REVUE DE
LINGUISTIQUE
FRANÇAISE
DIACHRONIQUE

7
2017

DIACHRONIQUES

LES ÉTATS ANCIENS DES LANGUES À L'HEURE DU NUMÉRIQUE

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

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

Herméneutique des similarités dans le *DFSM* : une expérience

Xavier-Laurent Salvador, Fabrice Issac & Marco Fasciolo
Université Paris XIII

Contexte

Nous nous proposons de donner ici les éléments théoriques et la méthodologie permettant l'élaboration d'un outil automatique capable d'explorer un dictionnaire en langage naturel. Nous nous intéresserons, pour ce faire, à la constitution d'une ressource numérique qui enregistre autant les propriétés sémantiques inhérentes au lexique à un instant donné de son histoire qu'un positionnement relatif des unités lexicales par rapport à l'ensemble des mots de la langue. S'agissant d'explorer un dictionnaire en langage naturel, nous distinguons de fait deux degrés dans l'identification de relations :

1. un niveau fondamental, qui identifie l'existence d'une connexion entre les mots ;
2. la caractérisation de ce lien de connexion.

Traditionnellement, les outils d'accès aux entrées d'un dictionnaire utilisent (i) l'ordre alphabétique des vedettes, (ii) un moteur de recherche, plus ou moins sophistiqué. Le premier revient à une simple transposition de l'utilisation d'un dictionnaire physique, le second permet d'effectuer des recherches potentiellement très fines, mais nécessitant une bonne connaissance de la macrostructure. Par ailleurs aucune information contextuelle n'est fournie, et le résultat prend la forme d'un article isolé.

Afin de contextualiser l'ensemble des informations, et non pas seulement la vedette, on pourrait concevoir l'élaboration

d'une représentation qui ait comme base celle que nous faisons de notre propre lexique, le lexique mental. En plus de relier sémantiquement les concepts les uns aux autres, l'accès à l'information linguistique s'en trouve amélioré. Lors de la production d'un discours, le choix d'un mot ou d'une expression se fait à un rythme très soutenu, avec un taux d'erreur très faible[1]. Des expériences en psycholinguistique portant sur l'organisation du lexique mental (Aitchison, 2003) montrent que les relations entre les éléments du lexique sont de deux types, soit intrinsèques, soit associatives (Levelt, 1989).

Les relations intrinsèques, ou catégorielles, contiennent des informations linguistiques sur l'unité lexicale elle-même. On peut décomposer les relations intrinsèques en relations :

– sémantiques (elles concernent la synonymie, l'antonymie, l'hyperonymie, l'hyponymie, la méronymie) ;
– morphologiques ;
– phonologiques (pour les mots commençant ou se terminant par les mêmes phonèmes : p. ex. *travail* et *traverse*).

Les relations associatives, de leur côté, regroupent les unités dont la fréquence d'apparition dans un même contexte est importante. À *ouvrier* on associera par exemple *usine* ou *travail*. Ce type de relations dépend de la connaissance qu'a le sujet du monde qui l'entoure. Les relations entre les éléments mentaux forment ainsi un réseau dans lequel les nœuds ne sont pas les mots eux-mêmes, mais leurs sens particuliers. Dans cette représentation, une collocation n'est pas stockée comme une association de mots, mais comme une unité individuelle à part entière.

L'analyse par la grammaire traditionnelle des relations entre les lexies en termes d'antonymie ou de synonymie s'intéresse au second niveau. En revanche, l'analyse grammaticale proposée par la grammaire scolaire, voire – dans une moindre mesure – par

1. Le choix d'un mot se fait au rythme de 2 à 5 mots par seconde avec un taux d'erreur inférieur à 1 pour 2 000, alors qu'on estime le nombre de mots que connaît un locuteur moyen à 35 000. Cette vitesse est encore plus élevée en lecture (Levelt, 1989).

la stylistique, en termes de champs sémantiques ou notionnels, s'intéresse au premier niveau. C'est à ce titre que, dans le discours critique du texte littéraire, *naufrage* et *débarquement* peuvent être analysés comme appartenant tous deux au « vocabulaire de la mer[2] » indépendamment du fait que ces deux termes renvoient à des réalités connotées de manière pour le moins antinomique, puisque le premier évoque l'échec et le second, le succès. Toutefois, la « relation de pertinence » entretenue par le discours critique entre cette paire nominale et l'hyperonyme maritime est validée par la présence manifeste, dans une décomposition des unités sémantiques de chacun des termes, du sème /mer/.

Nous présupposons donc qu'il est pertinent d'établir deux niveaux d'analyse des relations entretenues par les mots entre eux : la connexion d'une part, et d'autre part la caractérisation du degré de similarité. Le premier niveau est binaire : la connexion existe ou n'existe pas. Le second niveau, en revanche, est scalaire : il va d'une similarité nulle, d'un certain degré de pertinence, jusqu'à une similarité maximale, qui coïncide avec la synonymie. Dans le domaine de la dictionnairique, la distribution du vocabulaire en « domaines de sens » renvoie à une classification des objets de la langue en fonction de leur apparition dans des contextes d'emplois que le corpus d'exemples rend manifeste. Ainsi, si le mot *gingembre* est classé dans le domaine de la botanique (c'est une plante) et de la pharmacopée (c'est un ingrédient), c'est en vertu du fait que son contexte d'emploi fréquent justifie la caractérisation de la polysémie du terme. Il s'agit donc là d'une analyse de discours opérée par le lexicographe, et non pas de la manifestation d'une propriété inhérente ni au mot de la langue, ni à l'objet. Ainsi, ce n'est pas parce que l'on consomme du gingembre qu'il est consommable, mais l'emploi du terme dans des livres de recettes et de médecine témoigne de ce que cet aspect « consommable » est bien une propriété discriminante de la plante, et que le mode, ou plutôt la destination, de cette consommation varie en fonction des effets. À ce stade de la description de la relation entretenue

2. Voir Georges Molinié (1995).

par les sens d'une unité polysémique, il semble que nous pouvons considérer qu'au niveau (1) il existe une connexion entre *gingembre* (bot.) et *gingembre* (pharm.), et que cette connexion peut être caractérisée par (2) un haut degré de similarité, sans pour autant parler, en l'occurrence, de synonymie.

Il existe en ancien français le terme « zédoaire », répertorié par le *Dictionnaire du français scientifique médiéval* (*DFSM*), qui présente les mêmes propriétés que le gingembre aussi bien du point de vue de la nature (c'est une racine), du goût (c'est aigre) et de la distribution des emplois. Il existe donc (1) une forte connexion entre *gingembre* (bot.) et *zédoaire* (bot.), et cette connexion peut être caractérisée par (2) un très haut degré de similarité : c'est une synonymie. En revanche, la distribution des relations de connexion entretenues par les unités croisées par domaine est moins forte. L'ensemble du corpus des relations entretenues par chacun des niveaux sémantiques de chaque unité polysémique, et des relations entretenues entre chaque unité est documenté, dans la ressource comme dans le dictionnaire, par le corpus des exemples qui illustrent le travail définitoire du lexicographe.

