REVUE DE
LINGUISTIQUE
FRANÇAISE
DIACHRONIQUE

2021
8

DIACHRONIQUES

REGARDS LINGUISTIQUES SUR LES ÉDITIONS DE TEXTES MÉDIÉVAUX

Bazin-Tacchella & Souvay – 979-10-231-2174-2

SORBONNE UNIVERSITÉ PRESSES

Regards linguistiques sur les éditions de textes médiévaux

Regards linguistiques sur les éditions de textes médiévaux

SORBONNE UNIVERSITÉ PRESSES
Paris

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

Diachroniques n° 8

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

PDF complet – 979-10-231-2168-1

TIRÉS À PART EN PDF :
Glikman & Verjans – 979-10-231-2169-8
Bragantini-Maillard – 979-10-231-2170-4
Balon – 979-10-231-2171-1
Lavretiev, Guillot-Barbance & Heiden – 979-10-231-2172-8
Mazziotta – 979-10-231-2173-5
Bazin-Tacchella & Souvay – 979-10-231-2174-2

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

https://sup.sorbonne-universite.fr

Lemmatisation et construction automatique de ressources lexicographiques : les développements du lemmatiseur LGeRM

Sylvie Bazin-Tacchella & Gilles Souvay
Université de Lorraine, ATILF/CNRS

Le lemmatiseur hors contexte LGeRM[1] a d'abord été développé par l'ingénieur de recherche Gilles Souvay à l'ATILF dans le cadre du *Dictionnaire du moyen français* (*DMF*). Au départ, il s'agissait de faciliter la rédaction et la consultation du dictionnaire grâce à cet outil. Mais, depuis 2008, LGeRM a connu de nouveaux développements, en raison de son intégration dans d'autres projets. Il paraît intéressant d'en retracer l'histoire et de montrer son utilisation dans l'interrogation des bases textuelles et la constitution de lexiques ou glossaires informatisés.

La création du lemmatiseur LGeRM s'inscrit dans l'histoire du *Dictionnaire du moyen français*, aujourd'hui un dictionnaire électronique en ligne[2]. Au moment de la rédaction de cette publication, sa version 2012 était encore en ligne, mais la nouvelle version en préparation, la version *DMF 2015*, l'a depuis remplacée[3]. C'est le propre de la lexicographie électronique que de fournir des matériaux évolutifs. Mais cette évolution

1. Acronyme de Lemmes, Graphies lemmatisées et Règles Morphologiques. Pour une présentation détaillée du lemmatiseur, se reporter à Gilles Souvay, « LgeRM : un outil d'aide à la lemmatisation du moyen français », dans David Trotter (dir.), *Actes du XXIV^e Congrès international de linguistique et de philologie romanes* (Aberystwyth, août 2004), Tübingen, Niemeyer, 2007, t. I, p. 457-466 et Gilles Souvay et Jean-Marie Pierrel, « LGeRM. Lemmatisation des mots en moyen français », *Traitement automatique des langues*, ATALA, 50, 2009/2, p. 21 *sq*.
2. www.atilf.fr/dmf.
3. Les copies d'écran sont tirées de la version 2012. Même si l'utilisateur accède désormais à la version la plus récente sur le site, il lui est toujours possible de revenir à une version précédente grâce à un onglet « Versions antérieures » sur le bandeau de la page d'accueil.

passe par un développement cohérent et, de ce point de vue, il paraît intéressant de considérer avec attention l'origine et les évolutions de l'outil qui est désormais au cœur de l'ensemble du projet *DMF*.

Si la conception proprement dite d'un dictionnaire portant sur le moyen français remonte à une trentaine d'années[4], un tournant méthodologique et conceptuel important a été pris en 2001 au moment où son promoteur, Robert Martin, a pris la décision d'en faire un dictionnaire totalement et uniquement électronique. En effet, le dessein initial avait été de construire méthodiquement le dictionnaire à partir de la synthèse de lexiques préalables augmentés de « dossiers de mots », réalisés notamment à partir de glossaires d'éditions critiques, lemmatisés dans le « glossaire des glossaires ». Cela correspond à la première période de rassemblement des matériaux (1984-2001). Pour des questions de coût et de durée, il a donc été décidé en 2001 d'abandonner le mode traditionnel d'avancée d'un dictionnaire, lettre par lettre, au profit d'une progression par étapes ; la première étape devait alors consister en une transformation des lexiques déjà disponibles à cette date et des lexiques ultérieurs en une base de données lexicales balisées en langage XML-TEI, de manière à construire progressivement le dictionnaire. Des lemmes avaient alors été créés pour regrouper les articles des différents lexiques portant sur le même mot au moment de la consultation, mais on était encore loin de la mise en place d'une procédure automatique de lemmatisation. Une deuxième période, entre 2007 et 2012, a vu la synthétisation des articles rédigés au cours des étapes précédentes. À partir de la version *DMF 2012*, une nouvelle étape d'enrichissement sélectif du dictionnaire et de développement de ses outils et de ses bases a été inaugurée.

4. L'histoire du *DMF* se greffe sur celle du *Trésor de la langue française* (*TLF*) développé à l'Institut national (aujourd'hui le laboratoire ATILF). Dans sa conception initiale, le *TLF* devait couvrir toute l'histoire de la langue, des plus anciens textes au XX^e^ siècle. Mais comme un tel projet était irréalisable dans les délais impartis, le *TLF* s'est cantonné à la langue des XIX^e^ et XX^e^ siècles et Paul Imbs, fondateur et premier directeur du *TLF*, a souhaité que d'autres dictionnaires prennent le relai pour des périodes antérieures : c'est le cas du *DMF* pour le moyen français. Cette histoire commune explique aussi pourquoi le *DMF* s'appuie sur les mêmes principes lexicographiques que le *TLF*.

Le lemmatiseur LGeRM

La nécessaire gestion des variantes

LGeRM, conçu au départ pour apporter une solution aux difficultés de consultation rencontrées habituellement dans les dictionnaires qui portent sur des états anciens de la langue, devait rendre la consultation du *DMF* plus facile et par conséquent plus conviviale – ce qui est une attente forte pour un dictionnaire électronique. Ces difficultés sont liées au statut de la langue ancienne : en effet, le lecteur de textes médiévaux qui veut s'aider d'un dictionnaire traditionnel est souvent confronté à l'importance de la variation linguistique dans un état de langue non codifié. La langue médiévale est essentiellement variante, à la fois en raison de la transmission manuscrite et d'un système linguistique souple, sur les plans diachronique et diatopique, non normé, ce qui ne signifie pas pour autant aléatoire. Un dictionnaire construit selon les principes traditionnels ne permet pas d'identifier et de regrouper les formes variantes, ce qui rend son utilisation limitée pour le spécialiste et peu utile pour le néophyte.

Sous quelle entrée trouver les formes que l'on peut rencontrer dans un texte médiéval ? La variation peut être seulement graphique, ainsi *tens*/*tans* ; le copiste peut conserver des consonnes qui ne sont plus prononcées, comme *s* intérieure dans *teste*, ou choisir de les insérer pour rappeler l'étymologie, comme la labiale *p* dans *temps* ou *corps*, ou la palatale *c* dans *faict* ; il existe même de véritables équivalences graphiques, telles *c*/*k*, *ss*/*s* ou *ai*/*ei*/*e*, *i*/*y*, souvent liées à l'évolution phonétique, caractéristiques d'une zone géographique et/ou d'une période temporelle. Quelle est alors l'entrée choisie par le dictionnaire ? Celle qui se rapproche le plus du français moderne, celle qui est la plus fréquemment employée dans la période considérée ? Lorsqu'un dictionnaire comme le Godefroy[5] donne la liste

5. Frédéric Godefroy, *Dictionnaire de l'ancienne langue française et de tous les dialectes du IXe siècle au XVe siècle*, Paris, F. Vieweg, 10 vol., 1881-1902. Une version du Godefroy est accessible et interrogeable en ligne à l'adresse suivante : http://www.micmap.org/dicfro. On peut également y accéder à partir des liens présents dans les en-têtes

des formes rencontrées, il faut les retrouver sous une entrée qui n'est pas forcément celle qui pose problème à l'utilisateur. Il existe des renvois, mais ils ne sont pas systématiques.

