevals. Ses ongles sont fendues come de buef et danz come sengler et la coe retorte, [et mangue] ble[s] [de champ] ou il vet a recolons por les aguaiz des homes. (Brunet Latin, *Trésor*, I, 135, p. 240)

Les énoncés dénominatifs constituent le second type de matériau investigué en priorité. Ils présentent une forme alternative des énoncés définitoires. Ils permettent entre autre de situer les termes au sein de champs notionnels, ou bien les variantes d'un même terme entre elles (nous soulignons) :

> La voie de respirer on **l'appelle** trachee artere et le chief on **l'appelle** epiglote, la voie de la viande on **l'appelle** meri ou ysophagus et la voye moyenne entre ces deux conduis, on **l'appelle** gorgeron et la languette qui est sur ces deux conduis, on **l'appelle** la vule ou luette. (GORDON, Prat., c.1450-1500, IV, 1)

> Persicaire, c'est une herbe qui a les feules qui ressemblent à feules de de pechier ; l'en **l'appelle** currago ou currage. [...] Aucuns l'appellent sanguinaire. (Sec. Sal., éd. Camus, XV^e^ s., p. 102)

Il est possible de maîtriser le choix des définissants en se référant aux connaissances disponibles actuellement sur l'AF et sur les sciences et techniques au Moyen Âge. Toutefois, la configuration d'ensemble du concept reste sujette à discussion si elle ne se trouve pas attestée dans un énoncé. Souvent, comme c'est le cas pour le sens de *cyclamen* (cf. plus haut), elle résulte de recoupements d'énoncés. Pour cette raison elle se présente davantage comme une proposition que comme le reflet strict et exhaustif d'un état de conscience[15].

Comme on peut s'y attendre, le périmètre sémantique des termes médiévaux n'est pas isomorphe de celui de leurs équivalents modernes. Deux grands faisceaux d'explications peuvent être avancés pour justifier de ce constat.

---

14. *Apotaine* : hippopotame. Pour la conscience médiévale l'hippopotame est un poisson !
15. À supposer qu'un tel objectif soit tenable, et même qu'il ait un sens. Les définitions lexicales et lexicographiques en synchronie contemporaine sont-elles autre chose que de simples propositions ?

En premier lieu, l'évolution du monde réel et des représentations qui l'accompagnent, laquelle aboutira à ce que certaines dénominations voient leur représentation sémantique changer. Ainsi, on ne définit plus aujourd'hui la chirurgie comme on le faisait au Moyen Âge, la pratique et la profession qui l'exerce ayant évolué. Ainsi, si l'on peut définir en To la chirurgie, et le nom *chirurgie*, par :

> Partie de l'art médical qui consiste à faire avec la main ou à l'aide d'instruments certaines opérations sur le corps de l'homme. (Wiktionnaire)
>
> *A specialty in which manual or operative procedures are used in the treatment of disease, injuries, or deformities*. (Termscience[16])
>
> Partie de la thérapeutique qui met en œuvre des procédés manuels et l'usage d'instruments, et qui groupe elle-même diverses spécialités selon les organes ou appareils intéressés (chirurgie thoracique), les buts recherchés (chirurgie réparatrice), etc. (*Trésor de la langue française*)

en T-1 sa signification sera circonscrite à :

> Pratique de l'art de guérir les maladies et les maux du corps par des incisions, cautérisations et replacement des os, et autres opérations à l'aide d'instruments.

L'autre grand faisceau de causes tire ses références non pas de la mutation des pratiques mais de celle des savoirs eux-mêmes et des représentations qui y sont associées. Ainsi l'une des différences majeures entre les représentations médiévale et moderne tient-elle à la coupure épistémologique engagée à la Renaissance et qui a entraîné, jusqu'au XVIII^e^ siècle, une refonte des principes théoriques et méthodologiques des sciences et des techniques. Pour cette raison certains domaines de connaissances sont à repenser totalement, comme la botanique (et l'ensemble des sciences de la nature), pour être abordés dans leur configuration médiévale. Ainsi, le terme *cyclamen* se définit-il aujourd'hui par :

> Plante **herbacée** à gros **tubercule** d'où naissent des racines, à feuilles en cœur, d'un vert sombre moucheté de blanc et

16. En ligne : http://www.termsciences.fr/-/Index/Rechercher/Rapide/Naviguer/Arbre/.

> d'un rouge pourpre sur leur face inférieure, à fleurs solitaires pendantes, blanches, rosées ou violines, surplombant un long **pédoncule** recourbé, présentant une **corolle** à cinq **pétales** renversés, tordus sur eux-mêmes, et donnant naissance à un fruit en **capsule** arrondie. (*Trésor de la langue française*)
>
> Plante vivace tuberculeuse de la famille des Primulacées, selon la classification classique et selon la classification phylogénétique (APG III) (Anciennement Myrsinacées APG II). Le cyclamen n'a pas de parenté nette avec les autres primulacées, quoiqu'il ressemble aux *Dodecatheon* d'Amérique du Nord par ses pétales renversés. (Wikipedia)

Définitions que l'on peut compléter par la taxinomie botanique :

> Règne *Plantae*, Sous-règne *Tracheobionta*, Division *Magnoliophyta*, Classe *Magnoliopsida*, Sous-classe *Dilleniidae*, Ordre *Primulales*, Famille *Primulaceae*, Genre *Cyclamen L., 1753*.

Ces connaissances témoignent de savoirs modernes (nous soulignons ci-dessus), lesquels n'étaient pas disponibles au Moyen Âge. Dans le *DFSM*, le terme *cyclamen* se définira par :

> Plante des bois dont la racine est en forme de pomme et qui servait notamment de nourriture pour les porcs.

Cette définition est conforme aux informations livrées par les corpus. L'exemple de *cyclamen* ci-dessus illustre une évolution conceptuelle entre T-1 et To, laquelle se traduit par un accroissement, une précision plus grande et une systématisation des informations classifiantes à la période moderne. En l'occurrence, elles expriment les propriétés intrinsèques du référent. Mais pour certains termes[17], la définition impose une prise en compte de la signification spécifique des définissants en T-1, comme le montrent les exemples d'*écrevisse*, de *boterel*, d'*apostume*, d'*humeur* et d'*armoise* (ci-dessous). Sont marqués d'un astérisque à droite les définissants présentant une pertinence pour la conceptualisation du défini en T-1, mais dont

17. Aucune quantification ni proportion n'est actuellement possible, mais les études menées sur les lettres A, B et C montrent que, comme on pouvait s'y attendre, l'immense majorité du vocabulaire est affectée.

l'interprétation doit tenir compte de la signification dans cette même synchronie :

> *écrevisse* : Petit poisson* de mer protégé par une enveloppe rigide, qui se déplace en reculant.
>
> *boterel* : Ver* venimeux aux yeux rouges qui fréquente les lieux humides et subit une mue, crapaud.
>
> *abeille* : Mouche* qui fabrique le miel et la cire.
>
> *apostume* : Toute enflure, grosseur causée par une corruption* des humeurs*.
>
> *bègue* : Personne dont la parole est corrompue* comme par un tremblement.
>
> *armoise* : Herbe* chaude* et sèche* avec des feuilles plus grandes et grasses que l'absinthe*, de longues tiges et de petites fleurs blanches odorantes en été[18].

Pour ne prendre que deux exemples, la classification médiévale des animaux ne s'opérait pas sur la base de propriétés anatomiques, comme c'est le cas aujourd'hui, mais sur la prise en compte du milieu de vie dominant. Pour cette raison, tous les animaux vivant dans un milieu aqueux sont considérés comme des poissons, y compris la baleine et l'hippopotame ! Sur un modèle homologue les végétaux sont classés en agronomie selon leur localisation par rapport au niveau du sol : ceux qui, comme le cyclamen, poussent sous terre[19] sont des *racines* ; ceux qui croissent au-dessus de la surface de la terre sont des *fruits* ; ceux qui, comme le poireau, occupent une position intermédiaire, sont des *herbes*.

Restituer le concept dans sa configuration T-1 implique un certain nombre de contraintes, dont :

– celles portant sur le choix et l'emploi des définissants (cf. paragraphes précédents) ;
– celles pesant sur le contrôle de la métalangue de description en To. Elles affectent le schéma de la définition, le libellé de clauses définitionnelles de sorte à conférer à l'ouvrage une

18. *Chaud* et *sec* renvoient à la théorie des quatre qualités fondamentales, issue de la *Physique* d'Aristote.
19. Ou sont estimés comme tels.

valeur heuristique non pas seulement quant à la connaissance (en To) de la terminologie médiévale, de sa structuration sémantique, de sa polysémie, de sa construction morphologique ou polylexicale, mais aussi quant à son organisation en tant qu'ensemble de représentations culturelles hiérarchisées qui procèdent d'une conception du monde en même temps qu'elles la véhiculent et la font ou non évoluer.

