REVUE DE
LINGUISTIQUE
FRANÇAISE
DIACHRONIQUE

7
2017

DIACHRONIQUES

LES ÉTATS ANCIENS DES LANGUES À L'HEURE DU NUMÉRIQUE

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

séparatrice pour mettre en évidence cette distribution. Le retour aux données qui sont à l'origine de la construction du graphique permet d'interpréter la position originale des parties au discours direct du *Comput* de Philippe de Thaon (représenté par le cercle situé en haut à droite de la **fig. 1**). Cette position est liée à l'usage très particulier des guillemets dans ce texte : ils n'indiquent pas les segments au discours direct mais servent à citer des mots isolés. Les guillemets étant les marques formelles sur lesquelles a reposé l'encodage du discours direct et sa dissociation des autres parties de textes dans la base, l'usage déviant de ces marques dans le *Comput* explique son positionnement excentrique dans le graphique[3].

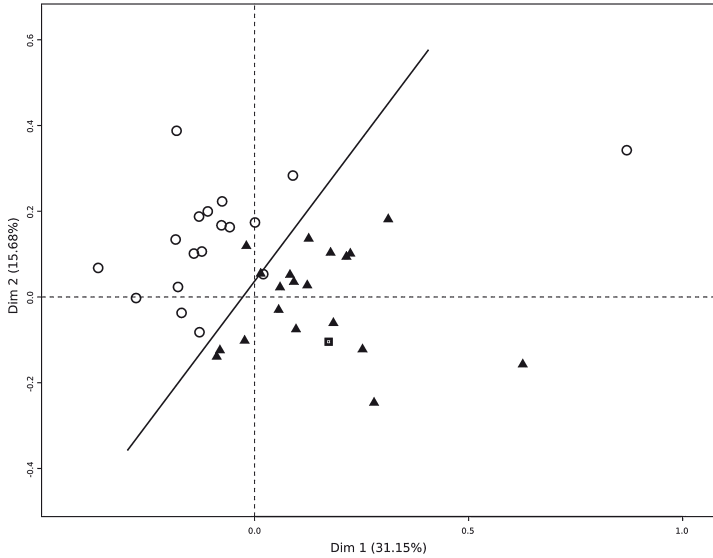

Fig. 1. AFC des textes du corpus, en distinguant les parties narratives *vs.* au discours direct. Pour le calcul, les textes sont modélisés par la fréquence des catégories morphosyntaxiques qu'ils utilisent. Les parties au discours direct sont représentées par des cercles, et les parties narratives par des triangles.

3. Le balisage du discours direct a été réalisé de manière semi-automatique dans tous les textes de la base. Il s'appuie sur les marques graphiques (guillemets ouvrants et fermants) insérées par les éditeurs modernes des textes médiévaux. Les limites de ce balisage automatique sont évidentes. Il permet néanmoins de dégager des tendances générales grâce à l'analyse d'un grand nombre de textes.

L'analyse des étiquettes morphosyntaxiques qui fondent la position relative de chaque point selon le premier axe[4] permet de se faire une idée assez précise des éléments les plus spécifiques au discours direct ou aux autres parties de textes. La fréquence élevée des pronoms personnels, conjonctions de subordination, négations, pronoms impersonnels, interjections et adverbes caractérise le discours direct, celle des noms propres, articles définis, contractions de l'article et des prépositions (*du*, *au*, etc.), noms communs, déterminants cardinaux et participes présents distingue tout ce qui lui est extérieur.

Les études qui sont menées dans un tel cadre d'analyse reposent sur des ressources riches et adaptées. Elles supposent la possibilité d'exploiter les données textuelles avec des outils numériques de synthèse et de recherche. Elles impliquent le traitement d'un volume de données suffisant et d'une diversité assez représentative pour qu'on parvienne à des résultats stables et généraux. Elles imposent aussi un équipement numérique relativement poussé des textes : encodage du discours direct, étiquetage morphosyntaxique de tous les mots, description des unités textuelles grâce à un système de métadonnées permettant l'interprétation des résultats et l'étude de la variation.