Une consultation facilitée du dictionnaire

Grâce au lemmatiseur, le *DMF* s'est transformé en un dictionnaire totalement électronique, au-delà de la collecte et de la manipulation des données lexicographiques, notamment des corpus d'exemples. En effet, le lemmatiseur intervient lors de la consultation du dictionnaire. Il permet d'interroger à partir de la forme rencontrée dans un document, sans prérequis ou analyse particulière. 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, ou forme au pluriel lorsque la déclinaison disparaît, n'est pas une entrée du dictionnaire, il faut la chercher sous la forme non marquée[6] *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[7]. Parfois, cela

---

des articles du *DMF*, sous les sigles « GD » (pour « Godefroy ») et « GDC » (pour « Complément Godefroy »). Voici par exemple le bandeau correspondant dans l'article CHERCHER : [TL : *cerchier* ; GD : *cerchier1* ; GDC : *cherchier1* ; *AND* : *cercher* ; *DÉCT* : *cerchier* ; *FEW* II-1, 695a : *circare* ; *TLF* : *chercher*]. On remarque dans la liste que seul l'article du Tobler-Lommatzsch (TL) n'est pas consultable en ligne. Les articles accessibles sont ceux du *Trésor de la langue française* (*TLF*), du *Französisches Etymologisches Wörterbuch* (*FEW*), de l'*Anglo-Norman Dictionary* (*AND*) et du *Dictionnaire électronique de Chrétien de Troyes* (*DÉCT*).

6. On appose en ancien français la forme marquée par un morphème flexionnel, en l'occurrence *-s*, au cas sujet singulier et au cas régime pluriel pour les masculins à une seule base, aux formes non marquées, cas régime singulier et cas sujet pluriel qui présentent la base « nue » ou suivie du morphème zéro. Par convention, c'est la forme non marquée qui sert d'entrée dans les dictionnaires ou les glossaires d'édition. Mais encore faut-il que le lecteur puisse retrouver à partir de la forme marquée qu'il rencontre dans un texte la forme non marquée du dictionnaire ! L'orthographe moderne, *vifs*, à travers l'insertion d'une consonne muette, souligne le lien entre le pluriel et le singulier.

7. On pourrait objecter qu'il en va de même des formes conjuguées modernes, qui présupposent la connaissance du système morphologique verbal. Cependant, en français moderne, s'il subsiste nombre d'exceptions ou d'irrégularités, les variantes se bornent à l'orthographe autorisée et ne touchent plus à la formation des paradigmes

peut sembler plus simple, 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 en ancien et moyen français, selon la coloration dialectale des témoins, ainsi *loi*/*lei*, *bel*/*biel*, *ceval*/*cheval* ou encore *chacier*/*cachier* (latin **captiare*)[8].

Le rédacteur du dictionnaire, confronté aux multiples graphies possibles d'un même terme[9], a dû lui aussi choisir une entrée. Il peut s'agir d'une entrée moderne, lorsque le terme s'est maintenu – c'est l'option qui a été suivie dans le choix du lemme pour le *DMF*, ainsi *segnor* est rangé sous le lemme SEIGNEUR. Mais tous les termes n'ont pas subsisté. Lorsque le mot a disparu, le lemme peut être reconstruit comme une forme moderne acceptable, mais non attestée, comme PLENTÉ pour les formes du substantif *plenté*/*planté*[10] ; ou alors la forme la plus fréquente dans les textes peut être choisie. Avec le lemmatiseur LGeRM, le choix des lemmes, même discutable, permet le regroupement des formes. L'enjeu de la lemmatisation ainsi comprise n'est plus tant d'offrir le seul accès pertinent aux informations contenues dans l'article que de permettre de naviguer d'une forme attestée à l'ensemble des formes auquel elle appartient. La lemmatisation ne se borne pas à une procédure de sélection d'entrées choisies une fois pour toutes, mais est convoquée à chaque interrogation sur une forme.

---

eux-mêmes, comme à l'époque médiévale où étaient attestés, par exemple, jusqu'à trois types de passés simples différents pour le verbe *voloir*.

8. La graphie *ei* pour *oi* est générale dans les textes en anglo-normand ; les autres variantes mentionnées se retrouvent dans les textes picards. Mais les textes peuvent être plus ou moins marqués et certains traits se retrouver isolés dans certaines copies.
9. Ainsi, pour les mots cités précédemment, on relève des formes très variées dans la base textuelle : *lei*, *ley*, *loi*, *lois*, *loix*, *loiz*, *loy*, *loys*, *loyx* ; *ceval*, *cevaulx*, *cevaus*, *chesval*, *cheval*, *chevalx*, *chevau*, *chevaulx*, *chevaulz*, *chevaus*, *chevaux*, *chevax*, *cheveaux*, *quevaulx*. Cependant la fréquence permet de faire le départ entre les formes usuelles et des formes très marginales, voire douteuses.
10. Du latin *plenitate*. Le substantif, usuel en ancien et en moyen français, au sens de « grande quantité ou grand nombre, abondance » (*DMF*), n'a pas subsisté, alors qu'il est passé dans l'anglais *plenty*.

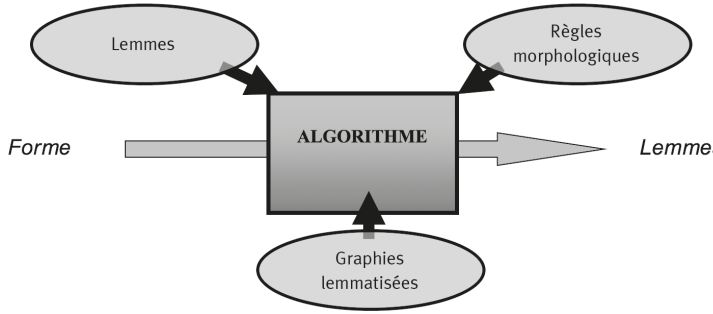
Le lemmatiseur est l'outil qui fait le lien entre les différentes composantes du *DMF* : dictionnaire, lexiques, base de textes et lemmatiseur. L'idée de départ était de permettre, par un simple clic sur un mot présent dans un exemple cité dans un article, d'accéder à l'article correspondant au lemme du mot. Mais il est apparu souvent difficile de savoir quelle entrée avait été retenue par les rédacteurs, si le mot avait subsisté ou non. Le lemmatiseur n'a donc pas permis d'interroger tous les mots des exemples rassemblés dans un article du *DMF*, parce qu'il aurait fallu lemmatiser tous les exemples. Mais il a facilité la consultation du dictionnaire : au lieu de chercher à deviner sous quelle entrée se cache telle ou telle forme, il offre la possibilité d'interroger par la forme rencontrée dans un manuscrit ou un texte édité. Le message qui s'affiche sur le site invite explicitement l'utilisateur à cette démarche : « Saisir un mot ou une forme sans se préoccuper de l'entrée du *DMF* ; des propositions s'afficheront ». Ainsi des formes telles que *destroit*, *destreit* ou *destroict* sont-elles automatiquement renvoyées à deux lemmes de forme moderne, DÉTROIT 1 (adjectif) et DÉTROIT 2 (adjectif et substantif masculin). À l'utilisateur de faire son choix, en fonction du contexte, entre l'adjectif ou le substantif masculin, voire de consulter, en cas d'hésitation, les deux articles du *DMF*. Il aurait pu lui-même saisir directement la forme moderne *détroit*. En revanche, devant une forme comme *menra*, le dictionnaire traditionnel n'est d'aucune aide ; le lecteur doit être familier des structures de la morphologie verbale médiévale ou, du moins, doit trouver des informations dans des ouvrages spécialisés à ce sujet. Dans un cas comme celui-ci, le lemmatiseur propose la réponse en ces termes : « La forme *menra* est connue du lemmatiseur avec l'analyse suivante : MENER, verbe ». Le lemmatiseur a reconnu la forme parmi l'ensemble des formes lemmatisées MENER et le lemme MENER sert alors de lien vers l'article du *DMF*. Lorsque la forme est ambiguë, le lemmatiseur propose plusieurs hypothèses : c'est le cas pour des formes telles que *destroit* ou *porte*[11]. Il peut

11. Pour *destroit*, voir *supra*. Pour *porte*, le lemmatiseur propose trois réponses possibles : PORTE 1, subst. fém. ; PORTE 2, subst. masc. ; PORTER 1, verbe. Si la deuxième hypothèse est peu vraisemblable, car elle correspond à une seule attestation dans les bases avec le

arriver que la forme ne soit pas connue du lemmatiseur, dans ce cas le programme applique des règles, qui sont des procédures formelles, afin de retrouver une forme connue.