C'est ce second faisceau de contraintes qui sera envisagé dans les paragraphes qui suivent.

Le schéma de définition : la logique des classes

La sémiotique d'un dictionnaire diachronique peut être perçue (et conçue) selon deux modalités :

(1) l'une, intralinguale, confronte une langue ou un état de langue avec un(e) autre. Elle viserait à exprimer en français moderne ce qui est énoncé et conçu en langue médiévale ;

(2) une seconde modalité, extralinguale, confronte le lexique et le monde, le texte de dictionnaire étant alors autosuffisant. Il n'a besoin d'aucun point de référence pour être interprété.

Une logique intralinguale repose sur plusieurs prérequis. Un premier prérequis considère que les états de langue sont suffisamment conjoints pour que l'un (To) fournisse l'interprétant de l'autre (T-1). Dans ce cas, l'état de langue To véhicule l'ensemble des paramètres (orthographiques, sémantiques, syntaxiques, culturels) nécessaires à l'interprétation de T-1. Un second prérequis, découlant du précédent, favorise dans ce cas l'adoption du modèle du dictionnaire bilingue : la traduction intralinguale de T-1 vers To s'opère dans le même schéma sémiotique que celui opposant deux langues différentes, l'une inconnue, et l'autre connue. Dans ce cas, le modèle de définition retenu est synonymique : un terme de la langue A (en l'occurrence T-1) est traduit par un terme en langue B (To), qui est son équivalent.

La définition par synonymie repose sur un postulat, difficilement défendable en soi et indéfendable dans le cas de la terminologie médiévale, selon lequel le concept associé au

terme serait un invariant diachronique. L'exploitation des corpus à la base du *DFSM* confirme que les concepts sont ancrés dans la langue et la culture de leur temps. On aboutit là à une forme de paradoxe dans la mesure où les référentiels terminographiques produits ces dernières décennies reposent précisément sur l'invariance du concept. À décharge, il convient de préciser qu'ils portent sur la synchronie contemporaine. Les dictionnaires spécialisés produits durant les siècles passés (notamment celui de Thomas Corneille et l'*Encyclopédie* de Diderot et d'Alembert) s'appuyaient sur le même principe car ils opéraient toujours en synchronie contemporaine. La terminologie diachronique est très jeune. Lorsqu'elle a débouché sur des référentiels, ceux-ci se sont inspirés de modèles qui n'intègrent pas la variation diachronique du sens, car ils se fondent sur la permanence[20] du réel et calquent le modèle du dictionnaire bilingue :

> *oestre*, n. m. : « grosse mouche, taon » ;
> *oeuchine*, n. f. : « atelier, officine », en général ; en particulier, « atelier de foulon, de teinturier, de brasseur[21] » ;
> *cyclamen*, subst. masc. : « Cyclamen » ;
> *apostume*, subst. masc. : Au propre, « Tumeur, abcès ». Au fig., « Blessure qui s'envenime[22] ».

Soit ils définissent le terme d'AF par son équivalent en FM (*oestre*, *cyclamen*, *oeuchine*, *apsotume* [sens propre]), soit ils le glosent quand le sens en AF ne dispose pas d'équivalent en FM (sens figuré d'*apostume*). Ces dictionnaires sont des objets hybrides, qui ne réservent la définition logique qu'à l'écart ontologique[23] (et non pas sémantique) entre les synchronies T-1 et To.

Le *DFSM* procède différemment. Optant pour une perspective sémasiologique, il ne considère pas la représentation du réel comme identique en T-1 et To. Pour cette raison, quelle que soit l'entrée, la définition qui lui est appliquée procède sur le modèle logique, par incluants et différences spécifiques, que

20. Souvent dans une représentation très approximative de celui-ci.
21. *Oestre* et *oeuchine*, extraits du dictionnaire de Godefroy.
22. *Cyclamen* et *apostume*, extraits du *DMF*.
23. Le terme renvoie à une réalité spécifique de T-1, orpheline en To.

ce mot soit inconnu en FM (*caladre*), qu'il soit connu avec la même signification[24] (*calcul*) ou avec une signification différente (*abominable, abonder*) :

> *caladre* : Oiseau blanc qui a le pouvoir de prédire la mort et de détruire certaines maladies par le regard.
> *abominable* : Qui provoque des nausées.
> *abonder* : Être en grande quantité ou en excès, en parlant d'humeurs*, de superfluités ou encore du lait maternel.
> *calcul* : Opération portant sur des nombres, des caractères numériques, ou des objets représentant des nombres et destinée à obtenir d'autres nombres.

Il peut se produire, localement, qu'un terme soit défini par un synonyme, mais uniquement si celui-ci présente une signification spécifique à la langue médiévale :

> accidents de l'âme : Passions*

mais le fait n'est pas une généralité à l'échelle du dictionnaire.

Adopter un format de définition logique impose que l'on se penche sur la distinction entre propriétés intrinsèques (PI) et propriétés extrinsèques (PE) du référent. Les PI décrivent l'entité sous son angle concret, perceptuel : dimensions, poids, texture, silhouette, couleur, aspect général, etc. Les PE l'envisagent sous son angle fonctionnel : usage, agents impliqués, résultats attendus, modalités d'action, etc. Une définition par PI est prédisposée à caractériser les entités concrètes, mais peut également s'appliquer à certains adjectifs ou à des verbes associés à des effectuations perceptuelles. Les PE, quant à elles, sont réservées à la seule caractérisation des artefacts, concrets ou abstraits. Pour cette raison, la définition d'un végétal différera selon qu'elle sera indexée en botanique, en agronomie ou en diététique. Dans ces deux derniers domaines, les PE jouent un rôle fondamental dans la caractérisation du référent, alors

24. D'après l'examen des lettres A, B et C du *DFSM*, les unités présentant une signification identique en FM et en AF sont en nombre extrêmement restreint. Par ailleurs, nous restons très prudent sur ce constat d'isomorphisme dans la mesure où il serait certainement contredit par une approche convoquant des niveaux de spécialité plus élevés.

qu'elles sont absentes de la description d'une espèce naturelle[25]. Ce type de caractérisation convient particulièrement aux noms[26] et à certains verbes, dont la contrepartie référentielle peut être décrite en termes perceptuels (*marcher*, *courir*, *s'agiter*) et orientée vers une finalité.

L'exemple ci-dessous montre la répartition des PE et PI dans la définition d'*ammoniac* :

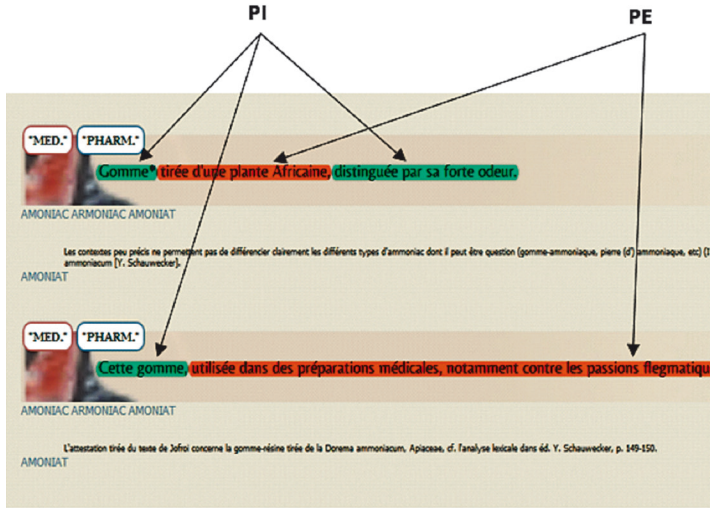


Fig. 2. Répartition des PF et PI dans *ammoniac* (*DFSM*, capture d'écran)

Les définitions de *caladre* et *calcul* montrent la limite extrême de cette répartition (nous soulignons en gras les PI) :

> *caladre* : **Oiseau blanc** qui a le pouvoir de prédire la mort et de détruire certaines maladies par le regard.
>
> *calcul* : Opération portant sur des nombres, des caractères numériques, ou des objets représentant des nombres et destinée à obtenir d'autres nombres.

La répartition entre PE et PI permet, pour la nomenclature des noms, d'opérer d'importants dégroupements entre domaines.

25. En tant que telles, les espèces naturelles sont totalement dépourvues de PI.
26. Aux noms concrets, mais aussi aux abstraits qui dénomment des processus (cf. *calcul*, *opération*, *amputation*…).