Une telle méthode de recherche, qui s'élabore dans un cadre de plus en plus expérimental, vise également à permettre à la communauté scientifique de reproduire les mêmes analyses, grâce à la diffusion, la documentation et la pérennisation des ressources. Cette méthode favorise l'enrichissement continu des textes numériques au fil des analyses linguistiques, les informations rendues disponibles par l'analyse ayant vocation à être associées aux données elles-mêmes pour pouvoir être réutilisées lors de recherches ultérieures. Dans le cas de l'étude citée ci-dessus par exemple, l'analyse des spécificités de l'oral représenté dans la Base de français médiéval amène à réinterpréter la fonction des guillemets dans le texte du *Comput*

4. Pour des raisons de lisibilité nous avons choisi de ne pas faire figurer ces étiquettes sur le graphique, mais l'outil donne directement accès aux informations qui sous-tendent la position des points du graphique.

et à revoir le balisage des séquences au discours direct dans ce texte.

L'amélioration continue des données et le coût inhérent à la préparation et à l'équipement numérique des textes rendent par ailleurs de plus en plus nécessaire le développement partagé et communautaire des ressources. Aux plans juridique et pratique, il est devenu indispensable de permettre la libre circulation et la rediffusion responsable de ces ressources parmi l'ensemble des partenaires qui participent, pour une part variable et à différents niveaux, à leur production et leur exploitation. Le respect de normes et standards internationaux de représentation numérique des textes, l'utilisation de licences de (re)diffusion ouvertes compatibles avec les juridictions et la jurisprudence internationales sont les principaux instruments de cette politique d'échanges communautaires. Nous essaierons de montrer, dans la suite de cet article, les implications très concrètes de cette méthode de travail concernant les aspects philologiques des textes (section 2) comme les outils d'analyse qui permettent de les exploiter (section 3).

Chaîne philologique ouverte pour l'établissement et l'annotation des textes de la BFM

Principes méthodologiques

Les recherches qui sont menées dans le cadre de la linguistique de corpus et que nous avons illustrées ci-dessus par l'étude des spécificités du discours direct en français médiéval exploitent le plus souvent un volume important de données textuelles. Elles permettent surtout d'appliquer les outils informatiques à l'analyse de données de type et de niveau très variés. Certaines de ces informations concernent les unités textuelles dans leur ensemble (métadonnées textuelles), d'autres sont internes aux textes ou à des parties de textes (structures textuelles, comme les passages au discours direct, les groupes de mots correspondant à des unités inférieures, les mots du texte, les caractères, etc.).

Le standard de balisage XML-TEI[5] permet d'encoder toutes ces informations aux niveaux qui leur correspondent (en-têtes pour les métadonnées textuelles, corps du texte pour tout le reste).

La méthodologie de corpus implique par conséquent que l'on identifie et traite séparément au moins trois types d'information : (i) les métadonnées textuelles, qui servent à caractériser les textes, à les regrouper ou à les dissocier, à interpréter les variations observées grâce à l'approche contrastive ; (ii) les structures internes ou unités linguistiques sur lesquelles portent les analyses et qui demandent à être clairement délimitées et balisées (dans l'exemple cité, les segments au discours direct, les limites de mots) ; (iii) les propriétés associées à ces unités linguistiques destinées à être mobilisées lors de l'analyse (étiquettes morphosyntaxiques, lemmes, annotations syntaxiques, etc.). C'est de la combinaison de ces informations multiples que dépendent l'interprétation des résultats et la richesse des analyses.