Conception de l'outil

Le schéma suivant résume le fonctionnement du lemmatiseur :



1. Architecture du système LGeRM. Gilles Souvay et Jean-Marie Pierrel, « Lemmatisation des mots en moyen français », art. cit., p. 154.

Ce programme a été conçu pour analyser les graphies du moyen français et éviter des manipulations fastidieuses en cas d'introduction de nouveaux mots par le biais des exemples. En effet, la simple projection du contenu de l'ensemble des lexiques sur les textes ayant servi à les constituer s'était avérée insuffisante, dans la mesure où tous les mots des exemples n'avaient pas été traités par les lexiques, qui avaient opéré des sélections sur des critères qui pouvaient d'ailleurs également varier d'un lexique à l'autre ; d'autre part, les mots grammaticaux n'avaient en général pas été traités dans les lexiques. Il y avait donc dans le corpus d'exemples des formes attestées sans lemme. L'outil d'analyse devait permettre de pallier ces manques : constitué d'une base de connaissances comportant les lemmes, les graphies lemmatisées et les règles morphologiques, qu'il

sens de « contribution », la première et la troisième renvoient à des lemmes usuels. Le lemmatiseur formule des hypothèses hors contexte, mais l'utilisateur cherche le sens de mots rencontrés dans un texte et, de ce fait, connaît leur catégorie grammaticale. Parfois cependant, même en contexte, subsiste une hésitation qui ne pourra être levée que par la prise en compte d'un contexte élargi, voire de connaissances extralinguistiques.

utilise pour fournir une liste d'hypothèses de lemmes associés à une graphie donnée, hors contexte. L'objectif est atteint si l'analyse juste est présente dans les hypothèses présentées ; l'outil ne gère pas lui-même les ambiguïtés, c'est le relecteur humain qui valide telle ou telle hypothèse en fonction du contexte d'utilisation de la forme et plus généralement de l'ensemble de connaissances auquel participe le texte de référence. Si la graphie rencontrée n'appartient pas à la base de connaissances, le programme applique les règles morphologiques pour la ramener à une forme connue, si possible à une graphie lemmatisée. Le bruit généré est important, notamment en l'absence de balisage grammatical des formes étudiées. Le lemme retenu pour le *DMF* est une forme moderne lorsqu'elle existe ; dans le cas contraire, la forme choisie est celle reçue par la tradition lexicographique. Parmi les règles morphologiques, un très grand nombre concerne la flexion verbale (environ 3 000) : elles ramènent en général la graphie verbale à un infinitif ou elles modernisent une forme ancienne pour la ramener à une forme connue ; d'autres règles portent sur la flexion nominale et adjectivale (environ 100) : il s'agit notamment de trouver la forme correspondante du masculin singulier ou encore de moderniser une forme ancienne. Les autres règles sont beaucoup plus diverses (environ 400) : modernisation ou vieillissement des graphies, règles de régionalisme, règles phonétiques… Le système permet enfin de prendre en compte des phénomènes de formation de mot, comme le lien entre l'adjectif et l'adverbe en *-ment*.

Élargissement des applications de LGeRM

Intégration dans de grands projets d'édition en ligne

Le lemmisateur LGeRM, conçu au départ pour le *DMF*, a vu ses applications s'élargir. En permettant la lemmatisation de textes de moyen français et, moyennant certaines adaptations, également la lemmatisation de textes des périodes antérieures et postérieures, il s'est transformé en outil au service de l'édition de texte. Ce développement a été favorisé par des collaborations scientifiques extérieures au laboratoire et des expérimentations

internes. La première d'entre elles a joué un rôle véritablement fondateur en rassemblant le *DMF* et des équipes britanniques travaillant sur de grands auteurs de moyen français : celles de l'université d'Édimbourg pour le projet consacré à Christine de Pizan, sous la direction de James Laidlaw, et les universités de Sheffield et Liverpool pour le projet dédié aux *Chroniques* de Froissart, sous la direction de Peter Ainsworth. Ce « *joint project* », intitulé « Medieval Vernacular Dictionaries: principles, methodology, practice, problems and solutions », s'est déroulé sur quatre années, entre 2008 et 2011. Les discussions sur les résultats de la lemmatisation des textes ont permis d'aller plus loin dans l'analyse des problèmes rencontrés et dans la recherche de solutions, notamment grâce à l'encodage des textes et au respect de la TEI. Chaque nouveau projet a offert la possibilité d'affiner le balisage, par exemple pour les noms propres. Ainsi l'utilisation de LGeRM dans d'ambitieux projets d'éditions électroniques a-t-elle permis de l'améliorer et d'en faire peu à peu un outil de construction de glossaire efficace et souple.

L'édition lemmatisée du Réceptaire de Jean Pitart

Parallèlement à notre participation à ces grands projets collaboratifs[12], nous avons lancé la lemmatisation de textes divers du moyen français, littéraires et non littéraires, notamment celle d'un recueil de recettes médicales, traditionnellement connu comme le « réceptaire de Jean Pitart ». Pour évaluer très concrètement les possibilités offertes par LGeRM sur un type de texte différent, d'ampleur limitée, une simple transcription du texte avait été lemmatisée avec l'outil[13]. Dans une première

12. Ces projets, toujours accessibles en ligne, renvoient au *DMF* par des liens sur chacun des mots du texte édité lemmatisé :
    a) Christine de Pizan, *The Making of the Queen's Manuscript*, édition électronique du ms. Harley MS 4431 de la British Library : http://www.pizan.lib.ed.ac.uk/ ;
    b) The Online Froissart, *Édition électronique des Chroniques de Jean Froissart* : http://www.hrionline.ac.uk/onlinefroissart ;
    c) édition électronique en ligne du *Mystère des Actes des Apôtres* de Simon Gréban : http://eserve.org.uk/anr/.
13. Toute l'étude a été menée avec la version 2010 du *DMF*. Voir Sylvie Bazin-Tacchella, « Le "Réceptaire attribué à Jean Pitart" (XIVe siècle) : projet d'une édition et d'un

étape, le texte ne présentait aucune espèce de balisage, sinon des éléments paratextuels habituels dans l'édition des manuscrits, indication de la foliotation et des colonnes, ponctuation et accentuation conformes aux normes en vigueur dans l'édition des textes médiévaux. Il avait été mis au format XML de façon à être lemmatisé avec LGeRM. Une première lemmatisation du texte avait mis à jour toute une série de termes pour lesquels LGeRM n'offrait pas de réponse, regroupés à part dans une « liste des mots inconnus ». Y apparaissaient quelques mots non identifiés par le programme, mais aussi des erreurs de transcription ou de saisie. Ainsi une première lemmatisation permet-elle de détecter et de corriger des erreurs souvent difficiles à déceler malgré les relectures, et de prendre conscience du balisage nécessaire pour que le traitement automatique isole de manière systématique un certain nombre d'éléments. Le texte au format XML TEI a donc été enrichi de nouvelles balises, avant qu'il ne soit procédé à une nouvelle lemmatisation :

<head><num>1.</num> La <rs type="remede">toile maistre <name>Jehan Pitart</name></rs> contre toutes bleceures de jambes et d'autres lieus et en ot la recepte du <rs type=»personne»>roy de <name>France</name></rs>. </head>

<c>P</c>renés oile d'olive <num>.iii.</num> livres ; suif de cerf <num>.i.</num> quarteron ; ceruse poudree, <num>.ii.</num> livres ; serapin, opopanac, armoniac, litarge, mummie, de chascun une once ; stafizagre, <foreign>bdellium</foreign>, ase fetide, orpiment, de chascun demie once ; encens, mastic, mirre, colofome, aloe, cicotrin, de chascun <num>.ii.</num> drames ; <foreign>ypoquistidos</foreign> une drame.