C'est à la définition de justifier le rattachement de tel terme à son domaine. À cet égard elle possède une fonction signalétique. Tout terme se rattache à un domaine : il faut entendre par là qu'il n'existe pas de terme sans domaine[27]. Ainsi, il n'est guère suffisant de définir *abject*, en Médecine, par « repoussant » seulement. Cette contrainte vaut aussi bien pour des termes dont l'emploi est strictement spécialisé (*alchane*, *bétoine*) que pour ceux qui disposent par ailleurs d'une signification en langue courante ou d'une saillance particulière dans d'autres domaines (*acteur*) :

> *alchane* (Médecine) : **La poudre** obtenue à partir de cette plante[28], en tant qu'elle purifie et soigne la peau, nourrit et teint les ongles et les cheveux.
>
> *bétoine* (Médecine) : **Cette herbe***, en tant qu'elle possède de nombreuses vertues médicinales, et notamment favorise la conception, améliore la vue, soigne les douleurs et les blessures à la tête, favorise la guérison des plaies, traite les écrouelles ou la goutte froide.
>
> *acteur* (Médecine) : Auteur d'un traité de médecine.

Quelques clauses définitionnelles

Les contraintes représentationnelles qui pèsent sur la métalangue de description rendent inévitable l'adoption de formules définitionnelles, lesquelles fonctionnent comme des constructions logico-syntaxiques permettant de traiter des phénomènes identiques, à l'échelle de toute la nomenclature et quels que soient les domaines concernés.

Le démonstratif et la relation anaphorique

Les articles du *DFSM* sont conçus sur le modèle de la polysémie, car le dictionnaire entend privilégier l'unité psycholinguistique du signe, par-delà la contrainte domainière. Se pose dès lors la question du principe de classification des sens. Plusieurs axes peuvent être retenus, privilégiant un classement :

---

27. Ce serait en propre la définition du lexique, aux yeux du terminologue : être un ensemble d'unités non classables dans un domaine, être à soi un non domaine. Nous avons déjà pris position sur ce point (Petit, 1995, 2010).
28. Définition anaphorique, voir plus bas.

(1) en fonction de la fréquence du sens dans le corpus ;

(2) historique croissant ;

(3) par ordre alphabétique de domaine ;

(4) par génération de sens, notamment dans le cadre d'une polysémie systématique ;

(5) par types d'emplois : libres *vs.* contraints, au plan syntaxique ou morphosyntaxique.

Opter pour un modèle unique s'avère une solution peu adaptée compte tenu de l'hétérogénéité du matériau à traiter. Pour cette raison, la configuration sémantique disponible pour chaque unité indique l'axe à privilégier. L'observation des données montre néanmoins que la polysémie systématique fournit un vecteur de classement récurrent, tous domaines confondus. Nombre d'articles présentent donc des définitions anaphoriques, fondées sur l'emploi du démonstratif. Ainsi à l'article *bette* (nous soulignons) :

> Herbe* qui pousse en s'étalant au sol, dont les feuilles ressemblent à celles du chou et qui présente une grosse racine.
>
> **Cette herbe** en tant qu'elle est cultivée pour ses feuilles dont se nourrissent les hommes.
>
> Les feuilles **de cette herbe*** en tant qu'aliment à privilégier dans le régime de santé, en particulier des hommes de complexion chaude.
> Les feuilles de **cette herbe***, en tant qu'elles peuvent servir pour confectionner un emplâtre.
> La partie dure des feuilles **de cette herbe***, côte de bette.

La détermination syntagmatique : « en parlant de »

Sont concernés les adjectifs et noms dont l'emploi est syntagmatiquement contraint. Ils fonctionnent de manière stable et récurrente comme expansions (arguments appropriés, pour le lexique-grammaire) d'un nom ou d'un paradigme de noms présentant un profil sémantique commun. Le tour « en parlant de » a pour fonction d'exprimer la solidarité syntagmatique entre l'entrée et l'unité terminologique à laquelle elle est liée (nous soulignons) :

commixtion : Union par mélange, **en parlant de** plusieurs ingrédients qui entrent dans une préparation médicale.

L'expression des propriétés extrinsèques : « en tant que »

Cette clause définitionnelle a pour fonction de justifier un dégroupement polysémique sur la base d'une acquisition de PE. Elle est utilisée pour marquer l'articulation d'un concept doté de PE spécifiques sur un autre qui en est *a priori* dépourvu – nom d'espèces naturelles, d'organes, de parties du corps (nous soulignons[29]) :

anis : Herbe* qui ressemble au fenouil.
Cette herbe* **en tant qu'**elle est cultivée.

L'expression de l'hyperonymie fonctionnelle : « tout »

La structure d'une définition logique impose que soit renseigné l'hyperonyme du défini, ou du moins un superordonné si celui-là n'est pas disponible. Lorsque l'accrochage hiérarchique entre concepts s'opère uniquement sur la possession de PE (et non pas d'un complexe de PI et de PE), l'hyperonyme est alors un nœud fonctionnel à l'extension vaste et privé de rattachement catégoriel précis. Il peut alors être occupé par une dénomination occurrentielle (Petit, 2009), pour peu que ses PE coïncident et fournissent un site d'accrochage valide. Au plan définitionnel, le marqueur *tout* vient manifester l'existence de ce nœud fonctionnel en précisant le choix opéré par le dictionnaire pour le saturer (nous soulignons) :

apostume : **Toute** enflure, grosseur causée par une corruption* des humeurs*.

L'hyponymie : « une des N de »

Dans un registre opposé, la glose *une des* N [*espèce*, *forme*] *de* indique un rattachement hyponymique en précisant la nature du lien. Ce type de précision a pour effet de marquer la spécificité de l'entrée dans le paradigme de ses cohyponymes :

29. Voir également les définitions 2, 3 et 4 de *bette*, plus haut.

> *ascite* : **L'une des formes d'**hydropisie, caractérisée par son origine essentiellement aqueuse ainsi que par sa localisation au niveau du ventre, et considérée comme la pire.

Conclure un si bref exposé serait une gageure. Nous avons laissé en chemin de nombreuses pistes dont nous évoquerons les deux principales, promises pour un exposé futur :

– l'identification des domaines de connaissance et le rattachement des termes susceptibles de les valider ;
– le traitement informatisé du dictionnaire et son influence sur la présentation de l'information, son codage et l'exploitation heuristique des données ;
– le contrôle de la métalangue de description dans une double perspective : (i) celle de l'harmonisation de la représentation sémantique des données de T-1 avec celles présentes en To ; (ii) celle du codage lexicosyntaxique de l'information en vue de l'exploitation heuristique (et informatisée) des données.

Néanmoins, les quelques réflexions menées ici montrent que la terminographie médiévale reste largement à construire, *a fortiori* si elle débouche sur une exploitation informatique (ce qui est le lot de tous les référentiels terminologiques actuels). Les remarques formulées dans ces pages n'engagent pas seulement le plan méthodologique de la description. En amont, elles mettent en cause la perspective sémiotique à projeter sur la terminologie diachronique, et en particulier sur celle du Moyen Âge.

Références bibliographiques

Béjoint, Henri et Thoiron, Philippe (dir.), *Le Sens en terminologie*, Lyon, Presses universitaires de Lyon, 2000.

Binon, Jean, Verlinde, Serge, Bertels, Ann et Selvat, Thierry, *DAFA = Dictionnaire d'apprentissage du français des affaires*. En ligne : http://www.projetdafa.net/.

Cabré, Maria-Teresa, *La Terminologie. Théorie, méthode et applications*, trad. J. Humbley et M. Cormier, Paris, Armand Colin/ Presses de l'université d'Ottawa, 1998.

Cadiot, Pierre, « Propriétés extrinsèques en sémantique lexicale », *Journal of French Language Studies*, n° 7, 1997/2, p. 127-146.

Cadiot, Pierre et Lebas, Franck (dir.), « La constitution extrinsèque du référent », *Langages*, n° 150, 2003.

Campenhoudt, Marc, « Le terme : condensation syntaxique et condensation des connaissances en langue spécialisée », 2010. En ligne : http://www.termisti.org/romanica_w.pdf.

Candel, Danielle, « La représentation par domaines des emplois scientifiques et techniques dans quelques dictionnaires de langue », *Langue française*, n° 43, 1979.

Condamines, Anne, Rebeyrolle, Josette et Soubeille, Annie, « Variation de la terminologie dans le temps : une méthode linguistique pour mesurer l'évolution de la connaissance en corpus », dans *Actes du colloque Euralex International Congress*, Lorient, Université de Lorient, 2004, p. 547-557.

CréaLScience, *Dictionnaire du français scientifique médiéval*. En ligne : http://www.crealscience.fr/.

