

REVUE DE
LINGUISTIQUE
FRANÇAISE
DIACHRONIQUE

7
2017

DIACHRONIQUES

LES ÉTATS ANCIENS
DES LANGUES À L'HEURE
DU NUMÉRIQUE

Martin – 979-10-231-2157-5



LES ÉTATS ANCIENS DES LANGUES À L'HEURE DU NUMÉRIQUE

JOËLLE DUCOS

Présentation

ROBERT MARTIN

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

SYLVIE BAZIN-TACHELLA & GILLES SOUVAY

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

XAVIER-LAURENT SALVADOR, FABRICE ISSAC & MARCO FASCIOLO

Herméneutique des similarités dans le *DFSM* : une expérience

ESTRELLA PÉREZ RODRÍGUEZ

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

ELISA GUADAGNINI

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

ANA GÓMEZ RABAL

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

MICHÈLE GOYENS & CÉLINE SZECEL

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

CÉLINE GUILLOT, SERGE HEIDEN & ALEXEI LAVRENTIEV

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

GÉRARD PETIT

Terminographie diachronique : le cas de la terminologie médiévale française

RAMON MASIÀ

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

EARL JEFFREY RICHARDS

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



LES ÉTATS ANCIENS DES LANGUES
À L'HEURE DU NUMÉRIQUE

Les états anciens
des langues
à l'heure du numérique



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

© Presses de l'université Paris-Sorbonne, 2018

© Sorbonne Université Presses, 2021

Diachroniques n° 7

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

PDF complet – 979-10-231-2155-1

TIRÉS À PART EN PDF :

Ducos – 979-10-231-2156-8

Martin – 979-10-231-2157-5

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

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

Pérez Rodríguez – 979-10-231-2160-5

Guadagnini – 979-10-231-2161-2

Gómez Rabal – 979-10-231-2162-9

Goyens & Szeceł – 979-10-231-2163-6

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

Petit – 979-10-231-2165-0

Masià – 979-10-231-2166-7

Richards – 979-10-231-2167-4

Maquette initiale : Compo-Méca (64990 Mouguerre)

Réalisation : Emmanuel Marc Dubois/3d2s

SUP

Maison de la Recherche

Sorbonne Université

28, rue Serpente

75006 Paris

Tél. (33) 01 53 10 57 60

sup@sorbonne-universite.fr

sup.sorbonne-universite.fr

À 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 inexistantes. 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. *sincoisier* », 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, commodées 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 *avait 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 bénéfiques 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-RUNGALDIER, 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.

Résumés / Abstracts

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

Résumé

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

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

Abstract

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

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

Résumé

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

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

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

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

Abstract

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

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

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

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

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

Résumé

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

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

Abstract

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

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

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

Résumé

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

Abstract

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

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

Résumé

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

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

Abstract

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

Robert MARTIN, À propos du *DMF*

Résumé

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

Abstract

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

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

Résumé

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

Abstract

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

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

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

Résumé

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

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

Abstract

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

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

Résumé

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

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

Abstract

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

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

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

Résumé

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

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

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

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

Abstract

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

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

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

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

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

Résumé

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

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

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

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

Abstract

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

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

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

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

COMITÉ SCIENTIFIQUE

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

COMITÉ DE RÉDACTION

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

COMITÉ ÉDITORIAL

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

Table des matières

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

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