<c>L</c>a maniere de confire est <pb n="3ra"/> tele : mettés vostre oile et vostre suif de cerf sus lent feu tant que le suif soit bien fondu et puis si mettés les poudres, c'est a savoir la poudre de ceruse premierement et mouvés bien tousjours tant qu'il se commence a espoissier et puis y mettés les gommes, c'est a savoir serapin, armoniac, opopanac, qui aient esté par

glossaire électronique », dans Joëlle Ducos (dir.), *Sciences et langues au Moyen Âge / Wissenschaften und Sprachen im Mittelalter*, Heidelberg, Winter, 2012, p. 269-286. Voir également notre présentation détaillée du projet et de ce type de texte (http://www.atilf.fr/dmf/JeanPitart).

<num>.iii.</num> jours temprees en <num>.iii.</num> voirres de bon vin aigre et quant ces gommes seront bien fondues sus le feu, si y mettés toutes les autres choses, ce sunt les autres gommes l'une aprés l'autre a petit feu et laissiés boulir tant que il soit espés et puis ostés du feu et y mettés les poudres chascune par soy delieement poudrees, puis prenés vostre toile et la moilliés dedens et puis la mettés sus une table qui soit ointe d'uile d'olive et soit bien frotee d'une part et d'autre d'une liche si que ele soit bien soueve et laissiés refroidier et la gardés.</p>

Aux balises liées à la structuration du texte (divisions, titres, capitales, changement de feuillet), ont été ajoutées des balises permettant d'isoler :

– les quantités, indiquées par des chiffres romains encadrés de points :

<num>.iii.</num> livres ;

– les termes latins, notamment dans les énumérations d'ingrédients :

<foreign>bdellium</foreign> ;

– des syntagmes considérés

comme des noms composés :

La<rstype="remede">toilemaistre<name>JehanPitart</name>
</rs> ;

ou comme des noms propres :

maistre <name>Jehan Pitart</name>.

Analyse des résultats de la lemmatisation

Après repérage et correction des erreurs, il reste des mots inconnus, c'est-à-dire des termes que le lemmatiseur ne possède pas dans sa base de connaissances et pour lesquels il est incapable de remonter à une forme connue en appliquant des règles. Ces mots inconnus peuvent constituer de véritables apports lexicaux et doivent donc être étudiés à l'aide de la documentation disponible au laboratoire. Il peut s'agir de nouveaux lemmes ou de nouvelles formes correspondant à des lemmes existants. Certaines de ces formes ont pu être déformées

par le copiste ou résulter d'un enchaînement d'erreurs. Dans un certain nombre de cas rencontrés dans le *Réceptaire de Jean Pitart*, les formes, bien que déformées, ont pu être mises en relation avec des formes attestées et être correctement lemmatisées par le programme, ce qui est apparu encourageant pour une utilisation dans le cadre de l'édition électronique comme aide à la lecture des textes ; ainsi les formes *stafizagre*, *colofome* ont-elles été lemmatisées STAPHISAIGRE, COLOPHANE (*DMF*) ; pour la forme *tenoisiee*, le lemmatiseur a proposé deux lemmes : le verbe NOISER et le substantif TANAISIE (*DMF*). C'est le second qui était pertinent.

Notre objectif en lemmatisant le *Réceptaire de Jean Pitart* était de préparer un texte électronique interrogeable en ligne, voire d'aider à la construction d'un glossaire systématique. Grâce aux interactions avec le *DMF*, les apports peuvent être réciproques :

– les éditeurs des textes médiévaux peuvent s'appuyer sur l'ensemble des ressources du *DMF* (le dictionnaire proprement dit, mais également toutes les bases qu'il exploite) ;
– le texte saisi peut offrir de nouvelles données lexicologiques qui enrichiront le *DMF* (compris au sens large).

Pour le premier type d'apport, il convient de distinguer ce qui relève du texte lemmatisé lui-même et ce qui relève du lemmatiseur ; en effet, il s'agit de deux procédures bien distinctes dans le formulaire de levée d'ambiguïté :

– la levée d'ambiguïté au sens strict, qui est opérée manuellement par la sélection d'un des lemmes proposés ou l'ajout d'un lemme nouveau ; elle ne concerne qu'une occurrence donnée dans le texte édité, mais elle peut être élargie aux autres occurrences ou à une partie de celles-ci ;
– l'enrichissement de la base du lemmatiseur par la nouvelle forme ou le nouveau lemme.

Confondre les deux procédures conduirait à limiter les possibilités du lemmatiseur. En d'autres termes, il s'agit de bien distinguer solution individuelle et programmation ; il faut laisser les choix ouverts, quitte à continuer à produire des ambiguïtés toujours en grand nombre – le but étant moins de parvenir toujours à une solution unique que d'obtenir, parmi les choix multiples, la bonne solution. Une solution ponctuelle, liée à un contexte particulier, qui permet la levée d'ambiguïté, ne saurait devenir une règle. En revanche, à travers telle ou telle solution que le lemmatiseur ne proposait pas, on peut formuler une règle qui faisait défaut au programme, notamment en matière d'équivalence graphique.

Les étapes du traitement

Le lemmatiseur applique un certain nombre de règles pour lemmatiser les formes inconnues : lorsque le nombre de règles appliquées ou le nombre de lemmes est élevé, on parle d'« analyses douteuses ». Le programme en donne la liste. L'utilisateur doit passer en revue tous les cas ambigus, dans certains cas valider un des lemmes proposés ou en proposer un autre. L'examen des analyses douteuses ne constitue qu'une première étape qui conduit à de nombreuses levées d'ambiguïtés. Les formes aberrantes sont détectées et corrigées. L'étape suivante sera l'examen des formes pour lesquelles le lemmatiseur propose un ou plusieurs lemmes ; dans le premier cas, il faudra valider ou non la proposition et s'interroger bien évidemment sur les « mauvaises solutions » proposées pour formuler des « parades » ; dans le cas de propositions multiples, l'examen des différentes attestations de la forme données en regard à l'écran devrait permettre de distinguer des homonymes, comme *face* (subst.) et *face* (forme verbale de FAIRE) ou *aux* (subst. pl. d'AIL) et *aux* (forme contractée de la préposition *à* et de l'article défini). Cet examen conduit souvent à établir des filtres pour éliminer des propositions non pertinentes pour le texte considéré. L'éditeur peut consulter en même temps le texte et le dictionnaire, mettre en regard les différentes attestations, ce qui est un gain de temps et

d'efficacité considérable au bénéfice de l'analyse proprement linguistique et sémantique du corpus.