La caractérisation des relations de similarité, ou *calcul de distance*, peut être schématisé de la façon suivante :

Mot 1 – (connexion, degré de similarité) – Mot 2

La similarité est une information pertinente pour la description d'un état de langue, car elle renseigne sur « ce que veut dire le lexicographe » et qui n'apparaît pas évident pour le lexicographe lui-même. Cela est d'autant plus vrai lorsque l'on travaille sur un état ancien de langue d'où les mots, mais aussi les choses ont disparu aujourd'hui et lorsque, comme dans notre cas, il y a deux niveaux de lexicographes. Nous rejoignons finalement la pensée du linguiste John Langshaw Austin qui écrivait (Austin, 1956) :

> *[…] our common stock of words embodies all the distinctions men have found worth drawing, and the connexions they have found worth marking, in the lifetimes of many generations: these surely are likely to be more numerous, more sound, since they have stood up to the long test of the survival of the fittest,*

> *and more subtle, at least in all ordinary and reasonably practical matters, than any that you or I are likely to think up in our armchairs of an afternoon – the most favoured alternative method.*

La méthode que l'auteur propose pour élucider les « distinctions conceptuelles que les hommes ont pu marquer au cours de nombreuses générations » est la suivante :

> *First we may use the dictionary – quite a concise one will do, but the use must be thorough. Two methods suggest themselves, both a little tedious, but repaying. One is to read the book through, listing all the words that seem relevant; this does not take as long as many suppose. The other is to start with a widish selection of obviously relevant terms, and to consult the dictionary under each: it will be found that, in the explanations of the various meanings of each, a surprising number of other terms occur, which are germane though of course not often synonymous. We then look up each of these, bringing in more for our bag from the « definitions » given in each case; and when we have continued for a little, it will generally be found that the family circle begins to close, until ultimately it is complete and we come only upon repetitions. This method has the advantage of grouping the terms into convenient clusters – but of course a good deal will depend upon the comprehensiveness of our initial selection.* (*Ibid.*)

Cette réflexion, que nous appliquons au domaine de la dictionnairique, permet d'établir des connexions entre les mots du lexique et des degrés de similarité entre eux.

Enjeux épistémologiques

Du côté philologique

Une partie de notre travail a pour point de départ la réalisation du *DFSM*. Cet ouvrage, consacré à la langue de spécialité médiévale, repose sur un corpus scientifique constitué par des traducteurs ou des vulgarisateurs qui travaillent presque comme des lexicographes au Moyen Âge, ce qui suppose de prendre en compte ces réflexions et pratiques de locuteurs.

La conception du dictionnaire permet ainsi une réflexion sur les processus de genèse et de néologie d'une terminologie.

Il doit aboutir à une meilleure connaissance du français médiéval dans ses usages spécialisés, mais aussi donner des outils d'étude et une méthodologie pour apprécier les modalités de création et en rendre compte dans un dictionnaire conçu comme une représentation d'évolution linguistique. Le développement technique au sein de l'équipe CréaLScience, de plus, complète les recherches épistémologiques, rendant au demeurant hommage à l'étymologie même du terme « informatique » – qui donne forme et corps aux pensées dont la nature est par essence informe.

Le *DFSM* présente donc une structure lexicographiquement stratifiée : d'un côté, il y a les rédacteurs médiévaux des définitions ; de l'autre côté, il y a l'équipe actuelle qui traduit ces définitions en français contemporain. Tout l'enjeu épistémologique du dictionnaire réside dans l'articulation de ces deux niveaux lexicographiques et dans la mise en forme des analyses du corpus multilingue qui en découle. Au premier niveau, parler de lexicographes pour les rédacteurs médiévaux peut étonner. Cet étonnement, cependant, est hors de propos. On trouve dans la littérature encyclopédique médiévale, comme dans les traductions de la Bible d'ailleurs, un ensemble de marqueurs caractéristiques de la glose encyclopédique : *c'est assavoir, si come dit le maistre, sicome est*,... Ces marqueurs embrayent sur une prise de parole originale des traducteurs afin d'introduire des paraphrases explicatives de certains mots, renvoyés en mention autonymique (Authier-Revuz, 1995 et 1998). Lorsque, dans un discours français, un terme français est autonyme et assorti d'une glose encyclopédique, n'est-on pas en droit de considérer que le traducteur a œuvré, en l'occurrence, en spécialiste de la langue ? Ne peut-on même considérer qu'il agit en lexicographe lorsqu'il renseigne son lecteur sur les emplois de *zurbe* ou sur les façons d'observer les apophtegmes ?

Dans le cadre de la définition des rapports qu'il entretient avec l'énonciation du texte traduit, le traducteur est indéniablement un sujet de l'énonciation. Il devient une forme d'interface de coïncidence entre le vouloir-dire du texte source et les horizons d'attente du texte traduit. Sa réflexion sur le lexique de la langue

source s'apparente fortement à celle menée par un lexicographe s'agissant de la recherche de la nature, du sens et des conditions syntaxiques d'avènement de ce dernier dans la langue cible.

Enjeu dictionnairique

Le traducteur intervient donc comme un lexicographe à part entière, chaque emploi qu'il fait de chaque unité du système étant le fruit d'une réflexion issue à la fois d'un enseignement universitaire et d'un souci d'enseignement scientifique ; et le texte traduit apparaît comme un recueil de prises de position lexicographiques à partir de relations méronymiques établies non pas entre deux langues, mais entre le latin, langue de communication savante, et le français en situation de diglossie. La traduction se conçoit *a fortiori* comme un discours rapporté, en témoigne le discours attributif qui l'introduit : « x (= l'auteur du texte original) dit que ». L'autorité de l'auteur / traducteur sur sa propre production est ainsi mise entre parenthèses, puisque le traducteur se présente comme citant et reprenant les mots du lexique de l'auteur premier, soit qu'il le suive au point de calquer le vocabulaire français sur le lexique latin – c'est le calque savant – soit au contraire qu'il le juge en inadéquation avec le lecteur français, et qu'il l'abandonne – c'est la glose paraphrastique. Ainsi, entre paroles rapportées et appropriation du discours d'un autre, la traduction scientifique engagée dans une réflexion métalinguistique sur le lexique construit un discours autonome. À l'intérieur de ce discours, nous remarquons des nœuds opaques qui posent le problème de la mention autonymique de quelques unités sémantiques, et c'est là sans doute que nous rejoignons de plein pied la problématique de la néologie et de son repérage. Un tel phénomène définit le paradoxe des unités lexicales placées en mention autonyme en contexte traductologique. Le traducteur travaille sans cesse à rendre son énoncé pertinent dans le cadre d'un enseignement fondé sur la transmission sémantique. Nous retrouvons dans ce phénomène la problématique du dictionnaire bilingue : construire un vocabulaire spécialisé (une définition), dont le sens est donné explicitement par l'introduction de xénismes en mention autonymique (l'entrée principale) et

dont le contexte se charge de saturer le signifié par le biais de mentions correctives.

Au second niveau lexicographique (celui de l'équipe actuelle), deux problématiques principales émergent.

