REVUE DE
LINGUISTIQUE
FRANÇAISE
DIACHRONIQUE

DIACHRONIQUES

7
2017

LES ÉTATS ANCIENS DES LANGUES À L'HEURE DU NUMÉRIQUE

Ducos – 979-10-231-2156-8

PUPS

7

DIACHRONIQUES

REVUE DE LINGUISTIQUE FRANÇAISE DIACHRONIQUE

LES ÉTATS ANCIENS DES LANGUES À L'HEURE DU NUMÉRIQUE

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

12 €

http://pups.paris-sorbonne.fr

Maquette : www.stephanemercier.fr

LES ÉTATS ANCIENS DES LANGUES À L'HEURE DU NUMÉRIQUE

Les états anciens des langues à l'heure du numérique

PUPS

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

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

PDF complet – 979-10-231-2155-1

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

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

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

Tél. (33) 01 53 10 57 60

sup@sorbonne-universite.fr

sup.sorbonne-universite.fr

Présentation

Joëlle Ducos
EA 4509 STIH
Université Paris-Sorbonne

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

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

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

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

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

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

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

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

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

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

3. En ligne : www.crealscience.fr.

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