Jacquart, Danielle et Thomasset, Claude (dir.), *Lexique de la langue scientifique (astrologie, mathématiques, médecine…) : Matériaux pour le « Dictionnaire du moyen français » (DMF)*, avec la collab. de Sylvie Bazin-Tacchella, Jean-Patrice Boudet, Thérèse Charmasson, Joëlle Ducos, Hervé L'Huillier (INALF-CNRS), Paris, Klincksieck, 1997.

De Bessé, Bruno, « Le domaine », dans Béjoint, Henri et Thoiron, Philippe (dir.), *Le Sens en terminologie*, Lyon, Presses universitaires de Lyon, 2000, p. 182-197.

DEAF = Dictionnaire étymologique de l'ancien français. En ligne : http://www.deaf-page.de/fr/.

DMF = Dictionnaire du moyen français. En ligne : http://www.atilf.fr/dmf/.

Dury, Pascaline, « Les noms du pétrole : une approche diachronique de la métonymie onomastique », 2008. En ligne : lexis.univ-lyon3.fr/IMG/pdf/Lexis_1_Dury.pdf.

GODEFROY, Frédéric, *Dictionnaire de l'ancienne langue française et de tous ses dialectes, du IXe au XVe siècle*, Paris, F. Vieweg, 1881. En ligne : http://micmap.org/dicfro/.

HUMBLEY, John, « La terminologie française du commerce électronique, ou comment faire du neuf avec de l'ancien. Vers une géomorphologie lexicale », dans *Terminologie et plurilinguisme dans l'économie internationale. Actes de la Ve Journée scientifique de REALITER*, Milan, 2009. En ligne : http://realiter.net/spip.php?article1847.

L'HOMME, Marie-Claude, *La Terminologie : principes et techniques*, Montréal, Presses de l'université de Montréal, 2004.

OTMAN, Gabriel, *Les Représentations sémantiques en terminologie*, Paris, Masson, 1996.

PETIT, Gérard, « Le traitement des variantes graphiques dans les dictionnaires Larousse et spécifiquement dans *Le Petit Larousse illustré* », dans *Langue française. La variation graphique et les rectifications de l'orthographe française (1990)*, Paris, Larousse, 1995, p. 40-51.

—, *La Dénomination. Approches lexicologique et terminologique*, Louvain, Peeters, coll. « Bibliothèque de l'Information grammaticale », 2010.

PICTON, Aurélie, *Diachronie en langue de spécialité. Définition d'une méthode linguistique outillée pour repérer l'évolution des connaissances en corpus. Un exemple appliqué au domaine spatial*, thèse de doctorat, Université de Toulouse 2, 2009.

REY-DEBOVE, Josette, *Étude linguistique et sémiotique des dictionnaires français contemporains*, Den Haag, Mouton, 1971.

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

Ramon Masià
Universitat Oberta de Catalunya

À peu près cent ouvrages de mathématique grecque écrits dans l'intervalle temporel du IVe siècle av. J.-C. au VIIe siècle apr. J.-C. sont parvenus jusqu'à nous. Ce sont des textes essentiellement rédigés en langue grecque : on n'y trouve pas de langage formulaire moderne (équations, intégrales, signes de racines, etc.). En fait, il est possible de définir la mathématique grecque comme un genre littéraire ancien, à l'instar de l'épique ou du comique, tant les traits stylistiques des textes mathématiques sont très bien définis et réguliers. En outre, les Anciens classaient ce type de textes par leur style, davantage que par leur contenu. C'est la raison pour laquelle il est très intéressant, et en réalité crucial, de connaître la langue des mathématiques grecques pour prétendre connaître la discipline elle-même. C'est d'ailleurs là une tâche réalisable, parce que le corpus mathématique grec est un corpus que l'on peut traiter, du point de vue numérique, avec des ressources informatiques et humaines limitées.

Il existe des projets pour le traitement numérique des textes grecs (le projet Perseus par ex., qui est le plus ambitieux), mais aucun d'eux ne se donne pour objectif d'analyser la langue de la science grecque. Notre travail représente le premier pas vers le traitement numérique de ce type de corpus, dans le cas des textes mathématiques.

Le Corpus des textes mathématiques grecs (CTMG)

Le Corpus des textes mathématiques grecs (CTMG) contient une centaine d'ouvrages, et l'intervalle temporel qu'il couvre va du IVe siècle av. J.-C. au VIIe siècle apr. J.-C. Ces textes ne recourent pas au langage formulaire moderne (équations, intégrales, signes de racines, etc.), mais présentent deux éléments originaux, que nous ne trouvons dans presque aucun autre ouvrage en grec ancien : des *lettres dénotatives* et des *diagrammes*. Les objets mathématiques sont usuellement codés par un groupe de lettres de l'alphabet grec (par ex. « un carré AB », où le groupe « AB » désigne le carré mentionné). Ces groupes de lettres dénotatives apparaissent dans le texte, mais aussi dans les diagrammes. Toutefois si on néglige ces deux éléments, lettres dénotatives et diagrammes, on peut dire que les mathématiques grecques sont écrites avec la langue usuelle des Grecs, comme tous les autres ouvrages grecs, qu'ils relèvent des genres épique, historique, médical, etc.

Les œuvres grecques sont segmentées selon leur genre littéraire : chacun de ces genres appelle un style spécifique, et chaque ouvrage relève spécifiquement de l'un d'entre eux ; il n'y a pas, en général, de mélanges de style. Les traits stylistiques du CTMG sont également bien définis, et c'est pourquoi on peut affirmer que la mathématique grecque présente un style clairement distingué ; les Anciens reconnaissaient un texte mathématique à son style, davantage qu'à son contenu – d'où la nécessité d'en connaître les traits stylistiques, ce qui ne représente pas de difficulté majeure en raison de la facilité attachée, aujourd'hui, au traitement informatique de ce corpus. En effet :

– le nombre d'ouvrages n'est pas si grand (environ une centaine),
– le nombre d'*occurrences* dans ces textes est « raisonnable » (à peu près 2,5 millions d'occurrences),

– le nombre de *lemmes* dans le corpus est bas (on pense qu'il ne comporte pas plus de 6 000 lemmes),
– la langue mathématique est quasiment dépourvue d'ambiguïtés,
– les structures syntaxiques de base sont réduites et ne présentent que peu de variantes.

Enfin, la langue des mathématiques grecques est plus répétitive et moins polysémique que la langue usuelle et, en conséquence, elle est plus aisée à traiter par des moyens informatiques. Nous présenterons ici les grands axes de nos travaux de numérisation et d'analyse de corpus.

La méthodologie

Notre recherche adopte les procédures de la linguistique de corpus (*Corpus Linguistics*), méthodologie de travail portant sur des *corpus* numérisés et présentant deux caractéristiques principales[1] :

– Les *corpus* numérisés contiennent, essentiellement, deux types d'informations ou de données : d'une part le(s) texte(s) qu'on veut analyser, et d'autre part les métadonnées associées à ces textes.
– L'objectif est d'extraire des informations concernant l'emploi et la structure de la langue représentée dans le corpus.

On utilise, aussi, des outils statistiques permettant de réaliser des analyses stylométriques.

La technique de base de notre méthodologie repose sur la lemmatisation, c'est-à-dire une procédure permettant de déterminer les occurrences, les formes et les lemmes de tous les textes du corpus :

– chaque mot du texte (chaîne de lettres de l'alphabet grec) constitue une occurrence (*token*) ;
– toutes les occurrences identiques du texte constituent une forme (*form*) ;

1. Voir Garside, Leech et McEcery (1997).

– l'ensemble des formes du même terme du lexique constitue un lemme. Le CTMG contient seulement un type de lemme qui n'appartient pas à strictement parler au lexique de la langue grecque : les lettres dénotatives.

Cette procédure de lemmatisation est semi-automatique : la première fois qu'apparaît une forme, il faut l'analyser (ou la valider) à la main. Ensuite, on la trouve automatiquement, parce que le corpus ne présente presque pas d'ambiguïté. Avec cette procédure semi-automatique, on peut garantir que les résultats de l'analyse ne contiendront pas beaucoup d'erreurs, tous les lemmes étant validés à la main, sans pour autant que le temps nécessaire à sa réalisation ne soit trop important, parce que le corpus contient un nombre « raisonnable » de lemmes. À ce stade de la recherche, l'annotation avec métadonnées est très simple, comme nous le verrons.

Problématique du CTMG

Plusieurs éléments problématiques du CTMG doivent être analysés et éclaircis avant de procéder à l'analyse du corpus. Il a fallu prendre des décisions préalables en ce qui les concerne. Il faut dire, d'ailleurs, que cette tâche inaugurale de résolution s'est révélée la plus longue et la plus pénible de notre travail. On peut classer ces éléments en deux catégories : les problèmes d'encodage d'un côté, et de l'autre les problèmes de marquage.