La première d'entre elles concerne le rapport avec le premier niveau de lexicographes-traducteurs, à savoir l'identification de « ce qu'ils veulent dire ». Cette question, cruciale, ne peut pas être laissée à l'appréciation de chacun, et un ensemble de contraintes doivent peser sur la mise en forme du corpus des définitions. C'est un élément d'autant plus important que le nombre d'auteurs est élevé. L'équipe du *DFSM* doit maîtriser le métalangage de description et s'accorder sur un ensemble de termes propres à la description lexicographique. L'idée consiste à créer un lexique des termes autorisés dans le cadre de l'article de manière à ce que chacun de ces termes fonctionne comme un signal adressé à l'*uptake* du calculateur pour l'enregistrement d'un ensemble de traits descripteurs des relations sémantiques et des propriétés.

Il est également essentiel d'assurer l'autosuffisance du dictionnaire par le recoupement des données. Ainsi, le traitement de la néologie sera d'autant plus facilité qu'il existera un ensemble de lemmes orphelins, ensemble qu'il sera alors aisé de traiter.

La mise en œuvre de l'architecture métalexicographique s'accorde avec un travail sur l'article. Une définition doit intégrer des propriétés intrinsèques : « ce qu'est le x », et des propriétés extrinsèques : « à quoi sert x », par exemple dans le cas d'un instrument, ou encore « d'où vient x » dans le cas d'une maladie. Ainsi, la définition de *zodiaque* ne peut-elle pas reprendre les termes d'un dictionnaire contemporain[3] :

> *Zodiaque* : cercle situé sur le plan de l'écliptique et autour duquel évoluent le Soleil, la Lune et les planètes.

3. Larousse en ligne : http://www.larousse.fr/dictionnaires/francais/zodiaque/83170?q=zodiaque#82170 [consulté le 21 juin 2017].

Elle doit restituer le champ des savoirs relevant de la période couverte. En l'occurrence, dans la définition contemporaine, le terme *écliptique* est hétérogène et anachronique, alors que l'absence des renvois vers *astre* ou *maison*, qui sont les hyponymes directs de *zodiaque* dans l'astronomie médiévale (Boudet, 2006), est dommageable pour la compréhension du concept médiéval.

L'autre problématique, qui touche le second niveau lexicographique susmentionné, concerne la question des « nomenclatures », terme par lequel nous entendons « la liste des lemmes présents dans le dictionnaire ». Dans le cadre de la rédaction d'un dictionnaire d'histoire des sciences anciennes, la première difficulté que nous rencontrons réside dans la constitution de ladite nomenclature (Ducos, 2006). La forme du lemme n'est pas problématique lorsque la langue du dictionnaire est homogène avec la langue décrite, même lorsqu'il y a xénisme. Les choses peuvent devenir plus compliquées dans le cadre de la rédaction d'un dictionnaire bilingue (Kocourek, 1991). Mais elle sont plus problématiques encore s'agissant de la période médiévale, où la langue connaît trois degrés de variation : une variation dans le temps, variation diachronique ; une variation dialectale, variation diatopique ; une variation idiolectale, liée à l'intervention des copistes qui sont eux-mêmes des adaptateurs du texte original. À celles-là s'ajoute la variation graphique, qui peut en premier lieu laisser croire à l'existence de plusieurs lexèmes, là où il ne s'agit que d'une variante dialectale, mais également rendre difficile la création de règles claires pour le choix du terme à intégrer à la nomenclature ; ainsi, au même moment, en France, trouve-t-on *bourraiche, bourrache, borrache, borraige* pour le lemme « *bourrache* ».

Du côté de l'exploration dictionnairique

Dans le cadre de notre travail, une question se révèle cruciale : comment explorer un dictionnaire de langue naturelle ? L'interrogation peut être précisée de cette manière : comment relier les définitions d'un tel dictionnaire ? Il existe des solutions impraticables. Considérons les définitions suivantes :

Couteau :
Instrument tranchant servant à couper, composé d'une lame et d'un manche.

Chat :
Petit mammifère familier à poil doux, aux yeux oblongs et brillants, aux oreilles triangulaires et griffes rétractiles, qui est un animal de compagnie.

Sosie :
Personne qui a une parfaite ressemblance avec une autre.

Envieux :
Qui éprouve de l'envie.

Vendre :
Céder à quelqu'un en échange d'une somme d'argent.

Rapidement :
D'une manière rapide.

Une première hypothèse consisterait à les rapprocher sur la base de leur forme par rapport à la catégorie grammaticale de l'entrée. Dans la définition d'un nom comme *couteau*, par exemple, on peut distinguer un hyperonyme (un autre nom) et une portion de différences spécifiques, alors que dans la définition d'un adjectif comme *envieux* on peut distinguer un terme de relation en première position (en l'occurrence le relatif *qui*…) et le corps de la définition, qui contient le signifié.

Une deuxième hypothèse consisterait à rapprocher ces définitions sur la base de leur forme par rapport au type d'objet décrit. La catégorie des instruments, par exemple, sera définie par un hyperonyme suivi de la fonction, alors que les éléments des listes naturelles (comme les fleurs) recevront une description centrée sur leur forme. Ce type de solution, cependant, n'est pas envisageable. Tout d'abord, ce niveau de finesse est un obstacle pour l'exploitation automatique. Comparons les définitions de *sosie* (qui est un nom relationnel) et *couteau* (qui est un nom ponctuel). Le premier terme de ces définitions (à savoir, *instrument* et *personne*) est-il un vrai hyperonyme ? Certes, instrument est un hyperonyme de *couteau*, car l'assertion « un couteau est un instrument » est informative. Mais *personne* ne peut pas être considéré comme un hyperonyme de *sosie* : c'est,

dans la définition, un argument de « avoir une ressemblance parfaite » (le signifié de *sosie*), qui remplit la fonction d'un pronom. La définition de *sosie* est donc plus proche de celle de *envieux* que de celle de *couteau*. Il nous paraît très difficile de modéliser informatiquement une telle intuition.

Ensuite – ce qui est plus important – le rapport entre la forme de la définition et la catégorie grammaticale de la vedette, ou le rapport entre la forme de la définition et le type d'objet décrit, sont précisément le genre de choses qu'on voudrait étudier à travers une exploration du dictionnaire. Si l'on veut fournir des outils pour une telle exploration, il faut par conséquent se fonder sur des critères plus élémentaires.

Nous proposons de nous inspirer des travaux menés dans le champ de la désambiguïsation, et notamment de l'algorithme de Lesk.