Des matériaux pour des utilisations diverses

Ainsi, pour ce réceptaire, l'utilisation expérimentale de l'outil d'aide à l'édition appuyé sur le lemmatiseur du *DMF* a-t-elle permis d'offrir un texte sûr et interrogeable en ligne[14]. Bien évidemment, il ne s'agit pas d'une édition définitive et exhaustive, il manque un certain nombre d'apports attendus. Cependant, les matériaux qui sont accessibles donnent l'état de la recherche au moment de la mise en ligne ; ils seront complétés ou corrigés au gré de nouveaux apports ; une introduction a ainsi pu être ajoutée plusieurs mois après, offrant au lecteur toutes les informations disponibles sur la transmission du texte et sur ses enjeux. L'essentiel est cependant le fait de permettre d'interroger le lexique par forme, par lemme, en navigant dans le texte et dans le dictionnaire lui-même. Le projet n'a pas conduit à la réalisation d'un glossaire en bonne et due forme, mais tous les matériaux sont disponibles pour le faire et en l'état, il offre la possibilité de récupérer des éléments intéressants pour d'autres projets, par exemple des entrées et des exemples de termes scientifiques pour le projet Crealscience[15], avec lequel le *DMF* a noué une convention permettant un échange de données :

Tableau 1. Extrait des exemples de termes de pharmacopée (lettre C) de Jean Pitart transmis au projet Crealscience

Lemme *DMF*	Forme attestée	Exemples sélectionnés dans le glossaire électronique de Jean Pitart
CALAMINE, subst. minerai de zinc, calamine	***calamine***	« Prenés couperose blanche, **calamine** cuite et en faites deliee poudre et qu'il ait de chascune egalment et mettés cele poudre temprer en belle yaue froide et clere. » 8vb « Et a faire ceste yaue faut a une chopine d'yaue le montant a une nois de noier de coperose et autant de **calamine**. » 9ra

14. http://www.atilf.fr/dmf/JeanPitart.
15. Il s'agit du projet de *Dictionnaire du français scientifique médiéval* mené dans le cadre de l'ANR Crealscience, codirigée par Joëlle Ducos (Sorbonne Université/Paris IV, EA 4509) et Xavier-Laurent Salvador (Paris I, I-DEF), voir http://www.crealscience.fr.

Lemme *DMF*	Forme attestée	Exemples sélectionnés dans le glossaire électronique de Jean Pitart
CAMOMILLE, subst. camomille	***camomille***	« Item prenés jus de **camomille** et oile d'olive et mellés ensemble et en frotés au feu avant l'accés l'eschine du malade et les poux des bras, ce vaut moult. » 9rb « Prenés cire virge demie livre, raisine .i. quarteron, oile d'olive le tiers de demie livre, gabaron, cyroigne, mastic, encens, mirre, therebentine, dyauté, oile de **camomille**, de chascun une once. » 9vb « Baigniés vous en la decoction de foilles de fresne et de mauves et de **camomille**. » 11ra « Prenés **camomille** et triblés et boulés en vin aigre et en lavés souvent le chief, nulle chose n'est meillieur. » 16vb
CAMPHRE, subst. substance aromatique extraite du *laurus camphra* ou d'autres plantes, huile ou substance brute	***camphre canfre***	« Prenés .iv. onces d'oile rosat, cire blanche une once, **camphre** une drame, ceruse once et demie et puis les aubuns de .ii. oex et faites oignement. » 24rb « En teles fievres et en toutes autres eschivés a vostre pooir ire et toutes choses qui peuent le cuer esmouvoir ou eschaufer et tenés diete et mettés paine a dormir et a suer selonc les remedes qui sunt dis par devant pour dormir et pour suer et flairiés roses et violetes et semblables choses, especiaument **canfre** et yaue rose. » 14va

Dans d'autres cas, l'utilisation de l'outil-glossaire a permis la construction du glossaire d'une édition parue sous un format papier. Chaque projet qui utilise l'outil a ses spécificités et les échanges avec nous sont déterminants. Des premiers essais jusqu'à la dernière version, les interfaces de l'outil ont constamment évolué pour s'adapter aux besoins, c'est pourquoi il est difficile d'en présenter une version figée. Nous nous bornerons à présenter ici les différentes étapes du travail avec l'outil d'aide à l'édition dans la version accessible au moment de la rédaction de cet article, et les différentes possibilités offertes à l'éditeur d'un texte en moyen français, ainsi que la richesse des matériaux à la disposition des utilisateurs.

Les possibilités de LGeRM

Mise en place

Des discussions préalables établissent les conditions de la collaboration entre un éditeur de texte médiéval et l'équipe chargée du projet *DMF*. Puis l'éditeur envoie le texte qu'il souhaite lemmatiser au format XML ; une fois la lemmatisation achevée, un mot de passe lui est fourni pour qu'il puisse accéder à son résultat. Sur la page d'accueil du *DMF*, il suffira alors de cliquer sur le menu « LGeRM : l'outil-glossaire » ou, mieux, d'utiliser un lien direct de type www.atilf.fr/dmf/pitart avec le nom du projet en collaboration[16]. Un formulaire s'affiche alors : à l'utilisateur d'indiquer son identifiant et son mot de passe, puisqu'il faut posséder les droits pour accéder au projet.

Toutes les données utiles sont rassemblées en ligne, sous la forme de synthèses accessibles sous le premier onglet du menu d'accueil « Présentation de l'outil-glossaire ». En voici la liste :

Principes généraux
Conditions d'accès
Exemples de réalisation
  Les étapes de la construction (1/3)
  1. Création du projet
  2. Lemmatisation du fichier
  3. Finalisation
  Légende des couleurs
  Gestion des mots grammaticaux
  Présentation du lemmatiseur
  – Principe de fonctionnement
  – Bibliographie sélective
  Prise en charge de la *TEI P5*
  – Généralités
  – Balises prises en charge
  – Segmentation des mots
  Crédits
  Projets en cours

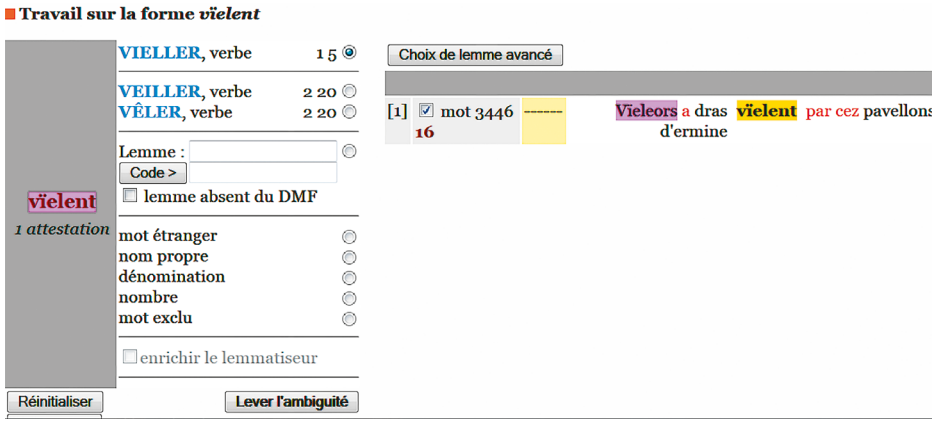
16. http://www.atilf.fr/LGeRM/glossaire.

Procédures

À l'ouverture d'un projet, ce qui s'affiche par défaut est une synthèse brute, une sorte de tableau de bord du projet, avec toutes les données chiffrées après lemmatisation par LGeRM : nombre de lignes, de mots, de ponctuations, de mots connus, de mots non résolus, de formes inconnues, de noms propres, de noms étrangers, de lemmes proposés par le lemmatiseur. La date de la lemmatisation est indiquée, ainsi que les droits d'accès. À droite de chaque information chiffrée, un bouton « Voir » permet d'accéder aux données proprement dites. Il s'agit d'une première étape, comme le rappelle la mention figurant au bas du tableau indiquant que « [l]e glossaire n'a pas été finalisé », à droite de laquelle se trouve un bouton « Glossaire finalisé » qui permettra d'accéder à un nouvel état du glossaire, obtenu après validation des interventions de l'utilisateur. À gauche de la synthèse brute apparaissent les fonctionnalités offertes au rédacteur, qui peut ainsi accéder, par exemple, à différentes listes : « Liste des formes inconnues », « Liste des formes » ou « Liste des lemmes ». Précisons à nouveau qu'il ne s'agit ici que d'une première étape-test, dont les résultats ne peuvent être que provisoires, destinés à être améliorés par la relecture et la vérification du texte, son éventuel amendement, et parfois par un balisage renforcé. Le premier essai offre cependant la possibilité de « dégrossir » considérablement les difficultés rencontrées par le lemmatiseur et, partant, d'améliorer à la fois l'outil – dans ses fonctionnalités notamment – mais aussi le texte en cours d'examen lui-même.