Lorsqu'il s'agit de corpus de textes médiévaux (ou plus généralement de ceux dont l'édition demande un travail philologique important), des questions méthodologiques supplémentaires doivent par ailleurs être résolues ou en tout cas prises en compte. Il convient en premier lieu de distinguer les sources primaires (les manuscrits, pour l'époque médiévale) des sources secondaires (éditions scientifiques). Il n'est pas envisageable de constituer de grands corpus de textes médiévaux directement à partir des sources primaires, car une bonne transcription de manuscrit est un travail philologique très laborieux qui comprend notamment l'étude de la tradition manuscrite et le choix d'un manuscrit de base, l'identification des erreurs scribales éventuelles, la résolution des nombreuses ambiguïtés des graphies médiévales (telles que les séries de jambages, agglutinations ou abréviations non univoques) et éventuellement la consultation d'autres manuscrits de la même œuvre afin d'éclaircir les passages difficiles. Un tel investissement

5. En ligne : http://www.tei-c.org.

est discutable si une bonne édition scientifique existe déjà pour un texte. Mais l'utilisation des éditions scientifiques comme source de données pour les corpus numériques pose, d'un autre côté, des problèmes importants.

Ré-ingénierie numérique d'éditions scientifiques existantes

On observe d'abord que les pratiques d'établissement du texte varient considérablement d'une tradition philologique à l'autre ; elles évoluent avec le temps, et dépendent dans une mesure non négligeable des choix personnels de l'éditeur. Même si, dans le domaine de l'édition de textes en français médiéval, la tradition « bédiériste[6] » (qui consiste à respecter autant que possible le manuscrit de base) est largement dominante, le degré de « liberté » que les philologues se donnent dans la correction du manuscrit originel est très variable. Ainsi, May Plouzeau (1994) a démontré que la dernière version de l'édition de *La Mort Artu* par Jean Frappier (1964) ne constituait plus une source de données linguistiques fiable.

Certains aspects de l'établissement de texte sont laissés entièrement à l'appréciation de l'éditeur scientifique. Il s'agit en particulier de la ponctuation et de la segmentation des locutions qui se sont figées et sont devenues des lexies uniques au gré de l'évolution de la langue (par ex. la locution prépositionnelle *par mi*, le groupe adverbial *ja mais* ou le syntagme nominal *bon heur*[7]). Cette hétérogénéité des pratiques pose des problèmes évidents pour l'annotation morphosyntaxique et la lemmatisation du corpus, ainsi que pour les recherches et l'analyse des données textuelles.

Une solution partielle aux problèmes posés par la diversité des pratiques philologiques repose sur la normalisation de l'encodage des textes grâce à l'application des recommandations du consortium TEI (*Text Encoding Initiative*). On peut ainsi neutraliser les différentes manières d'indiquer les mêmes types

6. Nommée ainsi en l'honneur de Joseph Bédier, qui en a formulé les principes dans son étude de la tradition manuscrite du *Lai de l'ombre* (1928).
7. Le processus inverse est possible, mais beaucoup plus rare : par exemple le préfixe *tres-* (du latin *trans-*) devenu l'adverbe *très*.

d'interventions éditoriales. Par exemple, les fragments restitués par l'éditeur scientifique à la place des lacunes peuvent être signalés, selon les éditions, par des crochets ou par des chevrons (plus rarement). La balise TEI <supplied> peut être utilisée dans les deux cas. Un autre exemple concerne l'indication des passages au discours direct. C'est toujours l'éditeur scientifique qui place les guillemets, car les manuscrits médiévaux n'utilisaient pas cette marque graphique dans cette fonction. En revanche, selon les traditions philologiques et les règles typographiques adoptées dans différents pays, les guillemets peuvent être français (« ») ou anglais (“ ”), être ou ne pas être fermés devant les incises ou entre les prises de parole dans les dialogues. La balise TEI <q> permet d'harmoniser toutes ces pratiques hétérogènes. Enfin, le balisage des mots du texte peut permettre de dissocier la segmentation visuelle réalisée à l'aide des blancs typographiques de la segmentation analytique utilisée dans l'annotation linguistique et dans les requêtes appliquées au corpus. On peut ainsi procéder à une normalisation massive tout en respectant les choix de l'éditeur dans la présentation graphique. L'harmonisation de la segmentation graphique étant cependant une tâche particulièrement lourde, elle n'a pas encore été réalisée dans le corpus de la BFM[8].