Problèmes d'encodage

L'encodage des textes n'est pas en soi un préalable à l'analyse. Les textes du CTMG ne sont pas tous encodés, et les textes encodés ne le sont pas toujours convenablement ni sous une forme homogène. En général, il est possible de distinguer, du point de vue de la numérisation :

– les éditions imprimées, critiques ou commentées : presque tous les ouvrages de mathématique grecque font l'objet d'une édition imprimée ;
– les éditions du *Thesaurus Linguae Graecae* (TLG) : la base de données numérique du TLG, la plus ancienne base de données

de textes grecs, contient presque tous les textes importants de l'Antiquité, également pour la mathématique grecque[2].

À partir de ces sources, nous avons commencé à créer une base de données d'ouvrages mathématiques « bien encodés ». Le TLG contient à peu près 80 % des ouvrages mathématiques imprimés, mais beaucoup de problèmes de codage y subsistent, parce que ce dernier se montre parfois aléatoire et pas toujours homogène. Il faut, donc, revoir l'encodage de ces textes et l'homogénéiser afin de le rendre utile dans la perspective du traitement stylométrique : il convient de réintégrer des mots, des abréviations, de ré-encoder les symboles, etc.

La conversion d'une édition imprimée en un texte numérisé et utilisable pour le traitement stylométrique est plus difficile encore à opérer : les procédures d'OCR (*Optical Character Recognition*) pour les textes grecs anciens commettent beaucoup d'erreurs, même si elles ont été constamment améliorées (voir Boschetti *et al.*).

Enfin, la correction (dans le cas du TLG) et la numérisation *via* OCR (dans le cas de textes imprimés) sont des tâches qui nécessitent des ressources humaines considérables, parce qu'elles sont en grande partie réalisées à la main ; en réalité, c'est l'opération qui demande le plus de temps.

Problèmes d'annotation

L'annotation de base est la lemmatisation : concrètement, il faut décider le lemme de chaque forme du texte. Mais il faut effectuer un balisage plus complet, et décider où il convient de s'arrêter. Deux types d'annotations sont nécessaires :

– L'annotation de la macro-structure du CTMG. Il faut introduire la division macroscopique des textes grecs, avec les données suivantes : auteur, ouvrage, livre, préface en forme d'épître, introduction avec définitions, propositions (avec leurs différentes parties : énoncé, exposition,

2. La mathématique grecque inclut des disciplines comme l'astronomie, la musique, la mécanique et d'autres, considérées comme relevant d'elle par les Grecs de l'Antiquité.

détermination, construction, démonstration, conclusion) et les démonstrations alternatives.

– L'annotation des micro-structures, concrètement la structure syntaxique, la structure mathématique et la structure logique.

Ces types de marquage sont encore très réduits, mais nous sommes en train d'accélérer le processus grâce au recours à des outils de marquage automatique (dont le *Natural Language Toolkit* de Python).

Résultats

Nous l'avons signalé, deux processus doivent être réalisés successivement : l'encodage/lemmatisation et l'analyse des textes codifiés.

L'encodage / lemmatisation

Nous avons encodé et lemmatisé intégralement les textes des auteurs suivants : Archimède, Apollonius, Diophante, Domninus, Pappus, Serenus et Théodose. En outre, nous avons encodé les textes qui suivent : les *Éléments* et les *Données* d'Euclide, les *Metrica* d'Héron d'Alexandrie, et les *Prolégomènes à l'Almageste*.

L'état actuel des données de base du corpus est le suivant :

– 733 003 occurrences sur un total de 2,5 millions dans le CTMG,
– 17 618 formes,
– 3 181 lemmes.

Le très petit nombre de lemmes est très remarquable, dans une partie aussi importante du corpus. En outre, il est vraisemblable que le total des lemmes dans le CTMG soit inférieur à 6 000.

Dans le tableau suivant se trouvent les données relatives aux 10 lemmes les plus fréquents. Il est remarquable de constater que près de 55 % des occurrences des textes lemmatisés appartiennent à un lemme de cette liste, c'est-à-dire que plus d'une occurrence sur deux d'un texte mathématique grec s'y trouve. Aucun autre texte en grec ancien ne présente cette concentration d'occurrences en un si petit groupe de lemmes.

Dans cette liste, on trouve l'article défini, avec plus de 20 % des occurrences, les lettres dénotatives, trois conjonctions, trois prépositions, les adjectifs numéraux, et, finalement, le verbe *être*, le seul représentant des catégories qu'on peut considérer comme sémantiquement pleines[3].

Tableau. 1. Les dix lemmes les plus fréquents

Lemme	733 003	%	% cumulé
o/article	158 904	21,68	21,68
lettre dénotative	104 893	14,31	35,99
και/et	29 625	4,04	40,03
ειμι/être	29 381	4,01	44,04
προς/préposition	17 592	2,40	46,44
δε/conjonction	13 959	1,90	48,34
nombre	13 534	1,85	50,19
αρα/par conséquent	11 916	1,63	51,81
αρο/préposition	11 003	1,50	53,32
υπο/préposition	10 626	1,45	54,77

Analyse des textes

Afin de parvenir à une analyse plus approfondie, nous avons procédé à une annotation plus précise pour l'œuvre intégrale de quelques auteurs et quelques livres isolés. Les auteurs traités intégralement sont Archimède et Apollonius, et les ouvrages isolés sont les *Éléments* et les *Données* d'Euclide ainsi que les *Metrica* d'Héron d'Alexandrie. Les données de base de ce corpus sont les suivantes :

– il contient 369 485 occurrences,
– le nombre de formes s'élève à 8 208,
– le nombre de lemmes est de 1 318.

Le balisage complémentaire inclut la catégorie grammaticale (presque 50 % des lemmes de cette partie du corpus ont déjà été marqués avec leur catégorie grammaticale) et les parties du texte

3. Il faut dire aussi que la fréquence de ce verbe est très élevée : elle représente plus de 4 % des occurrences (par ex. la fréquence de ce verbe dans Platon, l'auteur ancien qui utilise le plus le verbe être, est à peu près de 3 %). Tous les accents diacritiques en grec ont été éliminés dans notre texte et dans les tables pour faciliter l'édition.

(jusqu'à Proposition). Grâce à ce marquage, nous pouvons procéder à plusieurs types d'analyses : description statistique du lexique du corpus, analyses comparatives entre parties du corpus, analyse structurelle (syntaxique, logique et des unités mathématiques).

Nous montrerons maintenant quelques résultats intéressants parmi ceux obtenus grâce à ce traitement.

La langue d'Archimède vs. *la langue commune dans le corpus marqué*

Treize livres d'Archimède ont survécu et nous les avons tous numérisés, lemmatisés et annotés. Nous avons procédé à une comparaison de la langue d'Archimède avec celle de tous les textes du corpus encodé, lemmatisé et annoté (soit 36 livres). Pour faire une première comparaison, nous choisissons les lemmes communs de la langue d'Archimède et ceux de la langue du corpus déjà encodé.

Il y a 27 lemmes communs dans le lexique d'Archimède. Cette catégorie représente 3,41 % de l'ensemble des lemmes qu'il contient (761 lemmes au total), mais seulement 1,79 % des lemmes de la langue du corpus encodé sont communs à tout le corpus (1 226 lemmes). Ces lemmes représentent respectivement 62,71 % des occurrences chez Archimède (97 876 occurrences) et 64,20 % des occurrences dans le corpus (234 897 occurrences). Il faut noter que la proportion d'occurrences des lemmes communs aux œuvres d'Archimède est presque égale au nombre d'occurrences des lemmes communs aux ouvrages du corpus.

Si on analyse le type et les fonctions du lexique commun dans chaque cas, on peut noter que :

– le pourcentage d'éléments anaphoriques/déterminants (articles, pronoms, etc.) est très semblable dans la langue commune d'Archimède et dans la langue commune du corpus codifié (à peu près 40 %) ;
– les éléments de formation de *formulae*[4] mathématiques sont à la hauteur de 5,23 % dans la langue d'Archimède et de 5,57 %

4. Au sens du terme *formula* introduit par Reviel Netz (dans Netz, 2003, chapitre IV).

dans la langue commune du corpus encodé, c'est-à-dire une différence de 6,11 % ;

– il y a 9,03 % d'éléments logiques dans la langue d'Archimède et 10,31 % dans la langue commune, soit une différence de 12,45 % ;
– il y a dans la langue d'Archimède davantage d'éléments sémantiques pleins que dans la langue commune. On compte dans la langue commune d'Archimède quatre verbes (*être, avoir, conduire, couper*), deux tours relationnels (*égal à* et *plus grand que*), un substantif (*droite*), un adjectif (*ce qui reste*), alors que, dans la langue commune du corpus, il y a deux verbes (*être et avoir*) et deux tours relationnels (*être égal à* et *être plus grand que*).