Le texte étant débarrassé de ses erreurs de saisie – notamment grâce aux « mots inconnus » détectés par le lemmatiseur – et mieux préparé pour la lemmatisation, il est à nouveau lemmatisé et les résultats de la lemmatisation systématiquement contrôlés. L'éditeur dispose des différentes listes de formes ou de lemmes indiquées plus haut, qu'il peut parcourir en suivant l'ordre d'apparition dans le texte ou l'ordre alphabétique. Il peut également sélectionner une liste d'« analyses douteuses », c'est-à-dire des formes dont l'analyse implique l'utilisation

d'un nombre important de règles. Par défaut, le paramétrage porte sur l'utilisation de trois règles ou plus, mais l'utilisateur peut le modifier. Mais il peut d'emblée choisir de passer en revue toutes les formes ambiguës, c'est-à-dire pour lesquelles le lemmatiseur présente plusieurs lemmes possibles. Cette phase de « levée d'ambiguïtés », manuelle, demande beaucoup de temps et une très bonne connaissance de l'état de langue concerné. Pour chaque forme ambiguë, un formulaire permet de saisir le choix de l'utilisateur, en sélectionnant par un clic chaque attestation ou certaines seulement :



Cet exemple de formulaire présente une attestation unique dans le texte considéré. Le premier lemme proposé, qui est l'hypothèse pertinente, est doté d'une pondération faible, alors que les deux autres présentent une pondération importante liée à l'utilisation de règles par le lemmatiseur pour retrouver une forme connue. Il suffit à l'utilisateur de valider la première hypothèse proposée. Lors d'une nouvelle finalisation des données, l'ambiguïté disparaîtra et le seul lemme indiqué sera « VIELLER ». Lorsque la forme est attestée plusieurs fois dans le texte, le formulaire rassemble les différentes attestations et c'est l'utilisateur qui doit valider le lemme qu'il considère pertinent pour chaque exemple en cliquant dans la case placée à gauche du lemme retenu. En cas de doute, il peut toujours revenir

à un contexte élargi en cliquant sur le numéro de la page ou sur la référence indiquée. Dans certains cas, la levée d'ambiguïtés est la même pour l'ensemble des occurrences ; il est alors possible de sélectionner « Forcer l'analyse aux *x* attestations ». Mais il existe des formes véritablement ambiguës dans l'état de langue, pour lesquelles il faut examiner attentivement le contexte, en l'absence d'un « étiqueteur » morphosyntaxique précis et fiable. Il s'agit principalement des mots grammaticaux. Ainsi, seul le contexte permettra de dire si la forme *se* est un pronom réfléchi, un morphème introduisant une interrogative indirecte ou une hypothétique, voire une variante diatopique de l'adverbe *si*. Le lemmatiseur n'est d'aucune utilité dans ce cas ; il permet seulement de rassembler toutes les occurrences de la forme *se*. L'éditeur d'un texte en moyen français peut chercher surtout une aide pour construire son glossaire et rassembler toutes les données lexicales utiles à la compréhension du texte édité. Mais l'analyse des mots grammaticaux est utile dans les études linguistiques. Dans le formulaire de levée d'ambiguïtés, il est également possible de choisir d'autres options que les lemmes proposés, en saisissant un autre lemme de la base de connaissance ou en intégrant un nouveau lemme ; il est possible aussi d'abandonner la lemmatisation en sélectionnant par un clic l'un des cas suivants : nom étranger, nom propre, dénomination, nombre, mot exclu.

Ressources produites

Ajoutons que l'utilisateur de l'outil n'est nullement tenu de lever les ambiguïtés : il peut se contenter des ressources produites automatiquement pour générer de façon traditionnelle son glossaire ; il peut aussi lever les ambiguïtés de façon partielle et sélective ; le seul cas qui impose une levée systématique des ambiguïtés est celui de la diffusion du texte lemmatisé – ce que la plateforme en ligne du *DMF* nomme, par commodité, « édition électronique ». Depuis 2011, le texte au programme des agrégations de lettres est ainsi traité chaque année au laboratoire ATILF grâce à

l'outil-glossaire[17]. Le résultat du traitement automatique par LGeRM est vérifié complètement, les ambiguïtés sont levées et les lemmes contrôlés. C'est le projet finalisé qui est diffusé sur la plateforme du *DMF* : l'outil-glossaire offre un certain nombre de fonctionnalités liées à la lemmatisation du texte numérisé. Selon le statut du texte, le lecteur pourra ou non consulter le texte intégral : mais dans tous les cas, il aura la possibilité de connaître sa nomenclature, d'interroger par lemme ou par forme et d'exporter les résultats sous la forme d'un index lemmatisé ou d'une liste d'exemples. L'intérêt pour le linguiste comme pour le littéraire est évident. Le premier peut interroger les formes du texte, commenter leur distribution et leur fréquence. Le regroupement par lemme permet en outre d'avoir accès aux formes attestées des paradigmes morphologiques, notamment pour les verbes les plus usuels. Le second peut s'appuyer sur des données lexicales et sémantiques exhaustives et confronter différentes occurrences de mots-clés, des adjectifs tels que *courtois* ou *vilain* dans le *Roman d'Eneas*, ou un substantif comme *nonchaloir* dans les poèmes de Charles d'Orléans.

En deçà et au-delà du *Dictionnaire du moyen français*

Même si cette contribution, en insistant sur l'origine de l'outil, marque ses liens étroits avec le *DMF*, elle affirme aussi ses développements originaux. L'outil de lemmatisation, créé pour la gestion de la variation dans le dictionnaire, a élargi son champ d'intervention. Les textes médiévaux traités n'appartiennent pas tous à la période du moyen français, qui est la période de référence de notre dictionnaire ; des textes d'ancien français

17. Charles d'Orléans, *Poésies* (2010-2011) : http://atilf.fr/dmf/CharlesOrleans ; Béroul, *Tristan* (2011-2012) : http://atilf.fr/dmf/Beroul ; Guillaume de Lorris, *Le Roman de la Rose*, éd. Armand Strubel (2012-2013) : http://atilf.fr/dmf/RomanRoseStrubel ; *Le Couronnement de Louis*, éd. Ernest Langlois (2013-2014) : http://atilf.fr/dmf/CouronnementLouis ; *Le Roman d'Eneas* (2014-2015) : http://atilf.fr/dmf/RomanEneas ; Jean Renart, *Le Roman de la Rose ou de Guillaume de Dole* (2015-2016) : http://atilf.fr/dmf/RomanRoseGuillaumeDole ; Christine de Pizan, *Le Livre du duc des vrais amants*, éd. Dominique Demartini et Didier Lechat (2016-2017) : http://atilf.fr/dmf/PizanVraisAmans.

ont pu être lemmatisés avec des ajustements mineurs. L'outil-glossaire a également connu des développements au-delà de la période de référence : même si les vérifications et les interventions manuelles sont plus importantes que pour le moyen français, LGeRM peut s'adapter aux textes antérieurs (ancien français) comme aux textes postérieurs (XVI^e^ et XVII^e^ siècles).

Pour ces derniers, la participation de l'ATILF au projet européen Impact[18] a été déterminante. En effet, pour chacun des pays partenaires, la collaboration entre bibliothèque nationale et laboratoire de linguistique historique avait pour but d'améliorer les logiciels d'océrisation utilisés sur les documents anciens. Dans ce cadre, Gilles Souvay a adapté son outil pour répondre à cette nouvelle utilisation. L'ajout de formes modernes Morphalou, autre ressource de l'ATILF[19], lui a permis de construire un lexique capable de couvrir très largement la période choisie pour l'expérimentation (XVII^e^ siècle). Le lexique moderne « archaïsé » a été projeté sur un corpus textuel issu de Frantext et sur un corpus de seize textes ayant conservé leur graphie d'origine, numérisés dans le cadre du projet pour constituer la « vérité terrain ». L'utilisation combinée des ressources a montré son efficacité dans le processus d'adaptation à une période intermédiaire entre le moyen français et la langue couverte par le *TLF* et Morphalou. Un projet synthétique portant sur la variation dans les états anciens du français pourrait permettre de relier et d'approfondir les résultats déjà obtenus sur des périodes particulières.