Une seconde source de difficulté est liée au fait que l'état de la propriété intellectuelle de nombreuses éditions n'est pas clair. En France (à la différence de l'Allemagne et de l'Italie, par exemple), il n'existe pas de texte législatif spécifique concernant les éditions critiques (Margoni et Perry, 2011). Si on considère ces éditions comme des œuvres originales créées par les éditeurs scientifiques, les droits patrimoniaux restent protégés pendant soixante-dix ans après la mort de l'éditeur. Certaines maisons d'édition prétendent détenir les droits de diffusion numérique des textes dont les éditeurs scientifiques sont décédés depuis plusieurs décennies. La recherche d'éventuels ayants droit

8. En effet, les annotations syntaxiques réalisées sur certains textes de la BFM dans le cadre du projet SRCMF (Stein et Prévost, 2013) reposent sur la segmentation actuelle des mots. Or, toute modification des choix de segmentation lexicale suppose de réaligner ces annotations sur les nouvelles unités lexicales qui pourraient être créées.

de ces éditions s'avère souvent très longue et complexe. Les contrats d'édition récents prévoient généralement la cession exclusive des droits de diffusion numérique de toutes sortes à la maison d'édition, ce qui les rend inutilisables dans des corpus numériques, pour lesquels la libre diffusion des données est vitale (Guerreau, 2015). Même si la maison d'édition donne son accord pour l'intégration de « son » texte dans un corpus, elle peut le retirer à tout moment, ce qui risque de nuire à la reproductibilité et à la continuité des recherches basées sur ces données. Pour ne plus faire courir ce risque à ses utilisateurs, la Base de français médiéval a été contrainte de retirer un certain nombre de textes (plus d'un million d'occurrences mots au total) en août 2014 à la suite de la rupture d'une convention avec une maison d'édition.

Selon un autre point de vue, défendu récemment par l'une des parties dans un procès opposant deux maisons d'éditions, le « corps » du texte d'une édition scientifique (à l'exclusion de l'introduction, des notes, des variantes et des annexes de toutes sortes) n'est pas une création de l'éditeur scientifique au sens où l'entend le Code de la propriété intellectuelle, et n'est donc pas protégeable. Un jugement de première instance a confirmé cette position, mais la controverse est loin d'être close dans ce débat juridique. Par ailleurs, les notes du texte peuvent comporter des informations très importantes et nécessaires à son analyse (comme l'indication de variantes ou la justification d'une correction).

L'objectif de la Base de français médiéval étant d'offrir à la communauté des chercheurs la ressource la plus riche et la plus fiable possible pour étudier la langue française des premiers textes à la fin du XV^e^ siècle, de multiples facteurs sont pris en compte lors de la sélection des textes à intégrer au corpus. Les œuvres sont d'abord sélectionnées en fonction de leur intérêt linguistique (on cherche à équilibrer le corpus sur le plan diachronique en tenant compte des genres et domaines textuels). La qualité philologique des éditions et leur statut juridique (qui peuvent être facteurs d'exclusion) sont ensuite

évalués. Dans le cas d'éditions récentes dont les fichiers de saisie sous un logiciel de traitement de texte sont disponibles et dont les auteurs n'ont pas cédé l'exclusivité des droits à une maison d'édition, la BFM négocie directement avec les éditeurs scientifiques pour obtenir ces fichiers sources et pouvoir les diffuser sous une licence libre. Des éditions plus anciennes sont numérisées aux frais de l'équipe de la BFM, avec l'accord des ayants droit, lorsqu'on les trouve.

Création d'éditions numériques originales

Tous les problèmes liés à la réutilisation d'éditions traditionnelles peuvent être résolus dans des éditions « nativement numériques ». Il est possible, notamment, de fournir plusieurs niveaux de transcription dont chacun est adapté à des usages et à des catégories de lecteurs différents (Guillot *et al.*, 2017 ; Marchello-Nizia *et al.*, 2015).