Avec ces données, il semble évident que la langue d'Archimède a une articulation logique plus faible (moins d'éléments logiques) et que son style est moins formulaire (moins d'éléments de formation de *formulae* mathématiques et davantage d'éléments sémantiques pleins). Si on a lu Archimède et Euclide, ces conclusions semblent plausibles, mais il serait impossible de les obtenir au moyen d'un relevé manuel.

Quelques analyses comparatives plus complexes

Des outils existent qui autorisent des analyses comparatives plus approfondies basées sur la recherche des mots clefs (*keywords*), le calcul de la *keyness*, et les mesures de proximité lexicale.

La *keyness* d'un lemme, nommé *keyword*, dans une partie du corpus mesure l'importance de ce lemme dans cette partie du corpus par rapport à son importance dans une autre partie du corpus. La table qui suit présente par exemple la *positive keyness* des lemmes présents chez Archimède par rapport à ceux contenus dans les *Éléments* d'Euclide[5].

5. Nous n'avons pas intégré de diacritiques dans les mots employés pour énoncer le lemme, parce qu'il est plus facile de les traiter informatiquement ainsi et, aussi, parce que de cette manière il est très aisé de les distinguer des différentes formes et occurrences.

Tableau 2. Positive keyness

Freq	Keyness	Lemme
1063	1293.356	κωνος
580	951.424	επιφανεια
956	912.546	τμημα
1566	891.203	εχω
364	687.368	βαρος
513	684.329	αξων
468	615.457	τομη
481	570.49	σχημα

Le terme κωνος, « cône », est le lemme le plus caractéristique chez Archimède par rapport aux *Éléments* d'Euclide. Dans la liste des huit lemmes les plus caractéristiques d'Archimède par rapport aux *Éléments* d'Euclide, sept sont des substantifs : επιφανεια (« surface »), τμημα (« segment »), βαρος (« gravité »), αξων (« axe »), τομη (« section ») et σχημα (« figure »). Y figure aussi un verbe, εχω (« avoir »), très caractéristique d'Archimède. Dans la table suivante nous identifions au contraire les lemmes plus caractéristiques des *Éléments* d'Euclide, par rapport au corpus archimédien :

Tableau 3. Negative keyness

Freq	Keyness	Lemme
11906	1619.83	lettre dénotative
644	1267.767	αρα
11	683.935	μετρεω
6	658.03	συμμετρος
2	486.951	ασυμμετρος
118	433.117	αριθμος
161	408.984	γωνια
3352	379.373	ειμι

Euclide utilise beaucoup plus de lettres dénotatives qu'Archimède, et c'est le lemme le plus caractéristique d'Euclide par rapport à Archimède. En second rang nous trouvons la conjonction conclusive αρα, « par conséquent » (ce qui signifie que les propositions d'Archimède sont moins conclusives que les propositions d'Euclide, parce que cette conjonction est utilisée très souvent dans la conclusion d'une proposition mathématique). Il y a, aussi, deux verbes (μετρεω, « mesurer » et ειμι, « être »), deux adjectifs (συμμετρος, « commensurable » et ασυμμετρος, « incommensurable ») et deux substantifs (αριθμος, « nombre » et γωνια, « angle »), qui sont très caractéristiques d'Euclide par rapport à Archimède.

Avec la recherche de *keywords* et de leur *keyness*, on peut définir les grands axes thématiques (verbes et substantifs préférés), préciser l'usage de diverses particules logiques (très importantes dans les textes mathématiques), etc.

Mais on peut calculer d'autres mesures de proximité entre les textes en recourant aux outils stylométriques, car ils nous donnent une idée de la proximité stylistique entre parties du corpus. Différents types d'analyse peuvent être réalisés : *cluster analysis*, *multidimensional scaling* ou *principal component analysis*. Le but principal de ces techniques est de grouper les ouvrages, ou des parties de ces ouvrages, selon leur proximité lexicale, à partir de certains paramètres. Par exemple, le graphique suivant montre une *cluster analysis* des données lemmatiques tirées de quelques ouvrages de quatre auteurs grecs : Euclide (en bleu), Archimède (en vert), Apollonius (en rouge) et Héron d'Alexandrie (en noir).

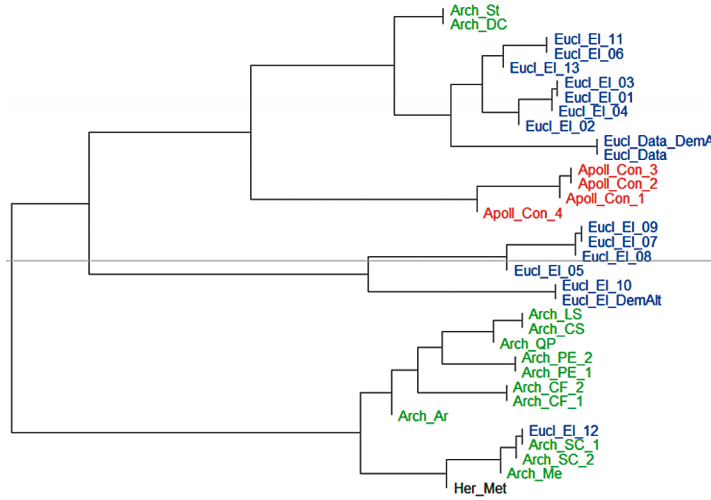


Fig. 1. *Cluster analysis*

Ce graphique montre un regroupement « objectif » de ces textes à partir des données lexicales. En cette phase de notre recherche, nous sommes en train d'évaluer ces outils : il s'agit de savoir si les résultats qu'ils nous offrent coïncident avec ce que nous savons de la mathématique grecque. Le graphique montre bien que les auteurs et les ouvrages constituent des regroupements de forme assez cohérente : tous les textes d'Apollonius sont contigus, presque tous ceux d'Archimède sont dans un même groupe, et aussi presque tous ceux d'Euclide. D'ailleurs, les sous-groupes de textes par auteur qui apparaissent dans le graphique sont aussi très cohérents. Ces faits nous montrent que, à première vue, la classification stylométrique nous offre une information pertinente.

On peut aller un peu plus loin : on peut chercher pourquoi des éléments apparaissent « déplacés » dans le graphique. Par exemple, Eucl_El_12 (Livre XII des *Éléments* d'Euclide) jouxte Arch_SC1, Arch_SC2 (Livres I et II de *Sur la sphère et le cylindre* d'Archimède). On peut expliquer ce fait du point de vue du contenu : ces livres se rapportent à des thématiques semblables et différentes de celles développées dans les autres livres des

Éléments ; en fait, certains chercheurs vont jusqu'à dire que ces œuvres d'Archimède sont la continuation naturelle du Livre XII des *Éléments*. De la même manière, Arch_St et Arch_DC (*Stomachion* et *Dimensio Circuli* d'Archimède) sont très excentriques par rapport aux autres textes d'Archimède, et effectivement, pour des raisons différentes, on les considère comme des ouvrages particuliers, et très différents du corpus central d'Archimède.

On constate, donc, que la classification stylistique résultant de ce type de graphique se reflète plus ou moins dans les caractéristiques du contenu et corrobore certaines intuitions des chercheurs quant aux ouvrages du corpus. En fait, ces résultats étaient prédictibles, parce que, nous l'avons dit, la mathématique grecque est un genre littéraire ancien, et par conséquent le contenu et le style doivent être très liés. On peut donc affirmer qu'en principe ces analyses sont susceptibles de donner des résultats intéressants et suggestifs, qui nous permettront d'étendre les recherches portant sur la mathématique grecque.

Pour finir, signalons que ces analyses ont été réalisées au moyen des logiciels suivants :

– pour la lemmatisation, les logiciels AntConc et AntWordProfiler,
– pour la recherche des *keywords* et de leur *keyness*, le logiciel AntConc également,
– pour les analyses stylistiques, avec *Cluster Analysis*, *Multidimensional Scaling* et *Principal Component Analysis* ; on a utilisé le programme statistique R et RStudio, et plus concrètement le *script* StyloR. On a recouru aussi au logiciel Gephi pour la visualisation et traitement de graphes.

