REVUE DE
LINGUISTIQUE
FRANÇAISE
DIACHRONIQUE

7
2017

DIACHRONIQUES

LES ÉTATS ANCIENS DES LANGUES À L'HEURE DU NUMÉRIQUE

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

PUPS

7

DIACHRONIQUES

REVUE DE LINGUISTIQUE FRANÇAISE DIACHRONIQUE

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-TACCHELLA & 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

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

12 €

http://pups.paris-sorbonne.fr

Maquette : www.stephanemercier.fr

LES ÉTATS ANCIENS DES LANGUES À L'HEURE DU NUMÉRIQUE

Les états anciens des langues à l'heure du numérique

PUPS

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

Diachroniques n° 7
ISBN papier : 979-10-231-0581-0

PDF complet – 979-10-231-2155-1

TIRÉS À PART EN PDF :
Ducos – 979-10-231-2156-8
Martin – 979-10-231-2157-5
Bazin-Tacchella & Souvay – 979-10-231-2158-2
Salvador, Issac & Fasciolo – 979-10-231-2159-9
Pérez Rodríguez – 979-10-231-2160-5
Guadagnini – 979-10-231-2161-2
Gómez Rabal – 979-10-231-2162-9
Goyens & Szecel – 979-10-231-2163-6
Guillot, Heiden & Lavrentiev – 979-10-231-2164-3
Petit – 979-10-231-2165-0
Masià – 979-10-231-2166-7
Richards – 979-10-231-2167-4

Maquette initiale : Compo-Méca (64990 Mouguerre)
Réalisation : Emmanuel Marc Dubois/3d2s

SUP
Maison de la Recherche
Sorbonne Université
28, rue Serpente
75006 Paris

Tél. (33) 01 53 10 57 60

sup@sorbonne-universite.fr

sup.sorbonne-universite.fr

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

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

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

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

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

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

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

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

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

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

Fig. 1. Formulaire d'interrogation

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

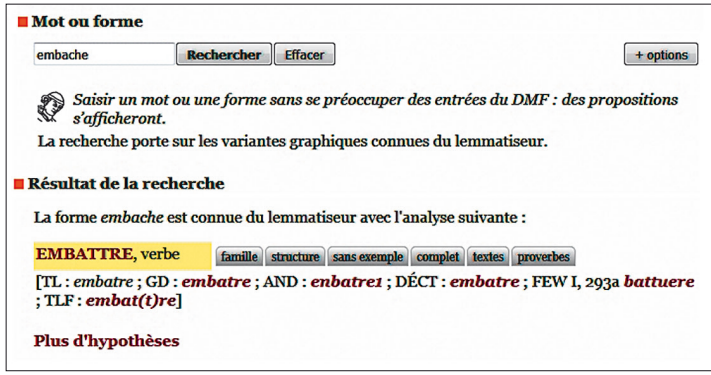

Fig. 2. Résultat de la recherche

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

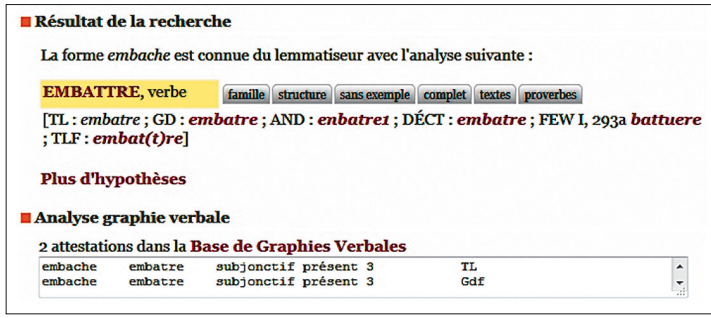

Fig. 3. Résultat de la recherche avec analyse de la graphie verbale

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

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

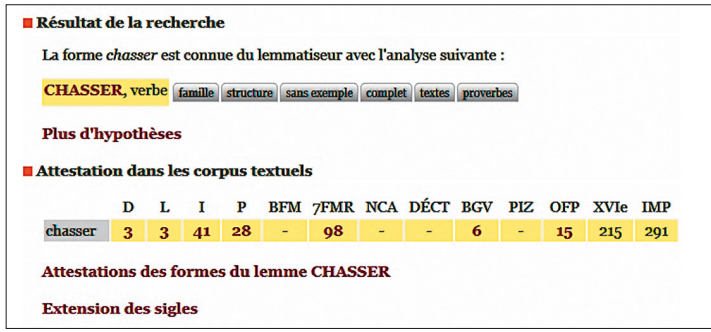

Fig. 4. Résultat de la recherche avec attestations dans les bases

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

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

CHASSER — Voir diachronie dans FRANTEXT

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

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

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

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

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

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

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

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

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

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

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

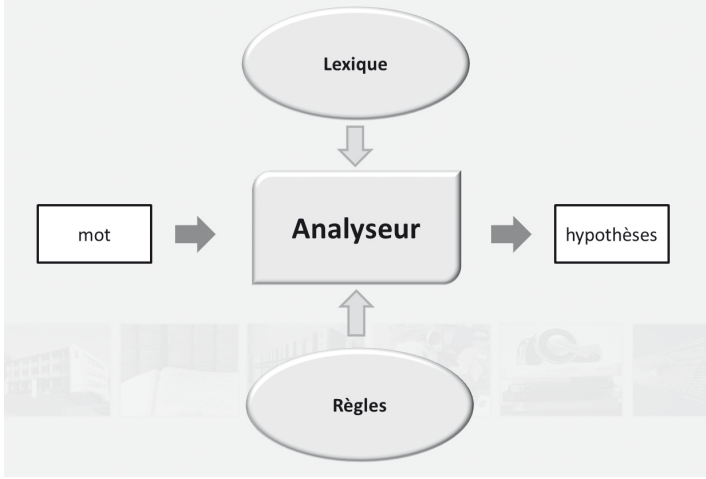

Fig. 6. Architecture du système

Lors de la recherche d'une entrée dans le dictionnaire, l'analyseur saisit un mot – hors contexte – et fournit des hypothèses de lemmes. Il utilise pour cela un lexique et des règles de flexion et de variation graphique.