Dans la pratique, il nous semble qu'une représentation à deux niveaux, qu'on peut qualifier de « normalisée » et « diplomatique », peut satisfaire la grande majorité des utilisateurs. Le niveau normalisé se rapproche de la tradition de l'édition des textes littéraires, avec toutefois l'application de règles plus explicites concernant notamment la ponctuation, la segmentation graphique des mots et la résolution des abréviations. Le niveau diplomatique se rapproche davantage du système graphique du document source : les lettres restituées à la place des abréviations sont signalées par des italiques, les distinctions « ramistes » (phonétiques) des lettres *i/ĩ* et *u/ṽ* ne sont pas introduites, les diacritiques modernes ne sont pas ajoutés et aucune marque de ponctuation n'est utilisée, lorsqu'il n'y en a pas dans le document transcrit. La segmentation graphique correspond dans la mesure du possible à celle du document source[9]. Ce type de transcription peut être indispensable pour certains types de recherche linguistique, en particulier dans le domaine de la morphologie (Schøsler, 2004,

9. Dans certains cas, faute d'instrument de mesure précis, la lecture et la décision de transcrire un blanc entre deux mots du manuscrit restent à l'appréciation de l'éditeur.

p. 463). Pour réaliser une transcription « à deux niveaux », il n'est pas nécessaire de transcrire deux fois le texte source. Il suffit d'utiliser un petit nombre de raccourcis typographiques, dans le cadre d'une convention de transcription utilisant un mécanisme de caractères spéciaux, qui permettent de générer automatiquement les deux types de transcription à partir d'un fichier unique. Par exemple, le caractère dièse permet de signaler dans *#Dieu* que la majuscule du nom propre est due à la normalisation éditoriale et que la graphie du document source comporte une minuscule.

Les principes de la segmentation lexicale pour les outils d'annotation linguistique et pour le moteur de recherche peuvent être clairement définis et appliqués dans le cadre de grandes collections d'éditions numériques et, idéalement, partagés par la communauté internationale des philologues. Il convient de souligner que la normalisation de la segmentation au niveau du codage informatique n'empêche pas l'éditeur scientifique d'appliquer ses propres choix de segmentation visuelle dans l'édition à l'écran ou imprimée. En règle générale, le codage de la segmentation la plus fine est préférable, car il est plus simple de regrouper que de découper des unités *a posteriori*.

Le modèle économique des éditions numériques diffère considérablement des éditions imprimées. Le coût de la fabrication et de la diffusion du livre est important et peut justifier la cession des droits à l'éditeur commercial. Pour les éditions numériques basées sur une chaîne de production bien réglée et disposant d'une plateforme de diffusion adaptée, c'est le travail philologique de l'éditeur scientifique qui représente l'investissement principal. Le coût d'hébergement d'une ressource sur la toile est relativement faible, et des services à forte valeur ajoutée (impression à la demande, export dans un format particulier) peuvent être proposés aux lecteurs. Ceci rend tout à fait possible la diffusion des éditions numériques sous une licence libre de type *Creative Commons* ou similaire. Cela est important non seulement pour faciliter l'accès à la lecture de

ces éditions par les membres de la communauté académique et un public plus large, mais aussi et surtout pour permettre leur intégration dans des archives ouvertes, dans des corpus divers et variés, ainsi que dans la toile de données. La possibilité d'accéder aux données primaires des travaux de recherche pour reproduire leurs résultats est un élément important de leur scientificité. Enfin, plus une ressource numérique est utilisée et reproduite, plus il y a de chances qu'elle puisse s'adapter aux évolutions technologiques constantes.