Tous ces développements se greffent sur un projet initial, celui du *DMF*, qui a su prendre le tournant d'une lexicographie véritablement évolutive, ce qui implique de dépasser non seulement la construction lettre par lettre du dictionnaire, mais aussi d'en faire éclater les limites dans la consultation, grâce au balisage des données et aux liens hypertextuels. Au cœur du

18. « Improving Access to Text » : http://www.impact-project.eu.

19. Le lexique Morphalou est un lexique des formes fléchies du français, à large couverture (540 000 formes). Les données initiales proviennent du *TLFnome*, la nomenclature du *Trésor de la langue française*. Il est en accès libre à des fins de recherche et d'enseignement et sa mise à jour est assurée par l'ATILF.

dispositif, le dictionnaire proprement dit demeure la source et la référence, à travers des versions datées qui sont archivées et une dernière version, directement disponible en ligne, qui reprend, corrige et enrichit la précédente.

Deux thèses en lien avec ce projet indiquent des champs encore à défricher et à labourer : la première, achevée, s'est emparée d'un des corpus initiaux du *DMF* abandonné en cours de route, celui des premiers récits de voyage en français (XIVe et XVe siècles) pour en établir le lexique en s'appuyant sur l'outil-glossaire[20]. Ce lexique des récits de voyage ne saurait ressembler aux lexiques initiaux qui devaient servir de ressources pour construire le futur *DMF* ; il se définit plutôt comme une ressource pour un enrichissement raisonné du dictionnaire. La deuxième porte sur les prépositions en moyen français[21], en lien avec le projet Presto auquel Gilles Souvay a également collaboré en fournissant des lexiques morphologiques[22] – ce qui devrait permettre de réfléchir aux mots outils jusque-là un peu délaissés[23]. À ces travaux étroitement liés au *DMF* s'ajoutent pour la lemmatisation de textes et le traitement des données lemmatisées des partenariats avec des projets extérieurs, l'un avec Michèle Goyens (Université catholique de Louvain),

20. Capucine Herbert, *Les Récits de voyage des XIVe et XVe siècles lemmatisés : apports lexicographiques au* Dictionnaire du moyen français [thèse de doctorat en sciences du langage sous la dir. de Sylvie Bazin-Tacchella, Université de Lorraine, 2016]. Voir également http://www.atilf.fr/dmf/RecitsVoyage.
21. Claire Schlienger, *Le Syntagme prépositionnel marquant l'inclusion : analyse diachronique de* en, dans, dedans *et* à [thèse de doctorat en sciences du langage sous la dir. de Sylvie Bazin-Tacchella, en préparation à l'Université de Lorraine].
22. Presto (*L'Évolution du système prépositionnel du français : approche diachronique et quantitative*) est un projet ANR/DFG (2013-2016), coordonné par Denis Vigier (Lyon II). Dans le cadre de ce projet, Gilles Souvay a participé au développement d'outils de lemmatisation adaptés à toutes les périodes du français en fournissant deux lexiques morphologiques, un lexique LGeRM médiéval et un lexique LGeRM XVIIe siècle : le premier, optimisé pour la période 1300-1500, comporte 66 976 lemmes et étiquettes *DMF*, soit 880 192 entrées dont 142 687 attestées dans la base Frantext ; le second, optimisé pour la période 1550-1700, comporte 89 754 lemmes et étiquettes *TLF*, soit 2 959 371 entrées dont 116 161 attestées (3,9 %).
23. Il existe un autre projet collectif lié au *DMF*, sous l'égide de Bernard Combettes, sur les locutions temporelles. Il faut également signaler que la nomenclature des mots grammaticaux et un lexique des préfixes et suffixes (Robert Martin) sont déjà accessibles sur le site à la rubrique « Compléments au *DMF 2012* », en attendant d'être publiés dans la version 2015.

sur la construction du vocabulaire médical au Moyen Âge et à la Renaissance[24] et l'autre avec les historiens médiévistes de l'Université de Lorraine[25].

Nous nous interrogions en 2013 sur les nouveaux défis que devait relever le *DMF*[26]. Cette contribution qui rappelle l'origine et les développements de l'outil-glossaire du *DMF* indique les chantiers actuels et à venir. L'adaptation d'un tel outil à la langue et aux textes du moyen français, puis l'élargissement de son domaine d'intervention s'inscrivent bien dans le dynamisme et le réalisme d'une lexicographie évolutive qui n'abandonne pas les projets ambitieux, mais choisit de les faire avancer par étape.

24. *Latin authority and constructional transparency at work: neologisms in the French medical vocabulary of the Middle Ages and the Renaissance and their fate.*
25. AMPLor : *Actes médiévaux des princes lorrains ; vers un corpus numérisé*, projet coordonné par I. Guyot-Bachy.
26. Sylvie Bazin-Tacchella et Gilles Souvay, « Avec la version 2012, la fin d'un projet ou de nouveaux défis pour le *DMF* ? », *XXVII^e Congrès international de linguistique et de philologie romanes*, *CILPR*, Nancy, 2013 [communication non publiée].

Résumés

Julie GLIKMAN et Thomas VERJANS, Regards linguistiques sur les éditions de textes médiévaux

Résumé

Cette contribution constitue l'introduction du volume. Elle présente le contexte dans lequel ce numéro a été préparé et la volonté des directeurs du volume d'interroger les rapports entre les pratiques philologiques et les études de linguistique diachronique. Ces rapports peuvent se mesurer dans la place accordée aux faits linguistiques dans les introductions d'édition, ou inversement la place accordée aux variantes et à l'apparat critique dans les corpus numérisés. Elle présente ensuite les différentes contributions du volume.

Abstract

This contribution is the introduction to the volume. It presents the context in which this issue was prepared and the willingness of the editors to question the relationship between philological practices and studies of diachronic linguistics. These relationships can be evaluated by considering the importance given to linguistic facts in the introductory sections of editions. Conversely, it can also be evaluated by according to the importance given to variants and critical apparatus in digitized corpora. The various contributions of the volume are also introduced.

Nathalie BRAGANTINI-MAILLARD, Suivre la lettre du copiste : l'édition critique au service de la linguistique diachronique et diatopique. L'exemple du ms. Paris, BnF, fr. 99

Résumé

La connaissance des modalités d'évolution du français à la fin du Moyen Âge ne peut désormais s'affiner sans une reconnaissance véritable du rôle crucial que jouèrent les copistes au plan linguistique dans la diffusion et la survie des textes anciens. L'action du copiste est en effet double, en s'exerçant à la fois sur le plan horizontal de la circulation des textes d'un espace linguistique à un autre et sur le plan vertical de la transmission des textes à travers les époques. Dans la pratique scientifique, la prise en compte de cet apport déterminant doit passer non seulement par une édition des textes plus respectueuse de la version procurée par un manuscrit donné, mais aussi par un examen documenté, exhaustif et précis des phénomènes linguistiques qui particularisent les témoins retenus et les modifications de scribe. À terme, l'information rassemblée par ces profils linguistiques devrait permettre de mieux appréhender les phénomènes d'adaptation, de rajeunissement et d'enrichissement du français au Moyen Âge. À titre illustratif, nous nous proposons de montrer l'intérêt que présente le ms. BnF, fr. 99 pour suivre de manière privilégiée certains phénomènes de modernisation du français dans la seconde moitié du XV^e^ siècle, ainsi que l'influence que put exercer le lieu de copie occitanisant sur l'adaptation linguistique du texte, autrement dit les conditions d'échanges entre oïl et oc.

Abstract

Knowledge of how French evolved in the late Middle Ages can no longer be refined without a genuine recognition of the crucial linguistic role played by copyists in the dissemination and survival of ancient texts. Copyists act both on the horizontal dimension of the circulation of texts from one linguistic space to another, and on the vertical dimension of the transmission of texts through

the ages. This decisive contribution must be taken into account, not only by providing edition of the texts that are faithful to the version of a given manuscript, but also by a comprehensive and precise examination of the linguistic phenomena that characterize the witnesses and scribal modifications. Ultimately, these linguistic profiles will provide information for a better understanding of the phenomena of adaptation, rejuvenation and enrichment of French in the Middle Ages. To illustrate this, we examine ms. BnF, fr. 99, which displays exceptionally well certain phenomena of the modernization of French in the second half of the 15th century. It also demonstrates the influence that the place of copying with an affinity for Occitan may have had on the linguistic adaptation of the text, i.e. the conditions of exchange between Oïl and Oc.