L'écrit mathématique en grec ancien est un genre littéraire de l'Antiquité : on peut donc le définir à partir de ses caractéristiques stylistiques. Mais on peut supposer aussi que les sous-genres de ce genre littéraire doivent présenter des caractéristiques stylistiques communes, et que par ailleurs chaque auteur peut avoir ses particularités stylistiques. Pour ces raisons, il est très intéressant de traiter numériquement le Corpus des textes mathématiques grecs (CTMG), afin d'obtenir les données

stylistiques des ouvrages et les comparer, ce qui nous permettra de mieux connaître la mathématique grecque.

Par ailleurs, le CTMG est un corpus que l'on peut traiter du point de vue numérique : le nombre d'ouvrages est restreint, le nombre de lemmes et d'occurrences est également de faible importance, le corpus ne comporte presque pas d'ambigüités et les structures syntaxiques de base y sont réduites.

Les résultats de nos premières analyses nous confirment que le CTMG est un corpus au lexique très réduit : nous recensons désormais 733 003 occurrences, et seulement 3 181 lemmes. Le lexique est lui aussi très concentré : 7 lemmes concentrent 50 % de toutes les occurrences. Ces caractéristiques ne sont communes avec aucun autre genre littéraire de l'Antiquité.

Les analyses comparatives des textes, utilisant différents outils (*keywords/keyness*, *cluster analysis*, *multidimensional scaling* et *principal component analysis*) nous confirment que des caractéristiques spécifiques peuvent être rattachées à chaque auteur/texte, et que les groupements d'ouvrages obtenus en utilisant ces outils sont cohérents avec ce que nous savons déjà des ouvrages et des auteurs. En outre, on constate bien une connexion entre les caractéristiques du contenu des ouvrages et leur style. Cette confirmation nous permet de conclure que l'utilisation de ces analyses nous permettra de mettre au jour des relations significatives entre textes, ou entre auteurs, inconnues jusqu'à présent.

Références bibliographiques

Textes

Acerbi, Fabio, « I codici stilistici della matematica greca: dimostrazioni, procedure, algoritmi », *Quaderni Urbinati di Cultura Classica*, n° 101, 2012, p. 167-214.

Aujac, Germaine, « Le langage formulaire dans la géométrie grecque », *Revue d'histoire des sciences*, n° 3, 1984/2, p. 97-109.

Bakker, Stéphanie J., *The Noun Phrase in Ancient Greek*, Brill, Leiden, 2009.

Boschetti, Federico *et al.*, « Improving OCR Accuracy For Classical Critical Editions », dans *Research and Advanced Technology for Digital Libraries. 13th European Conference*, ECDL 2009, Berlin/Heidelberg, Springer-Verlag GmbH, 2009, p. 156-167.

Crane, Gregory, « Generating and Parsing Classical Greek », *Literary and Linguistic Computing*, n° 6, 1991/4, p. 243-245.

Deodati, Sara et Kindt, Bastien, « La lemmatisation automatisée des sources en grec ancien : présentation de ressources linguistiques et d'outils de traitement », dans Corino, Elisa, Marello, Caria et Onesti, Cristina (dir.), *Atti del XII Congresso Internazionale di Lessicografia*, Alessandria, Edizioni dell'Orso, 2006, t. II, p. 1137-1143.

Eder, Maciej, « Style-markers in Authorship Attribution. A Cross-Language Study of Authorial Fingerprint », *Studies in Polish Linguistics*, n° 6, 2011, p. 99-114.

—, « Mind Your Corpus: Systematic Errors », *Authorship Attribution. Literary and Linguistic Computing*, n° 28, 2013/4, p. 603-614.

—, « Stylometry, Network Analysis and Latin Literature », *Digital Humanities 2014: Book of Abstracts,* EPFL-UNIL, 2014, p. 457-458.

Garside, Roger, Leech, Geoffrey et McEnery, Tony, *Corpus Annotation: Linguistic Information from Computer Text Corpora*, London/New York, Longman, 1997.

Gelbukh, Alexander et Sidorov, Grigori, *Procesamiento automático del español con enfoque en recursos léxicos grandes*, México, Instituto Politécnico Nacional, 2006.

Manning, Christopher et Schütze, Heinrich, *Foundations of Statistical Natural Language Processing*, Cambridge, MIT Press, 1999.

Mugler, Charles, *Dictionnaire historique de la terminologie géométrique des grecs*, Paris, Klincksieck, 1959.

Netz, Reviel, « Proclus' Division of the Mathematical Proposition into Parts: How and why was it formulated? », *Classical Quaterly*, n° 49, 1999/1, p. 282-303.

—, *The Shaping of Deduction in Greek Mathematics*, Cambridge, Cambridge University Press, 2003.

Vitrac, Bernard, *La Transmission des textes mathématiques grecs*. En ligne : https://www.academia.edu/16162595/La_transmission_des_textes_math%C3%A9matiques_grecs.

Pages web

AntConc : http://www.laurenceanthony.net/software/antconc/

AntWordProfiler : http://www.laurenceanthony.net/software/antwordprofiler/

R : https://www.r-project.org/

RStudio : https://www.rstudio.com/

StyloR : https://sites.google.com/site/computationalstylistics/scripts

TLG Perseus Digital library : http://www.tlg.uci.edu/

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 controverse[1]. 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 démontrer[2]. 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 ans[3]. 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 allemand[4].

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 déchiffrage 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 vrayes », alors qu'il s'agit très simplement de « bastides, douves et brayes [= *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 Galderisi[7] distingue cinq « grands moments » dans ce processus de *translatio* : le moment anglo-normand du XII^e^ 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'hommie 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 sainctes, tant a preuves vraies, comme legiste de la prodommie du noble et vertueux homme » (Vatican Reg lat. 1238, fol. 2r°). 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 livres[8] ».

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 latin[9]. »

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 fautive[10]. 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^ siècle[11]. 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 salutate* », 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 religieux[12], 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 : « B*onus 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 politique[13]. 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 éloquentes que fragmentaires, communauté dont Paul Zumthor a déjà envisagé l'existence[14].

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.

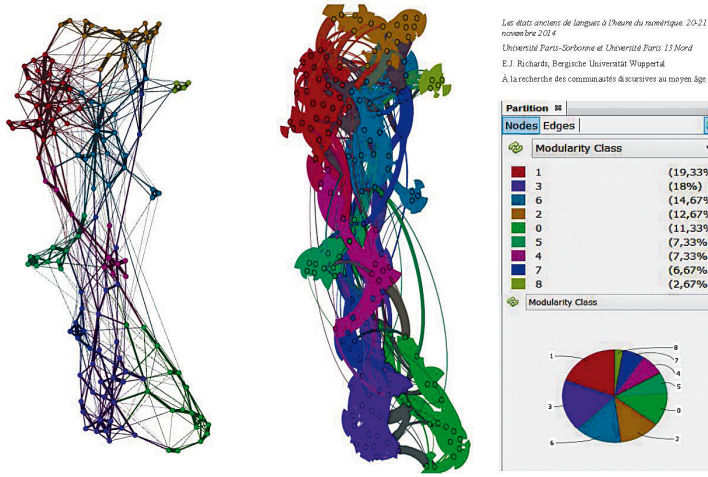


Fig. 1. Les 9 communautés discursives (projection Gephi)

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-italien[15]. 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).

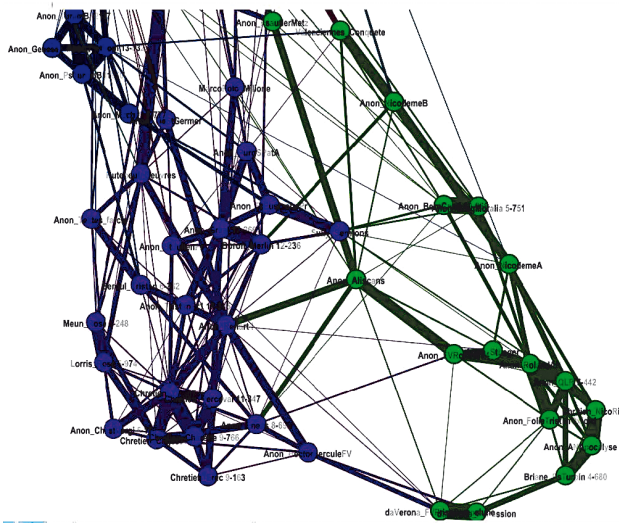


Fig. 2. Communautés nº 0 et 3

Le groupe de couleur violette situé au bas de la figure (nº 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 XIVe 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).