La diffusion ouverte des données implique l'utilisation de formats de représentation ouverts, le respect des normes et standards d'encodage et la documentation des pratiques particulières à une équipe. Pour ce qui concerne l'encodage d'éditions scientifiques numériques, le cadre proposé par le consortium TEI (déjà évoqué plus haut) semble à ce jour le mieux adapté. Les avantages de la TEI sont sa riche expérience (plus de vingt-cinq ans d'existence), la variété des types de textes et d'éditions pris en charge, la souplesse des schémas de balisage proposés, sa documentation extensive et sa communauté active. Certains des points forts de la TEI peuvent également devenir ses faiblesses. Le très grand nombre de balises disponibles pour l'encodage et le fait qu'il existe toujours plusieurs façons de faire pour encoder un même phénomène rend difficile la mise au point d'outils d'analyse. La documentation fournie par la TEI ne suffit pas toujours pour expliciter les choix faits au niveau d'un projet de recherche particulier. Pour cette raison, la TEI recommande de personnaliser le schéma de balisage utilisé par un projet ou par une communauté et fournit un mécanisme facilitant cette personnalisation et sa documentation. La BFM utilise le balisage TEI pour ses éditions numérisées depuis le début des années 2000 et documente ses pratiques de manière précise (Bertrand *et al.*, 2014). Les éditions nativement numériques appliquent le même schéma de base que le reste de la BFM, mais utilisent un certain nombre de balises supplémentaires, notamment pour la représentation des transcriptions multi-niveaux.

Chaîne analytique ouverte pour l'exploitation textométrique des textes de la BFM

L'application d'un sous-ensemble précis des recommandations de la TEI documenté dans le cadre de la BFM a non seulement rendu possible la mise en place un réseau d'échanges de textes entre partenaires partageant les mêmes pratiques philologiques, mais elle a également permis à ce corpus de textes d'être intégré dans la plateforme TXM pour son analyse et sa diffusion.

Développée initialement dans le cadre d'un projet financé par l'ANR en 2007-2010, cette plateforme a pour objectif de pérenniser et de mutualiser les développements informatiques d'outils textométriques comme *Hyperbase, Lexico 3, Le Trameur, DTM* et *Weblex*. La textométrie est une méthode d'analyse de corpus textuels développée depuis les années 1980[10] combinant des outils statistiques appliqués aux différentes unités des textes (analyse factorielle, calcul de spécificités, classification, analyse de co-occurrents) et des outils documentaires (listes de mots, recherche plein texte de patrons de mots pour l'établissement de concordances, lecture des éditions de textes du corpus). Son implémentation dans la plateforme TXM a été l'occasion d'élargir la méthode aux corpus textuels richement encodés en XML-TEI et annotés par différents outils de traitement automatique de la langue (comme le lemmatiseur *TreeTagger*) et de produire une version pour poste Windows, Mac OS X ou Linux (appelée « logiciel TXM ») ainsi qu'une version serveur pour l'accès par internet (appelée « portail TXM »), les deux versions partageant la même plateforme de base pour l'exploitation des corpus.

La mutualisation de la construction et de la maintenance de la plateforme est obtenue par un mode de développement ouvert appelé « *open-source* », bien établi dans les projets de recherche en informatique depuis vingt ans, qui fonctionne sur deux plans. D'une part, tout partenaire peut accéder aux sources du logiciel pour l'adapter ou l'améliorer en respectant les termes de sa licence

10. En ligne : http://textometrie.ens-lyon.fr/spip.php?rubrique80.

de diffusion[11]. Et d'autre part, la plateforme intègre elle-même de nombreux composants logiciels développés par d'autres projets *open-source*, en particulier l'environnement de calcul statistique R[12], le moteur de recherche CQP[13] et la plateforme Eclipse[14] pour la version pour poste de TXM. Le fait de pouvoir accéder aux sources du logiciel TXM est par ailleurs un gage de scientificité des travaux réalisés grâce à cet outil, parce qu'il ne fonctionne pas comme une boîte noire. Tous ses calculs sont décomposables et vérifiables à partir de ses sources. Le fait de déléguer certains calculs à d'autres composants *open-source* permet de profiter de leurs performances et de leurs améliorations constantes par leur communauté de développement. Mais il faut s'assurer que chaque composant soit bien maintenu par une communauté de développeurs dynamique, par des institutions ou des entreprises au risque qu'il ne soit un jour plus développé et ne puisse plus suivre les évolutions technologiques et continuer à fonctionner. Auquel cas on doit soit le remplacer par un composant *open-source* équivalent, soit le maintenir soi-même.