> *I am trying to decide automatically which sense of a word is intended (in written English) by using machine readable dictionaries, and looking for words in the sense definitions that overlap words in the definition of nearby words. […] To consider the exemple in the title [How to tell a « pine cone » from an « ice-cream cone »], look at the definition of pine in the* Oxford Advanced Learner's Dictionary of Current English*: there are, of course, two major senses, « kind of evergreen tree with needle-shaped leaves… » and « waste away through sorrow or illness… » And cone has three separate definitions: « solid body which narrows to a point… » « something of this shape whether solid or hollow… » and « fruit of certain evergreen trees… » Note that both evergreen and tree are common to two of the sense definitions: thus a program could guess that if the two words pine cone appear together, the likely senses are those of the tree and its fruits*. (Lesk, 1986)

Le problème ici abordé peut être formalisé de la façon suivante. Soit :

> un mot *M* qui connecte deux sens différents : s_1 et s_2

Imaginons rencontrer *M* dans deux co-textes :

> $C_1 = m_{1.1}, m_{1.2}, \ldots$ où les $m_{1.i}$ sont des mots
> $C_2 = m_{2.1}, m_{2.2}, \ldots$ où les $m_{2.i}$ sont des mots

En ce cas, nous sommes confrontés à des trigrammes, comme :

$m_{1,j}\ M\ m_{1,j+2}$ et $m_{2.k}\ M\ m_{2.k+2}$

La question est la suivante : comment déterminer le sens de *M* dans chaque séquence ? Si à chaque sens s_1 et s_2 de *M* correspond une définition D différente : D_1 et D_2, alors on cherchera la présence des trigrammes des co-textes dans les définitions, et on choisira le sens correspondant.

Cette solution présuppose une homologie entre l'emploi d'un mot (*i.e.* son occurrence dans un co-texte) et les mots de la définition qui préside à cet emploi. Ce présupposé mérite qu'on s'interroge : quel est le rapport existant entre les mots employés dans la définition et dans les exemples ? En réalité, ce rapport ne peut pas se réduire à un simple recouvrement, parce qu'une définition doit fournir un schéma (un modèle) qui permette de générer tous les emplois possibles.

Cependant et quoi qu'il en soit, la solution proposée par Michael Lesk nous offre une suggestion importante. Explorer un dictionnaire en langage naturel signifie : (i) prendre une entrée ; (ii) décomposer sa définitions en mots ; (iii) en envisageant chaque mot selon trois dimensions :

– occurrences (ou *tokens*) ;
– lemmes ;
– catégories morphosyntaxiques ;

et enfin (iv) regarder comment les séquences de mots se distribuent dans les définitions des autres entrées.

Le mode opératoire est le suivant : à partir d'une vedette, on décompose sa définition en atomes de sens, et on explore la façon dont ceux-ci se propagent dans les définitions des autres vedettes. Cela revient à explorer leurs échos sémantiques dans le dictionnaire. Il est à noter que cette exploration n'est pas statique et globale, mais dynamique et locale, car elle varie selon le point d'entrée choisi.

Remarquons tout d'abord que, d'un point de vue linguistique, la solution des « segments partagés » est grossière, car elle se fonde sur des séquences de mots sans présupposer aucune

structure. Il faut cependant noter que ce défaut est corrigé par un principe de coopération *à la Grice*. Nous supposons que les lexicographes sont des êtres rationnels et coopératifs, et donc que leurs définitions ne sont pas des suites de mots mis au hasard, mais bien des textes cohérents. Cette assomption autorise à émettre l'hypothèse que le critère de rapprochement des définitions, en pratique, peut se réduire simplement aux plus longs segments partagés. Nous faisons le pari que ces segments constituent des morceaux pertinents de la syntaxe des définitions. Le principe de coopération susmentionné joue un rôle fondamental en TAL, et la plupart des critiques qui reprochent au traitement automatique du langage un manque d'esprit linguistique (ou une perspective aveuglément extensionnelle) gomment ce point.

Il faut ensuite remarquer que l'écho sémantique que nous évoquions se différencie de la troisième investigation dictionnairique, envisagée par Jean Pruvost, en ce qui concerne le focus. Cette dernière est décrite de la façon suivante :

> La troisième approche est celle qui correspond à l'analyse des différents emplois du mot *norme* tout au long du dictionnaire : il s'agit d'établir un concordancier de l'usage du mot dans le corpus défini par tous les articles du dictionnaire où on trouvera le mot recherché. Ainsi apparaît l'usage dictionnairique du mot, au-delà de l'article qui lui est consacré, révélant par les co-textes de ce mot, c'est-à-dire ce qui le précède et ce qui le suit, une palette d'emplois, d'usages, propres à mieux en cerner la nature sémantique et syntaxique. Les agents de la norme que sont les dictionnairistes livrent ainsi à leur insu une illustration sémantique et syntaxique du mot qui complète heureusement l'article consacré à un mot. (Pruvost, 2005)

La troisième investigation considère le dictionnaire comme un corpus des environnements distributionnels du mot en examen : ce mot est le pivot du concordancier construit en explorant les définitions des autres mots. Dans l'approche de l'écho sémantique, en revanche, le pivot n'est pas constitué par la vedette examinée, mais bien par les mots de sa définition qui peuvent être communs avec celles des autres vedettes du dictionnaire.

L'articulation entre les deux niveaux lexicographiques susmentionnés et la perspective sur l'exploration dictionnairique implique que, concrètement, l'accès aux notions part d'un travail sur les mots eux-mêmes, et sur l'encadrement de ce travail.

Le résultat se présentera sous la forme d'un dictionnaire traditionnel doublé d'une représentation des relations sémantiques au sein d'un maillage partant des catégories épistémologiques et techniques pré-construites exposées plus haut, pour arriver à des champs sémantiques où la nature des relations envisagées sera représentée par des codes de couleurs. Ainsi, partant du domaine des sciences qu'est la « Médecine », l'utilisateur déroulera les champs constitutifs du domaine : Médecine, Anatomie, Botanique, Chirurgie.

Partant de « Botanique », un ensemble d'entités :

> Botanique → Objet de la science (phénomène observable, phénomène positif, phénomène négatif), instrument, corps constitué, partie du corps, action, état, événement, qualités (positives, négatives), caractères.

Enfin, de « phénomène observable » on retrouvera l'ensemble des champs sémantiques de chaque plante faisant apparaître des relations de synonymie entre les nomens et les noms français (*borrago*, *bourrache*) ou entre des unités lexicales dialectales (*borrache*, *bourache*). On peut ainsi imaginer établir des relations de méronymie ou d'hyperonymie entre des termes attestés dans le corpus, disparus du français moderne, représentant des *realia* non répertoriées aujourd'hui.

Parmi les premières conséquences d'une telle démarche, il y a la double possibilité de naviguer à travers une ressource et de la valider. En fait, plus les termes d'une ressource seront intégrés, plus il sera aisé d'en conclure que sa construction est valide. Une seconde conséquence renvoie à l'interopérabilité des ressources. Ainsi, les calculs de distance partant d'un lexique donné de l'ancien français, par exemple, pourront progressivement s'étendre à d'autres ressources de l'ancienne langue, voire à des ressources modernes qui intégreraient dès lors la similarité *gingembre* et *zédoaire*.