Laurent BALON,
Pour une « troisième voie » en matière d'édition de textes d'ancien et de moyen français

Résumé

La pratique de l'édition de texte se trouve face à un dilemme : en partant des conseils trouvés dans les quelques articles sur la question et les manuels récents donnant des principes d'édition, on observe que les critères actuels de choix des variantes aboutissent à l'exclusion du matériau intéressant le linguiste qui, de son côté, aurait besoin d'un exposé intégral de toutes les données, sans tri. Ce besoin d'un non-choix est important, mais peu facile à satisfaire, voire impraticable à l'écrit, et la présentation des données intégrales du manuscrit se heurte à la lisibilité et à l'intelligibilité. L'objet de cette contribution est de présenter une méthode d'édition constituant un compromis entre l'édition critique traditionnelle et la transcription dite diplomatique, reposant sur un protocole de choix de variantes permettant de mieux satisfaire certains besoins des linguistes. Afin de fournir au linguiste des informations immédiatement exploitables et utiles à l'avancée de la discipline, le principe méthodologique proposé consiste à signaler dans l'édition

certains faits de langue relevant de la ponctuation du mot par l'emploi d'un code graphique qui en conserve la trace, à savoir un système de « tirets » déjà suggéré par Jacques Monfrin pour la transcription des documents d'archives, mais complété et appliqué pour la première fois à un texte littéraire par Nelly Andrieux-Reix. Le bien-fondé et l'intérêt de cette méthode seront illustrés par des études de cas en lien avec notre propre travail de recherche.

Abstract

Editors must cope with a dilemma: according to publishing principles in recent papers and textbooks, the current criteria for choosing variants excludes materials of great interest to linguists. They would need a comprehensive view of the data, without sorting. This is not easy to achieve, and even impossible on paper. The full presentation of the data of the manuscript hampers legibility and intelligibility. The purpose of this contribution is to present a compromise between traditional critical editing and diplomatic transcription, based on a protocol of choice of variants that better satisfies linguistic investigations. The proposed methodological principle aims at providing information that is immediately usable and useful for the advancement of the linguistics. This purpose is achieved by indicating facts relating to the punctuation of the word by using a graphic code that keeps track of them: a system of "dashes", suggested by Jacques Monfrin for the transcription of archival documents. This system is expanded and applied for the first time to a literary text by Nelly Andrieux-Reix. The merits and interest of this method will be illustrated by case studies related to our own research work.

Alexei LAVRENTIEV, Céline GUILLOT-BARBANCE et Serge HEIDEN, Enjeux philologiques, linguistiques et informatiques de la philologie numérique : l'exemple de la segmentation des mots

Résumé

Les linguistes travaillant sur l'histoire de la langue ont toujours exploité et utilisé comme principale source d'exploration les éditions « classiques », bien que depuis longtemps on connaisse leurs limites pour la recherche linguistique. Le développement des technologies modernes a d'un autre côté rendu le recours à de nouveaux outils (concordances, index, calculs statistiques) peu à peu indispensable à la recherche en langue, et plus récemment, les progrès continus de la technologie ont également permis d'envisager la réalisation d'éditions d'un nouveau type. L'édition numérique, qui a déjà donné lieu à plusieurs réalisations concrètes, a ainsi permis aux linguistes auparavant bridés par le papier et les techniques traditionnelles d'exprimer plus librement leurs besoins et leurs exigences. Plusieurs recherches récentes déjà publiées montrent l'efficacité de ce mouvement et le caractère novateur des acquis ainsi obtenus. À partir d'un exemple concret d'édition numérique interactive, notre présentation détaillera les enjeux méthodologiques liés à ces nouveaux outils et à ces nouvelles pratiques, en proposant une réflexion sur le concept de « philologie numérique » et en montrant ses principaux apports pour la recherche diachronique. Cette question sera illustrée en particulier par la question de la segmentation des mots.

Abstract

Linguists working on the history of language have always exploited "classical" editions as their main source of exploration, although the limits of such resources for linguistic research have long been known. On the other hand, modern technology has gradually offered new tools (concordances, indices, statistical calculations), that now prove to be indispensable. More recently,

the continuous progress has also made it possible to produce new types of editions. Digital publishing, which has already produced several achievements, has thus enabled linguists to express their needs and requirements better than before, freed from the constraints of paper and traditional techniques. Several recent studies demonstrate the efficiency of digital publishing and the innovative nature of the results obtained. Based on an example of interactive edition, we survey the methodological issues related to these new tools and practices, by investigating the concept of "digital philology", and by evaluating how it contributes to diachronic research. The specific issue of word segmentation will illustrate our point.

Nicolas MAZZIOTTA, L'activité éditoriale comme démarche de représentation de la connaissance : l'exemple de la ponctuation médiévale

Résumé

Cette contribution concerne le traitement éditorial de la ponctuation médiévale, selon une approche de la philologie comme activité de représentation des connaissances. Après une présentation des concepts de *connaissance* et d'*inscription* (des connaissances), le traitement de la ponctuation médiévale sert d'exemple aux questionnements que soulève toute activité éditoriale. Dans la démarche ecdotique, il s'agit d'identifier des classes de signes, pour distinguer ce qui est différent et rapprocher ce qui est similaire, mais également de segmenter correctement les unités observées. En outre, éditer consiste à « donner à lire », ce qui se manifeste par l'importance de choix ergonomiques importants pour garantir l'accessibilité de la connaissance inscrite. À bien des égards, l'inscription informatique de l'édition a beau ouvrir le champ des possibles, elle ne résout pas tout. Pour inscrire, il faut d'abord comprendre. L'édition ne pourra jamais se passer des *choix* foncièrement humains qui fondent le travail de construction de la connaissance.

Abstract

This contribution focuses on the editorial treatment of medieval punctuation, according to an approach of philology as an activity of *knowledge representation*. After a brief presentation of the concepts of *knowledge* and *inscription* (of knowledge), the treatment of medieval punctuation serves as an example for the questions raised by any editorial activity. Identifying classes of signs and distinguishing between what is different and what is similar are key parts of the ecdotic process. Moreover, by editing a text, one actually *makes it readable*. Consequently, ergonomic choices are prominent in this process, in order to guarantee the accessibility of the knowledge inscribed. In many respects, digital publishing opens up the field of possibilities, but it does not solve the fundamental problems. Understanding the text stands as the first step into building any valuable critical edition. Human *choices* will always remain the basis of any elaboration of knowledge.

Sylvie BAZIN-TACCHELLA et Gilles SOUVAY, Lemmatisation et construction automatique de ressources lexicographiques : les développements du lemmatiseur LGeRM

Résumé

Le lemmatiseur LGeRM, conçu à l'origine pour faciliter la consultation du *Dictionnaire du moyen français*, a connu depuis 2008 de nouveaux développements et est aujourd'hui utilisé dans de nombreux autres contextes, notamment dans l'interrogation de bases textuelles et la constitution de lexiques ou glossaires informatisés, autant d'outils qui peuvent servir d'aide à l'édition, le lemmatiseur ayant été intégré depuis à plusieurs grands projets d'édition en ligne. Cette contribution se propose de retracer l'histoire de la conception de LGeRM et de ses développements successifs, en montrant les différentes possibilités de l'outil illustrées à partir des projets récents.

Abstract

The LGeRM lemmatizer, originally designed to facilitate the consultation of the *Dictionnaire du moyen français*, has undergone new developments since 2008. It is now used in many other contexts. In particular, it helps the interrogation of textual bases and the constitution of computerized lexicons or glossaries. Additionally, the lemmatizer has also been integrated into several major online publishing projects in order to help the publishing process. This contribution retraces the history of the conception of LGeRM and its successive developments, by showing how recent projects make use of it.

Table des matières