La pérennisation du développement repose sur deux plans : d'une part l'utilisation d'un langage de programmation correspondant à un standard industriel reconnu et développé selon un mode communautaire ouvert (Java[15]) et une architecture logicielle standard (OSGi[16]), d'autre part l'utilisation d'une plateforme de versionnage des sources du logiciel, qui permet la traçabilité de l'attribution et de la datation de toute modification apportée aux sources et offre la possibilité de revenir à une version antérieure, quelle que soit sa date.

Conçue dès l'origine comme devant être capable d'exploiter des corpus textuels richement encodés en XML-TEI et annotés finement au niveau des mots, la plateforme TXM a pu utiliser la BFM comme corpus de validation de ses capacités d'intégration

11. La licence GNU GPL V3. Voir en ligne : http://www.gnu.org/licenses/gpl-3.0.fr.html.
12. Voir en ligne : http://www.r-project.org.
13. Voir en ligne : http://cwb.sourceforge.net.
14. Voir en ligne : https://eclipse.org.
15. Voir en ligne : https://www.jcp.org.
16. Voir en ligne : http://www.osgi.org.

et d'exploitation de corpus textuels riches en encodage et annotations.

La chaîne analytique de la BFM commence par un processus d'importation des fichiers sources encodés en XML-TEI dans la plateforme TXM à l'aide d'un module d'importation de sources appelé « XML-TEI BFM ». Ce module a été développé spécialement pour ce corpus à partir de la documentation des pratiques d'encodage XML-TEI des textes de la BFM telle qu'elle est publiée sur le site du projet de la Base. Il est chargé d'interpréter les fichiers source de façon à construire le « modèle de corpus » exploité par TXM. Les métadonnées nécessaires et utiles à l'analyse des textes sont extraites des en-têtes TEI, les éléments TEI pertinents pour l'analyse (comme par exemple les éléments <q> contenant le discours direct) sont indexés et certaines informations sont projetées au niveau des unités lexicales afin de simplifier les requêtes de recherche. D'autres éléments (comme les notes éditoriales) sont exclus de la surface du texte afin de ne pas être mélangés avec le corps du texte dans les recherches et les décomptes, mais sont intégrés aux éditions pour aider à la lecture des textes. Les éditions sont paginées en fonction des sauts de page encodés dans les sources numérisées. Une fois importée dans TXM à l'aide de ce module d'importation, la BFM bénéficie de tous les services d'analyse offerts par la plateforme dans sa version pour poste ou dans sa version portail. Le portail BFM[17] est un portail TXM hébergeant le corpus BFM. Il offre des services supplémentaires par rapport à la version pour poste de personnalisation de pages d'accueil ou de documentation, de création de comptes utilisateurs et de contrôle d'accès texte par texte.

Conclusion : une synergie entre les chaînes philologique et analytique pour une ressource libre

Aujourd'hui la BFM est consultée et analysée au moyen d'un logiciel libre (la plateforme TXM) et offre un accès libre aux sources de ses textes. Ces sources sont établies par une chaîne philologique complète et ouverte, de façon analogue à la chaîne

17. Voir en ligne : http://txm.bfm-corpus.org.

d'analyse qui repose sur le logiciel TXM, lui-même développé en *open-source*. L'emboîtement entre ces deux chaînes est rendu possible par un usage précis du standard de représentation des textes XML-TEI. Développée à l'origine pour l'échange de représentations numériques de textes entre partenaires, la TEI commence donc à mettre en relation des projets d'établissement de corpus de textes avec des projets de développement d'outils d'analyse et d'exploitation qui relèvent pourtant souvent de communautés de recherche très différentes en termes d'objectifs et de mode de fonctionnement. L'adoption parallèle d'un mode de fonctionnement ouvert par les deux chaînes pour faciliter la mutualisation et la traçabilité des développements (établissement de texte d'un côté, implémentation de méthode de l'autre) nous semble être un gage de pérennité et de scientificité des travaux pouvant être réalisés à l'aide de la BFM.