Protocole

Le dictionnaire

« L'investigation dictionnairique » désigne l'ensemble des moyens destinés à organiser la représentation du contenu d'un dictionnaire existant ou en cours de rédaction afin d'en faciliter l'accès à l'utilisateur final, ce dernier pouvant être (i) un lecteur humain ou (ii) une machine. Dans le premier cas, l'investigation dictionnairique consiste à construire une interface informatique de représentation herméneutique des contenus d'un dictionnaire guidant le lecteur du connu vers l'inconnu (Issac et Salvador, 2010), avec pour objectif de réduire la part d'inaccessible dans la culture médiévale pour un lecteur néophyte. Dans le second cas, l'investigation dictionnairique consiste à automatiser la création de ressources à partir d'un dictionnaire en cours de rédaction de manière à réduire la distance qui existe traditionnellement entre la ressource informatique, impropre à la consultation humaine, et le « dictionnaire papier », inexploitable. Les points importants liés au projet consistent à réfléchir à l'organisation herméneutique du projet lexicographique et de la représentation des données pour l'homme et la machine.

Il existe de nombreux dictionnaires en ligne (Caruso, 2011), 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 complexes. Le processus de constitution est totalement déconnecté du processus de consultation. Le principe – ou scénario – le plus fréquemment rencontré en termes d'interface est un calque – ou 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.

Édition collaborative

Un dictionnaire est le fruit du travail de plusieurs rédacteurs. Le recours à l'outil informatique, et plus particulièrement à sa dimension collaborative *via* le réseau, est dans ce cadre tout à fait pertinent. La notion même de travail collaboratif utilisant les nouvelles technologies fait l'objet de nombreuses expérimentations, notamment dans la réalisation de données dictionnairiques ou encyclopédiques. L'encyclopédie *Wikipédia* est à ce titre emblématique : elle fait l'objet de nombreuses études comme de controverses (voir par exemple Barbe, 2010 ou Endrizzi, 2008), alors même que son modèle éditorial est en perpétuelle évolution. L'édition collaborative dans le cadre de *Wikipédia* s'appuie sur un outil et des règles éditoriales. Celles-ci sont au nombre de cinq, et désignées comme « principes fondateurs » :

> Les principes fondateurs de *Wikipédia* fixent les grandes lignes qui définissent *Wikipédia* et les conditions de son élaboration. Ils constituent le fondement intangible de toutes les règles et recommandations du projet et sont au nombre de cinq : encyclopédisme, neutralité de point de vue, liberté du contenu, savoir-vivre communautaire et souplesse des règles[4].

L'outil utilisé est un *wiki*, qui répond au nom de *Mediawiki* et propose, outre un mode de saisie de l'article encyclopédique lui-même, un historique de ses versions et une page de discussion. Nous caractérisons les systèmes d'édition collaborative d'après trois critères : rédacteurs, structure et publication (RSP). Nous les expliciterons ici en les illustrant avec l'exemple de *Wikipédia*.

– Rédacteurs : Quelle est la politique en ce qui concerne l'identification des rédacteurs ? *Wikipédia* propose soit l'anonymat, soit une identification faible.
– Structure : Quelles sont les contraintes portant sur la production du rédacteur, et comment sont-elles exercées ? Le langage interne de *Mediawiki* propose un certain nombre de balises structurelles et de mise en forme, mais aucune contrainte

4. Extrait de : https://fr.wikipedia.org/wiki/Wikipédia:Principes_fondateurs [consulté le 29 décembre 2021].

n'est imposée. La normalisation est donc effectuée non pas par le système, mais par les mises à jour et les corrections successives des relecteurs.

– Publication : Quelle est la politique appliquée à la publication des productions des rédacteurs ? *Wikipédia* n'impose pas de relecture *a priori* ; toute production peut être directement publiée. Les éventuelles relectures et modifications sont effectuées *a posteriori*, aucun contrôle n'est fait sur le degré d'expertise, seul le consensus valide (ou invalide) un article.

Nos besoins en termes de diffusion et de travail collaboratif nous ont tout naturellement amenés à choisir une architecture client / serveur web. L'outil développé utilise donc les technologies standards XML / (X)HTML / CSS / JavaScript du côté du client, et un moteur de base de données du côté du serveur.

Les informations à manipuler étant d'une part de nature complexe, et d'autre part variables suivant la nature de l'information lexicographique répertoriée – les macrostructures des dictionnaires de spécialités doivent refléter la spécificité du domaine visé –, il est impossible de construire une structure capable de rendre compte de tous les cas de figure à l'aide d'un gestionnaire de base de données classique. À l'inverse, le langage XML permet une grande latitude dans la structuration de l'information ; c'est donc sur cette technologie que s'est porté notre choix. L'ensemble est rendu accessible par un serveur de base de données, BaseX[5], développé à l'université de Konstanz et structuré autour d'un moteur Xquery. Voilà un exemple de requête rendant accessible la nomenclature des vedettes du dictionnaire dont l'initiale est « A » :

```
for $x in distinct-values(//entry)[./form/orth contains text «^A»
using wildcards]
  order by $x collation «?lang=fr»
  return <a>{$x}</a>
```

On distingue principalement trois types d'utilisateurs : les lecteurs, qui n'ont aucun contrôle sur les contenus, les rédacteurs

5. En ligne : https://basex.org.

et les administrateurs. Les rédacteurs sont des lexicographes ou des spécialistes du domaine qui disposent d'une interface de saisie dont la structure est imposée. La saisie effectuée, le rédacteur a la possibilité de solliciter une validation de l'article auprès d'un administrateur. Ce dernier peut demander des modifications à l'auteur, ou publier l'article en l'état. Tant qu'une fiche n'est pas validée, sa version antérieure et sa nouvelle version cohabitent jusqu'à la validation finale, qui rejette l'ancienne version dans les limbes. Il existe ainsi trois états :

– état corrigé et validé,
– état en cours de rédaction,
– dans les limbes.

L'ensemble des fiches est au format XML, compatible TEI[6]. Les balises utilisées sont les suivantes :

– form : décrit la forme de l'entrée, *i.e.* la vedette (balise orth) ainsi que cela a été évoqué dans la section précédente,
– gramGrp : décrit les informations grammaticales (nature et genre) attachées à l'entrée,
– etym : donne la référence étymologique,
– sense : liste l'ensemble des sens possibles, chacun étant associé à un domaine (attribut n), une définition (balise def), un ensemble de citations (balise cit), un ensemble de liens (balise xr).