L'analyseur

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

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

Le lexique

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

amer, AIMER, verbe
amer, AMER, adj.

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

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

Les règles

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

si conditions *alors* action *finsi*

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

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

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

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

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

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

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

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

Y → I
fayre → (*faire*, FAIRE, verbe)

Elle contient des règles d'équivalence graphique :

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

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

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

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

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

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

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

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

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

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

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

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

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

Fig. 7. Extrait d'un index lemmatisé

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

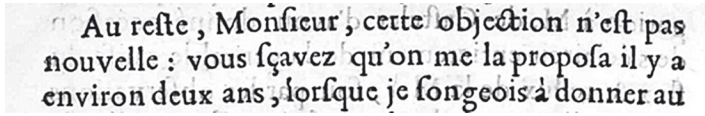
Au reſte, Monſieur, cette objection n'eſt pas
nouvelle : vous ſçavez qu'on me la propoſa il y a
environ deux ans, lorſque je ſongeois à donner au

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

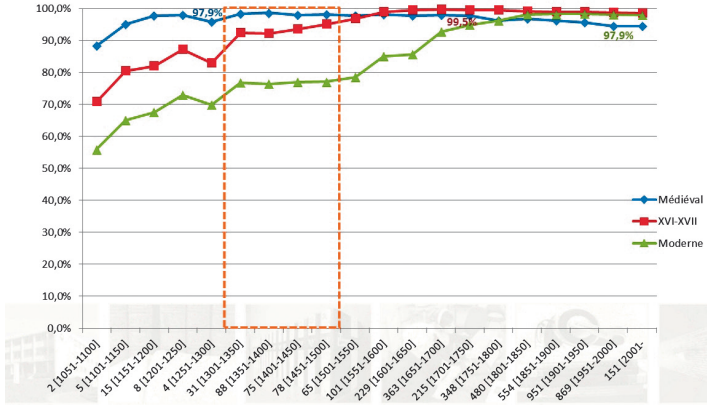

Fig. 10. Taux de couverture des lexiques : fréquences

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

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

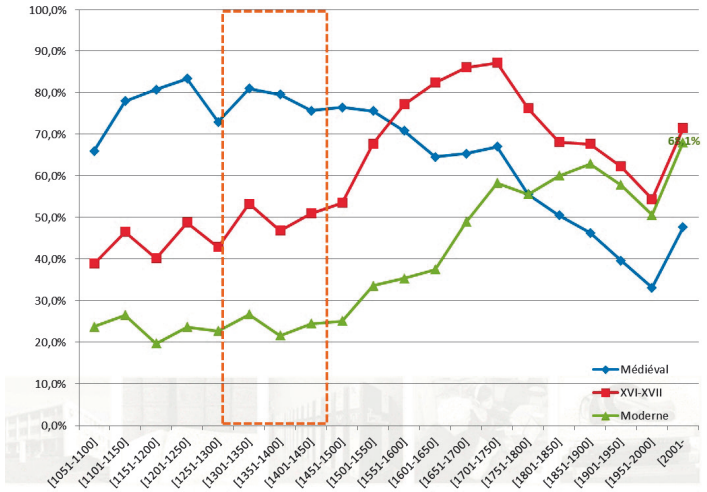

Fig. 11. Taux de couverture des lexiques : graphies

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

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

Références bibliographiques

Ressources électroniques

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

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

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

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

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

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

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

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

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

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

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

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

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

Collaborations

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

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

Communications et articles

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Résumés / Abstracts

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

Résumé

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

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

Abstract

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

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

Résumé

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

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

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

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

Abstract

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

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

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

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

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

Résumé

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

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

Abstract

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

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

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

Résumé

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

Abstract

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

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

Résumé

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

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

Abstract

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

Robert MARTIN, À propos du *DMF*

Résumé

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

Abstract

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

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

Résumé

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

Abstract

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

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

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

Résumé

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

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

Abstract

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

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

Résumé

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

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

Abstract

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

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

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

Résumé

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

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

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

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

Abstract

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

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

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

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

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

Résumé

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

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

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

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

Abstract

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

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

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

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

Table des matières

Présentation
Joëlle Ducos 7

À propos du *DMF* :
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 *DFSM* : une expérience
Xavier-Laurent Salvador, Fabrice Issac & Marco Fasciolo 49

Le *Lexicon Latinitatis Medii Aevi Regni Legionis*
(VIII[e] siècle-1230) : caractéristiques et quelques exemples
(*ventrescas, iera, cumbo, plentum*)
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 *Glossarium Mediae Latinitatis Cataloniae* :
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 Szecel 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**