Références bibliographiques

Bédier, Joseph, « La tradition manuscrite du *Lai de l'Ombre*, réflexions sur l'art d'éditer les anciens textes », *Romania*, n° 54, 1928, p. 161-196 ; p. 236-356.

Bertrand, Lauranne, Lavrentiev, Alexei, Pincemin, Bénédicte, Guillot, Céline, Heiden, Serge et Lascar, Justine, *Tutoriel TXM pour la BFM*, Version 2.0, Lyon, ENS de Lyon, 2014. En ligne : http://txm.bfm-corpus.org/files/Tutoriel_TXM_BFM_V1.pdf.

Frappier, Jean (éd.), *La Mort Artu*, Genève/Paris, Droz/Minard, 1964.

Guerreau, Alain, *L'Avenir de la philologie. Textes anciens et domaine public*. En ligne : halshs-01112213, 2015.

Guillot, Céline, Lavrentiev, Alexei, Rainsford, Thomas, Marchello-Nizia, Christiane et Heiden, Serge, « La "philologie numérique" : tentative de définition d'un nouvel objet éditorial », dans Trachsler, Richard, Duval, Frédéric et Leonardi, Lino (dir.), *Actes du XXVII^e Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013). Section 13 : Philologie textuelle et éditoriale*, 2017.
En ligne : http://www.atilf.fr/cilpr2013/actes/section_13/Guillot_Heiden_Lavrentiev_Marchello-Nizia_Rainsford.pdf.

Guillot, Céline, Heiden, Serge, Lavrentiev, Alexei et Pincemin, Bénédicte, « L'oral représenté dans un corpus de français médiéval (IXe-XVe) : approche contrastive et outillée de la variation diasystémique », dans Jeppesen Kragh, Kirsten et Lindschouw, Jan (dir.), *Les Variations diasystémiques et leurs interdépendances dans les langues romanes. Actes du colloque DIA II à Copenhague (19-21 novembre 2012)*, Strasbourg, Éditions de linguistique et de philologie, 2015, p. 15-28.
En ligne : halshs-00760647.

Lebart, Ludovic et Salem, André, *Statistique textuelle*, Paris, Dunod, 1994.

Marchello-Nizia, Christiane, Lavrentiev, Alexei et Guillot-Barbance, Céline, « Édition électronique de la *Queste del saint Graal* », dans Trotter, David (dir.), *Manuel de la philologie de l'édition*, Berlin/Boston, De Gruyter, 2015.

Margoni, Thomas et Perry, Mark, « Scientific and Critical Editions of Public Domain Works: An Example of European Copyright Law (Dis)Harmonization », *Canadian Intellectual Property Review*, n° 27, 2011, p. 157. En ligne : http://ssrn.com/abstract=1961535.

Plouzeau, May, « À propos de *La Mort Artu* de Jean Frappier », *Travaux de linguistique et de philologie*, n° 32, 1994, p. 207-221.

Schøsler, Lene, « Historical corpora. Problems and Methods », dans Bozzi, Andrea, Cignoli, Laura et Lebrave, Jean-Louis (dir.), *Digital technology and philological disciplines*. *Linguistica computazionale*, t. XX-XXI, Pisa/Roma, Istituti editoriali e poligrafici internazionali, 2004, p. 455-472.

Stein, Achim et Prévost, Sophie, « Syntactic Annotation of Medieval Texts: the Syntactic Reference Corpus of Medieval French (SRCMF) », dans Bennett, Paul, Durrell, Martin, Scheible, Silke et Whitt, Richard (dir.), *Corpus Linguistics and Interdisciplinary Perspectives on Language (CLIP)*, n° 3, « New Methods in Historical Corpora », 2013, p. 275-282.

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**