Voilà un exemple de fiche :

```
<entry>
  <form>
    <orth>ZEDOAIRE</orth>
  </form>
  <gramgrp>
    <gram type=”pos”>subst.</gram>
    <gram type=”gen”>masc.</gram>
  </gramgrp>
  <etym>
    <bibl><etymosrc>FEW XIX, 201b</etymosrc></bibl>
```

6. *Text Encoding Initiative*.

```
      <mentioned>Zadwar</mentioned>
    </etym>
    <sense n="BOT.">
      <def>Graine aromatique qui ressemble au gingembre mais qui
      est d'un goût moins âcre et de meilleure odeur.</def>
      <note id=»Struct»>
        <xr><ref>zédoaire,Curcuma,zedoaria</ref></xr>
        <xr><ref>Hyponyme</ref></xr>
        <xr><ref>Cohyponyme</ref></xr>
        <xr><gloss>Texte encyclopédique</gloss></xr>
      </note>
      <cit>
        <quote>Le foie confortent ces choses xilobalsamum, carvi,
        cubebes, zedoaire, alemandes, chastaingne en petit numbre
        [...]</quote>
        <bibl><author>WATERFORD, COPALE,</author>Le Secré
        des Secrés,
        <nonit>fin xii<sup>e</sup>s., p. 131, LIII</nonit></bibl>
      </cit>
    </sense>
  </entry>
```

Le corpus est constitué de l'ensemble des fiches dont la définition n'a pas systématiquement fait l'objet d'une validation. Le principe d'association s'effectue à partir d'un calcul de distance, l'idée étant de comparer chaque fiche avec la totalité du corpus. La nature même des liens qui seront calculés dépend bien évidemment de la manière dont les informations sont sélectionnées dans chacune des fiches. Ne pas effectuer cette sélection reviendrait à mettre au même niveau tous les éléments de la macrostructure : le résultat obtenu ne serait dès lors pas susceptible d'être interprété. Il s'agit donc non pas de concaténer, au prétexte que plus il y a d'information, plus le résultat serait pertinent, mais bien de choisir pour, le cas échéant, combiner.

Notre expérimentation s'est concentrée sur les définitions proposées par le dictionnaire CréaLScience dans sa version non aboutie. Nous avons donc extrait du corpus total un sous-corpus ne contenant que la vedette et les définitions qui lui sont rattachées :

<resultat>AIGUISER = Rendre aigu, intense</resultat>
<resultat>ALBUGINÉ = Humeur aqueuse de l'oeil</resultat>
<resultat>ANARCOSITÉ = Pouvoir narcotique</resultat>
<resultat>APLOMB = À l'aplomb, verticalement»</resultat>
<resultat>ADHÉRER = S'unir, se souder</resultat>
<resultat>ALPHOS = Ulcération de la peau</resultat>
<resultat>AIABLE = Propre à, apte à</resultat>
<resultat>ASCARIDE = Ver ascaride</resultat>
<resultat>ANNULEUX = Formé d'anneaux</resultat>
<resultat>ABONDANT = Abondant</resultat>
<resultat>APOSTOLICON = Onguent dit en lat. apostolicum</resultat>
<resultat>ADMINISTRATION = Action de faire absorber</resultat>
<resultat>ACCÈS = Accès d'une affection morbide, paroxysme d'une fièvre</resultat>
<resultat>ATORNER = Préparer</resultat>
<resultat>AGE = Âge, portion déterminée de la vie d'un homme Phase de la lune</resultat>
<resultat>ATEMPRANCE = Modération Équilibre de la complexion, du tempérament</resultat>
<resultat>AUDITIF = Qui sert à l'audition</resultat>
<resultat>ARGILLEUX = Argileux, de la nature de l'argile</resultat>

Similarités

Les définitions, on le voit, sont disparates, allant d'un simple mot redondant par rapport à l'entrée (comme pour ABONDANT), à un ensemble de définitions (comme pour ÂGE). Cet état de fait est attendu, le dictionnaire étant en cours d'élaboration / validation. L'un des résultats escomptés est justement la détection semi-automatique des incohérences ou des erreurs tant lexicographiques que définitionnelles.

Chaque définition est analysée (normalisation de la casse, découpage en mots[7]) et associée à sa vedette. Puis un étiqueteur morpho-syntaxique est appliqué de manière à obtenir pour chaque mot son lemme et sa catégorie grammaticale.

Plusieurs niveaux d'association ont été envisagés :

7. Nous utilisons ici le terme *mot* bien que les seuls critères retenus soient d'ordre typographique.

1. le lemme,
2. le mot,
3. l'unité polylexicale (mots simples et lemmes).

Les choix effectués déterminent là encore le sens donné, et par conséquent la grille d'analyse à utiliser. Ainsi la co-occurrence des termes de la recherche dans la définition, comme « plante » (subst.) et « à » (prép.), met en relation des vedettes de plantes individuelles (le géranium et la passiflore), alors que le même patron de recherche « plante à », en tant que lemme, met aussi en relation non plus un individu, mais des familles de plantes (plante(s) à feuilles caduques, plante(s) à feuilles lancéolées, plante(s) à feuilles persistantes). Il est ainsi possible d'envisager de combiner différents traits morphologiques ou sémantiques en imposant *a priori* certains motifs :

> [« à feuille »+ADJ+« utilisé dans »] où ADJ est un adjectif [ingrédient] ou [instrument]

Le résultat obtenu est un fichier XML représentant un graphe où les nœuds sont les vedettes (par exemple <li name=»LIEURE» num=»3»/>), et la nature du lien de similarité quand celui-ci est pertinent (par exemple <li name=»sans levain» num=»32» key=»key»>). Chaque nœud se voit associer un identifiant numérique qui est utilisé pour définir les liens. Ainsi <li deb=»1» fin=»21»/> relie l'entrée LIEURE avec FORSENERIE.

<li name=»LIEURE» num=»3»/>	<li deb=»1» fin=»21»/>
<li name=»LIBÉRALEMENT» num=»2»/>	<li deb=»1» fin=»9»/>
<li name=»CONTUSION» num=»3»/>	<li deb=»5» fin=»6»/>
<li name=»ALIS» num=»0»/>	<li deb=»0» fin=»5»/>
<li name=»INTÉGRAL» num=»7»/>	<li deb=»1» fin=»41»/>
<li name=»LEVAIN» num=»6»/>	<li deb=»1» fin=»10»/>
<li name=”CONTINU” num=”4”/>	<li deb=»1» fin=»20»/>
<li name=“SORBILE“ num=“8“/>	<li deb=»1» fin=»7»/>
<li name=“ANTHORA“ num=“9“/>	<li deb=»1» fin=»23»/>
<li name=”sans levain” num=”32” key=”key”>	<li deb=»1» fin=»34»/>
<li name=”levain” num=”5” key=”key”/>	<li deb=»1» fin=»30»/>
<li name=”INCESSIVEMENT” num=”11”/>	<li deb=»1» fin=»3»/>
<li name=”DESESPERÉ” num=”10”/>	<li deb=»1» fin=»14»/>
<li name=”INALTÉRÉ” num=”12”/>	<li deb=»1» fin=»25»/>
<li name=”INCORPOREL” num=”14”/>	<li deb=»1» fin=»36»/>
<li name=”STINCUS” num=”13”/>	<li deb=»1» fin=»35»/>
<li name=”ESTOILE” num=”15”/>	<li deb=»1» fin=»28»/>
<li name=”TRACE” num=”16”/>	<li deb=»1» fin=»19»/>
<li name=”TÉNESME” num=”18”/>	<li deb=»1» fin=»18»/>
<li name=”ESPERIT” num=”17”/>	<li deb=»1» fin=»16»/>
<li name=”SOURD” num=”20”/>	<li deb=»1» fin=»26»/>
<li name=“SOUFRE“ num=“19“/>	<li deb=»1» fin=»8»/>
<li name=“FORSENERIE“ num=“21“/>	<li deb=»0» fin=»1»/>
<li name=“AVEUGLE“ num=“23“/>	<li deb=»1» fin=»43»/>
<li name=“VIN“ num=“22“/>	<li deb=»1» fin=»22»/>
<li name=“ERRATIQUE“ num=“25“/>	<li deb=»1» fin=»2»/>
<li name=“DESESPOIR“ num=“24“/>	<li deb=»1» fin=»15»/>
<li name=“INSIPIDITÉ“ num=“27“/>	<li deb=»1» fin=»29»/>
<li name=“INSIPIDE“ num=“26“/>	<li deb=»1» fin=»13»/>
<li name=“MODÉRÉ“ num=“28“/>	<li deb=»1» fin=»4»/>
<li name=“SAUVAGE“ num=“29“/>	<li deb=»32» fin=»33»/>
<li name=“SEC“ num=“30“/>	<li deb=»1» fin=»12»/>
<li name=“VAIN“ num=“35“/>	<li deb=»1» fin=»39»/>
<li name=»MOULE» num=»34»/>	<li deb=»1» fin=»37»/>
<li name=»TINTINAILLE» num=»31»/>	<li deb=»1» fin=»42»/>
<li name=”PAIN” num=”33”/>	<li deb=»1» fin=»11»/>
<li name=”APPÉTIT” num=”36”/>	<li deb=»1» fin=»27»/>
<li name=”FAUCON” num=”39”/>	<li deb=»1» fin=»38»/>
<li name=”OPIATE” num=”38”/>	<li deb=»1» fin=»40»/>
<li name=”OLY” num=”37”/>	<li deb=»1» fin=»24»/>
<li name=”sans” num=”1” key=”key”/>	<li deb=»1» fin=»31»/>
<li name=»SAIN -2» num=»40»/>	<li deb=»1» fin=»17»/>
<li name=»ABISME» num=»42»/>	<li deb=»0» fin=»32»/>
<li name=»ÉTOILE» num=»41»/>	
<li name=»TENESME» num=»43»/>	