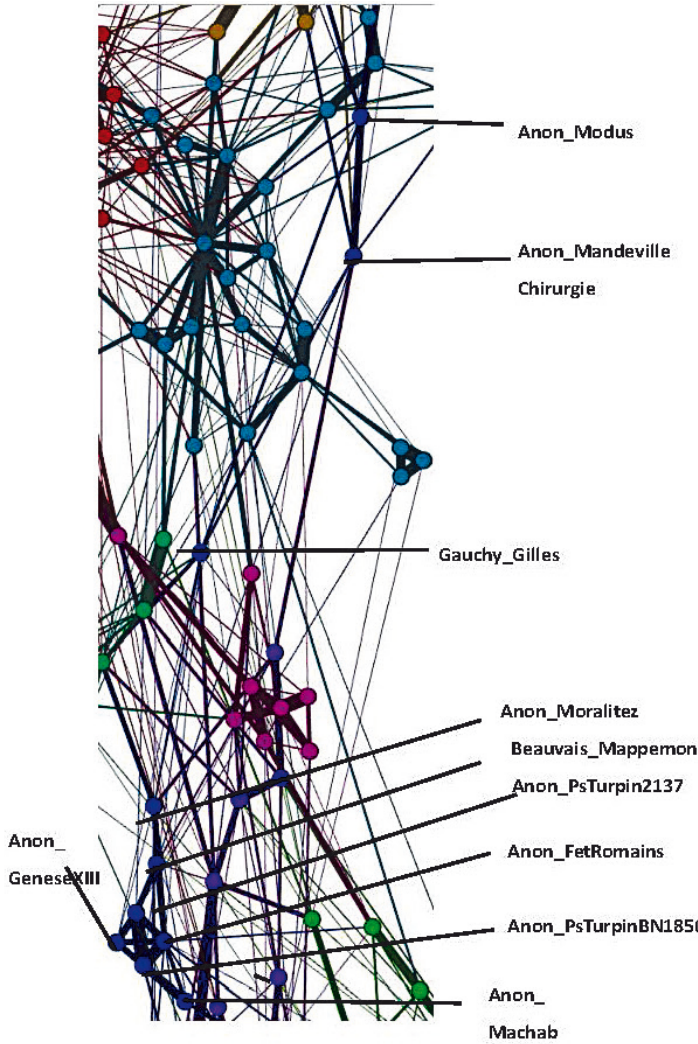


Fig. 3. Communauté n° 7

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

consciemment imité le style des chroniques de la Troisième Croisade ? On a jusqu'ici considéré qu'il s'est inspiré des *Chroniques* de Jean le Bel, puisque Froissart fait allusion à une mise en garde portée par ce dernier contre les chroniques en vers avec leur « grand plenté de parolles controuvées et de redictes pour embelir la rime[17] ». Cette remarque est finalement ironique, parce que la diversité lexicale – *la plenté de parolles* – des *Chroniques* de Froissart est l'une des plus élevées que présentent les œuvres originales. Cela dit, aucun lien stylométrique n'est révélé entre Jean le Bel et Froissart.

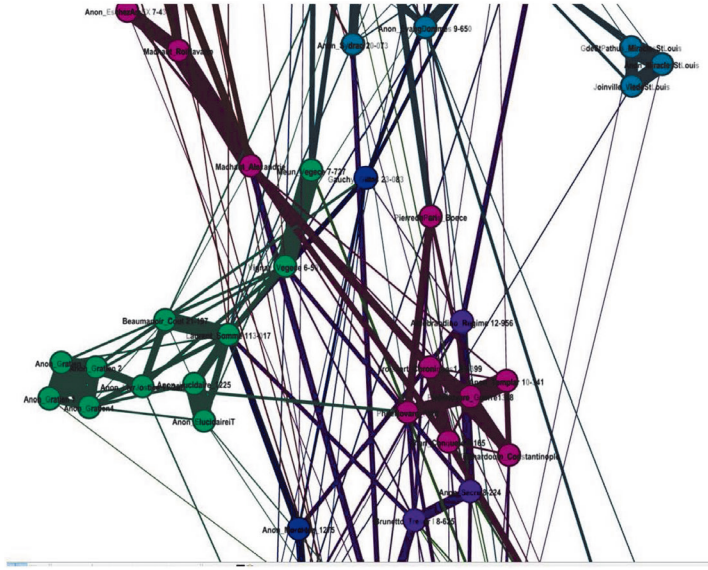


Fig. 4. Communautés n° 4 et 5

Toujours sur la même projection, en bleu-vert (n° 5 – **fig. 4**) se distingue un regroupement de textes légaux, d'un côté, qui comprend aussi deux traductions de Végèce, de Jean de Meun et de Jean de Vignay, et un ensemble de textes théologiques (deux traductions de l'*Elucidarium* d'Honorius Augustodunensis)

17. Voir la discussion que Peter Ainsworth propose de cette question sur « The Online Froissart » (en ligne : http://www.hrionline.ac.uk/onlinefroissart/apparatus.jsp?type=intros&intro=f.intros.PFA-Froissart).

de l'autre. L'écart relativement petit entre les textes dans cette seule communauté semble attester l'existence d'un français « scolastique ». L'écart entre les œuvres individuelles, en effet, ne correspond pas là à l'ordre chronologique dans lequel elles furent composées, un phénomène qui se manifeste ailleurs dans la projection – surtout à travers la présence du *Lancelot en prose*, composé au XIII^e^ siècle, et du *Miroir des simples ames* de Marguerite Porete dans la communauté dominée par les œuvres de Jean Gerson. Il est intéressant que le *Grand Coutumier* de Philippe de Beaumanoir, daté de 1283, soit regroupé ici avec la traduction de Gratien de la fin du XII^e^ siècle, une œuvre qui ne survit d'ailleurs que dans un seul manuscrit. La projection montre seulement la parenté linguistique intracommunautaire, qui semble – et cette conclusion provisoire doit être examinée de plus près – sans relation avec l'ampleur de la diffusion manuscrite des oeuvres concernées.

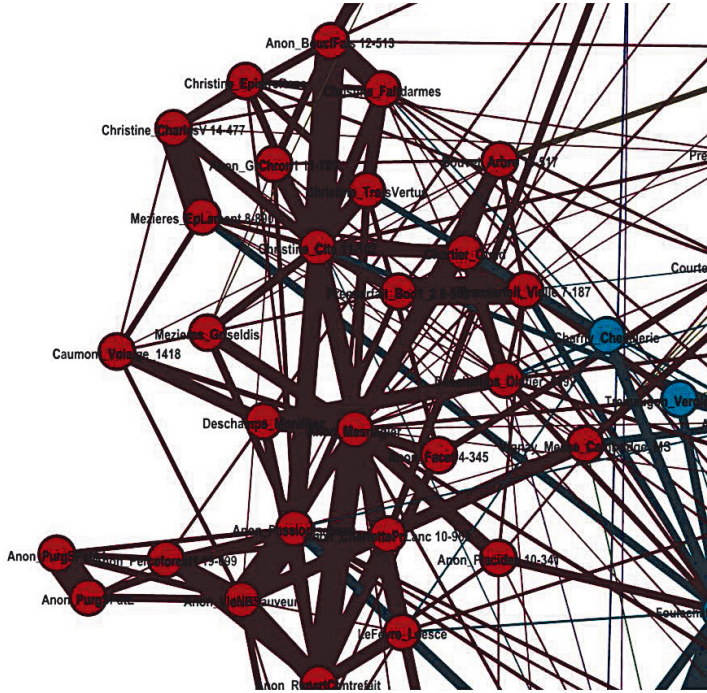


Fig. 5. Communauté n°1

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.

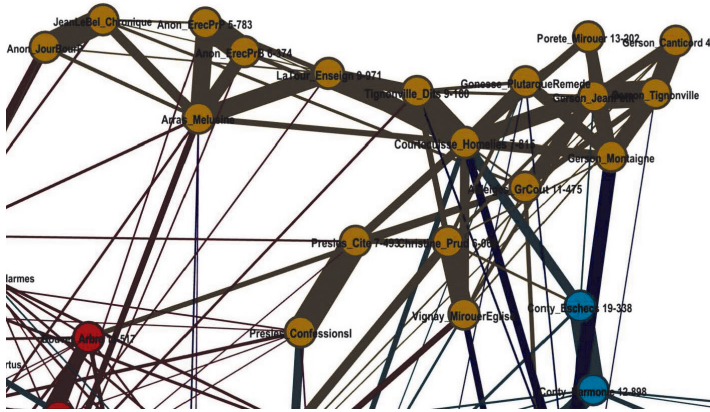


Fig. 7. Communauté n° 2

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, Marilynn (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-TACCHELLA 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 IXe et le XIIe siècle, est réalisé grâce à la collaboration de la section de lexicographie latine du département d'Études médiévales de l'Institut Milá 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.ovi.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.ovi.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 IVe 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 relativamente 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 IXe 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 dictionnairique. 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 dictionnairique 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 CréaLScience 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 THIBAULT (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