Ce résultat est traité par un script Javascript afin d’offrir un affichage et une navigation graphique dans n’importe quel navigateur (cf. *infra*).

Résultats, perspectives

Voici un premier exemple de graphe produit par le Dicoscope :

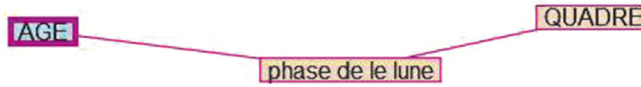

Fig. 1. Graphe ÂGE/QUADRE

Ce graphe révèle un lien entre ÂGE et QUADRE par le biais du segment « phase de la lune ». Ces deux mots connaissent chacun deux acceptions :

ÂGE
I. Âge, portion déterminée de la vie d'un homme
II. Phase de la lune

QUADRE
I. Aspect carré, ou quadrature (90.), jugé défavorable en astrologie
II. Phase de la lune, quartier

Le Dicoscope relie donc la deuxième acception de ÂGE avec la deuxième acception de QUADRE en relevant une intersection sémantique pointée par le lexicographe à travers le segment partagé analysé[8], la première étant un hyperonyme de la seconde. Le graphe détermine une relation de synonymie probable, la réciproque étant qu'il est possible d'exploiter le graphe pour effectuer une désambiguïsation entre les acceptions des lexèmes.

Dans un second temps, la figure suivante montre de manière détaillée les entrées liées à AZIMUT par le biais du segment lexicographique « de la sphère céleste ».

8. Par « analysé », nous évoquons le travail préalable de lemmatisation et de *tokenization* qui a permis la projection du Dicoscope. Nous rappelons que notre expérience est conduite sur le *Dictionnaire du français scientifique médiéval* (programme ANR CréaLScience) et que les segments sont lemmatisés.

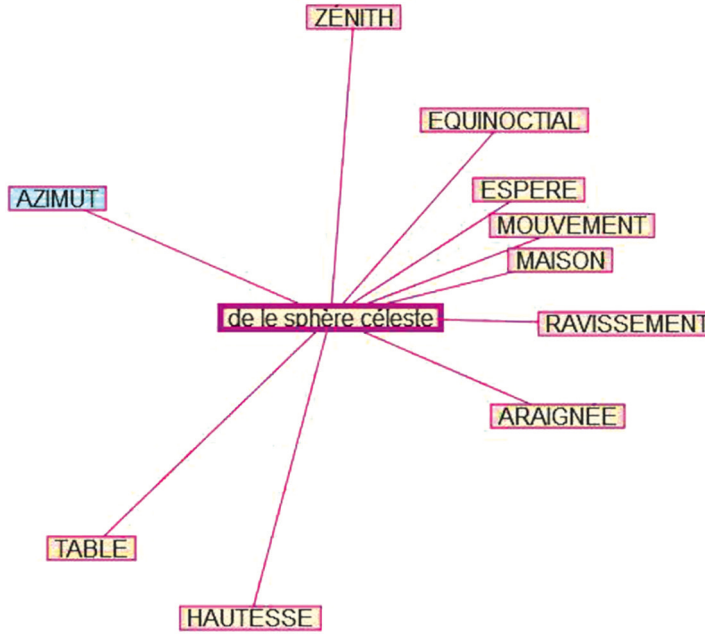

Fig. 2. Graphe «de la sphère céleste»

En l'occurrence, on obtient une liste de lemmes partagés par une série de fiches :

Céleste	(lié à SIGNAL, ACCIDENT, ECLIPSE…)
Cercle de le sphère céleste	(lié à AZIMUT MERIDIANE, ARIMILLE…)
Cercle de le	(lié à AZIMUT, ASCENSION)
Cercle	(lié à AZIMUT, HORIZON, CONE, CLIMAT…)
De le sphère céleste	(lié à AZIMUT ZENITH TABLE, ESPERE…)
De le sphère	(lié à AZIMUT, ARTICUS)
Le sphère	(lié à AZIMUT, COMETE, FIRMAMENT)
Passer par le zénith	(lié à AZIMUT, DIRECTION)
Passer par	(lié à AZIMUT, CENTRE, ARGUMENT…)
Passer	(lié à AZIMUT, SETON, ESPERIT…)
Sphère céleste	(lié à AZIMUT, ETOILE, INTELLIGENCE…)
Sphère	(lié à AZIMUT, POLE, APPROCHEMENT…)

À gauche, nous trouvons les segments partagés, et à droite les entrées qui les partagent, avec les informations concernant leur catégorie grammaticale et leur domaine. En construisant cette liste pour toutes les entrées du dictionnaire, nous obtenons l'ensemble des modules dictionnairiques. Or, si le dictionnaire peut être considéré comme un « trésor de la langue », le Dicoscope nous présente un trésor du dictionnaire, soit un méta-trésor. La perspective qui s'ouvre, ici, est celle d'une topographie du dictionnaire lui-même.

Il existe au moins deux façons de concevoir une topographie du dictionnaire. Tout d'abord, on peut étudier la distribution des segments selon les domaines, en explorant, par exemple, les moules caractéristiques d'un domaine, plutôt que d'un autre. Cela nous permettra de découvrir des regroupements ou des inclusions, fondés non pas sur une catégorisation *a priori*, mais bien sur les moules signifiants des définitions naturelles. Ainsi, nous pourrons évaluer, justement, la distribution des étiquettes de domaines accomplie par les lexicographes. D'ailleurs, dans le cas spécifique du *DFSM*, lorsque l'on dispose des noms des domaines anciens et modernes, on peut aussi envisager d'explorer leur évolution, en fournissant des données lexicographiques pour l'histoire des sciences. Ensuite, il sera intéressant d'explorer les segments partagés qui ne se révèlent attribuables à aucun domaine spécifique, mais qui manifestent un caractère trans-domaine. Ces segments-là permettront de faire émerger le socle du lexique indispensable pour construire tous les autres concepts. En ce qui concerne plus spécifiquement CréaLScience, nous attirons l'attention sur la possibilité d'investiguer les concepts qui sont restés constants à travers le lexique médiéval et ont perduré jusqu'à nos jours, c'est-à-dire le socle des concepts dont on ne peut pas se passer, au Moyen Âge comme aujourd'hui, pour définir tous les autres : des primitifs lexicographiques.

Observons enfin l'exemple suivant :

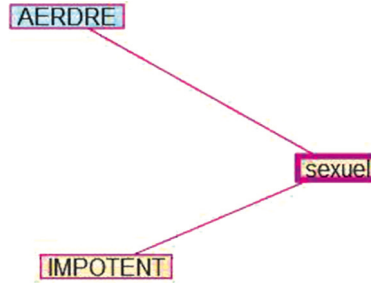

Fig. 3. Graphe « sexuel »

La définition de AERDRE proposée dans le *DFSM* est la suivante : « Être en contact lors de rapports sexuels ». Or, le Dicoscope montre ici une relation d'opposition, notamment avec « impotent », à travers le noyau commun partagé par les deux lexèmes. À la lumière de ce schéma qu'il est nécessaire de typer progressivement, probablement à la main, les liens mis en évidence par le Dicoscope permettent de réaliser un étiquetage des relations lexicales et conceptuelles du dictionnaire. Ainsi, nous aurons un outil efficace de validation du travail des équipes lexicographiques.

Finalement, l'outil mis en place peut aussi avoir tout simplement une fonction d'investigation dictionnairique qui obéisse précisément à la curiosité du lecteur, et qui le guide tout naturellement vers la connaissance de signifiants répertoriés difficilement accessibles. C'est le cas du lien entre le champ sémantique de HABITATION et « *avoir un rapport sexuel* », qu'un non spécialiste aura difficilement pu imaginer :

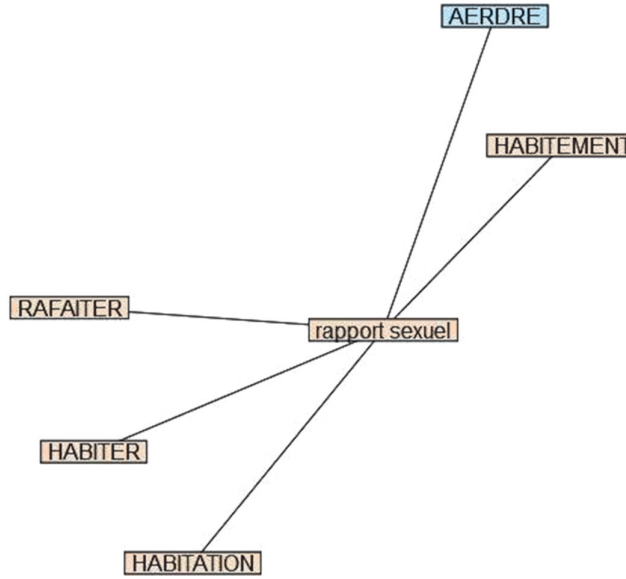

Fig. 4. Graphe « rapport sexuel »

Références bibliographiques

Aitchison, Jean, *Words in the Mind: An Introduction to the Mental Lexicon*, Oxford, Blackwell, 2003.

Austin, John Langshaw, « A plea for excuses » [1956], dans Urmson, James Opie et Warnock, Geoffrey James (dir.), *Philosophical papers*, London, Oxford University Press, 3e éd., 1979.

Authier-Revuz, Jacqueline, « Le guillemet, un signe de "langue écrite" à part entière », dans Defays, Jean-Marc, Rosier, Laurence et Tilkin, Françoise (dir.), *À qui appartient la ponctuation ?*, Louvain-la-Neuve, De Boeck/Duculot, 1998, p. 373-388.

—, *Ces mots qui ne vont pas de soi. Boucles réflexives et non-coïncidences du dire*, Paris, Larousse, 2 t., 1995.

Barbe, Lionel, « Wikipédia, un trouble-fête de l'édition scientifique », *Hermès*, n° 57, Paris, CNRS Éditions, 2010/1, p. 69-74.

Boudet, Jean-Patrice, *Entre science et nigromance. Astrologie, divination et magie dans l'Occident médiéval (XIIe-XVe siècle)*, Paris, Publications de la Sorbonne, 2006.

Caruso, Valeria, « Online Specialised Dictionaries: A Critical Survey », dans Kosem, Iztok et Kosem, Karmen (dir.), *Electronic Lexicography in the 21st Century: New Applications for New Users. Proceedings of eLesc 2011, Bled, Slowenien, 10-12 November 2011*, Ljubljana/Brighton, Trojína (Institute for Applied Slovene Studies)/Lexical Comuting Ltd., 2011.

Ducos, Joëlle, « Le lexique de Jean Corbechon : quelques remarques à propos des livres IV et XI », dans Van den Abeele, Baudoin et Meyer, Heinz (dir.), *Bartholomeus Anglicus, « De proprietatibus rerum », texte latin et réception vernaculaire*, Brepols, Turnhout, 2006, p. 101-115.

Endrizzi, Laure, « Le transfert des savoirs et le cas de Wikipédia », dans Schöpfel, Joachim (dir.), *La Publication scientifique. Analyses et perspectives*, Paris, Hermès/Lavoisier, 2008, p. 171-202.

Issac, Fabrice et Salvador, Xavier-Laurent, « Modèles théoriques inductifs et propositions d'applications aux données textuelles de l'ancien français », dans *JADT, Actes des 10 journées internationales d'analyse statistique des données textuelles*, Milano, LED, 2010.

Kocourek, Rostislav, *La Langue française de la technique et de la science. Vers une linguistique de la langue savante*, Wiesbaden, Oscar Brandstetter, 2e éd. augmentée, refondue et mise à jour avec une nouvelle bibliographie, 1991.

Lesk, Michael, « Automatic Sense Disambiguation. Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone », dans *SIGDOC Conference*, Toronto, Ontario, 1986.

Levelt, Willem J.W., *Speaking: From Intention to Articulation*, London, MIT Press, 1989.

Molinié, Georges, « Stylistique et tradition rhétorique », *Hermès*, n° 15, 1995, p. 119-128.

Pruvost, Jean, « Quelques concepts lexicographiques opératoires à promouvoir au seuil du XXIe siècle », *ELA. Études de linguistique appliquée*, n° 137, 2005/1, p. 7-37.

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**