REVUE DE
LINGUISTIQUE
FRANÇAISE
DIACHRONIQUE

DIACHRONIQUES

7
2017

LES ÉTATS ANCIENS DES LANGUES À L'HEURE DU NUMÉRIQUE

Masià – 979-10-231-2166-7

PUPS

7

DIACHRONIQUES

REVUE DE LINGUISTIQUE FRANÇAISE DIACHRONIQUE

LES ÉTATS ANCIENS DES LANGUES À L'HEURE DU NUMÉRIQUE

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

12 €

http://pups.paris-sorbonne.fr

Maquette : www.stephanemercier.fr

LES ÉTATS ANCIENS DES LANGUES À L'HEURE DU NUMÉRIQUE

Les états anciens des langues à l'heure du numérique

PUPS

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

© Presses de l'université Paris-Sorbonne, 2018
© Sorbonne Université Presses, 2021

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

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

Ramon Masià
Universitat Oberta de Catalunya

À peu près cent ouvrages de mathématique grecque écrits dans l'intervalle temporel du IVe siècle av. J.-C. au VIIe siècle apr. J.-C. sont parvenus jusqu'à nous. Ce sont des textes essentiellement rédigés en langue grecque : on n'y trouve pas de langage formulaire moderne (équations, intégrales, signes de racines, etc.). En fait, il est possible de définir la mathématique grecque comme un genre littéraire ancien, à l'instar de l'épique ou du comique, tant les traits stylistiques des textes mathématiques sont très bien définis et réguliers. En outre, les Anciens classaient ce type de textes par leur style, davantage que par leur contenu. C'est la raison pour laquelle il est très intéressant, et en réalité crucial, de connaître la langue des mathématiques grecques pour prétendre connaître la discipline elle-même. C'est d'ailleurs là une tâche réalisable, parce que le corpus mathématique grec est un corpus que l'on peut traiter, du point de vue numérique, avec des ressources informatiques et humaines limitées.

Il existe des projets pour le traitement numérique des textes grecs (le projet Perseus par ex., qui est le plus ambitieux), mais aucun d'eux ne se donne pour objectif d'analyser la langue de la science grecque. Notre travail représente le premier pas vers le traitement numérique de ce type de corpus, dans le cas des textes mathématiques.

Le Corpus des textes mathématiques grecs (CTMG)

Le Corpus des textes mathématiques grecs (CTMG) contient une centaine d'ouvrages, et l'intervalle temporel qu'il couvre va du IVe siècle av. J.-C. au VIIe siècle apr. J.-C. Ces textes ne recourent pas au langage formulaire moderne (équations, intégrales, signes de racines, etc.), mais présentent deux éléments originaux, que nous ne trouvons dans presque aucun autre ouvrage en grec ancien : des *lettres dénotatives* et des *diagrammes*. Les objets mathématiques sont usuellement codés par un groupe de lettres de l'alphabet grec (par ex. « un carré AB », où le groupe « AB » désigne le carré mentionné). Ces groupes de lettres dénotatives apparaissent dans le texte, mais aussi dans les diagrammes. Toutefois si on néglige ces deux éléments, lettres dénotatives et diagrammes, on peut dire que les mathématiques grecques sont écrites avec la langue usuelle des Grecs, comme tous les autres ouvrages grecs, qu'ils relèvent des genres épique, historique, médical, etc.

Les œuvres grecques sont segmentées selon leur genre littéraire : chacun de ces genres appelle un style spécifique, et chaque ouvrage relève spécifiquement de l'un d'entre eux ; il n'y a pas, en général, de mélanges de style. Les traits stylistiques du CTMG sont également bien définis, et c'est pourquoi on peut affirmer que la mathématique grecque présente un style clairement distingué ; les Anciens reconnaissaient un texte mathématique à son style, davantage qu'à son contenu – d'où la nécessité d'en connaître les traits stylistiques, ce qui ne représente pas de difficulté majeure en raison de la facilité attachée, aujourd'hui, au traitement informatique de ce corpus. En effet :

– le nombre d'ouvrages n'est pas si grand (environ une centaine),
– le nombre d'*occurrences* dans ces textes est « raisonnable » (à peu près 2,5 millions d'occurrences),

– le nombre de *lemmes* dans le corpus est bas (on pense qu'il ne comporte pas plus de 6 000 lemmes),
– la langue mathématique est quasiment dépourvue d'ambiguïtés,
– les structures syntaxiques de base sont réduites et ne présentent que peu de variantes.

Enfin, la langue des mathématiques grecques est plus répétitive et moins polysémique que la langue usuelle et, en conséquence, elle est plus aisée à traiter par des moyens informatiques. Nous présenterons ici les grands axes de nos travaux de numérisation et d'analyse de corpus.

La méthodologie

Notre recherche adopte les procédures de la linguistique de corpus (*Corpus Linguistics*), méthodologie de travail portant sur des *corpus* numérisés et présentant deux caractéristiques principales[1] :

– Les *corpus* numérisés contiennent, essentiellement, deux types d'informations ou de données : d'une part le(s) texte(s) qu'on veut analyser, et d'autre part les métadonnées associées à ces textes.
– L'objectif est d'extraire des informations concernant l'emploi et la structure de la langue représentée dans le corpus.

On utilise, aussi, des outils statistiques permettant de réaliser des analyses stylométriques.

La technique de base de notre méthodologie repose sur la lemmatisation, c'est-à-dire une procédure permettant de déterminer les occurrences, les formes et les lemmes de tous les textes du corpus :

– chaque mot du texte (chaîne de lettres de l'alphabet grec) constitue une occurrence (*token*) ;
– toutes les occurrences identiques du texte constituent une forme (*form*) ;

1. Voir Garside, Leech et McEcery (1997).

– l'ensemble des formes du même terme du lexique constitue un lemme. Le CTMG contient seulement un type de lemme qui n'appartient pas à strictement parler au lexique de la langue grecque : les lettres dénotatives.

Cette procédure de lemmatisation est semi-automatique : la première fois qu'apparaît une forme, il faut l'analyser (ou la valider) à la main. Ensuite, on la trouve automatiquement, parce que le corpus ne présente presque pas d'ambiguïté. Avec cette procédure semi-automatique, on peut garantir que les résultats de l'analyse ne contiendront pas beaucoup d'erreurs, tous les lemmes étant validés à la main, sans pour autant que le temps nécessaire à sa réalisation ne soit trop important, parce que le corpus contient un nombre « raisonnable » de lemmes. À ce stade de la recherche, l'annotation avec métadonnées est très simple, comme nous le verrons.

Problématique du CTMG

Plusieurs éléments problématiques du CTMG doivent être analysés et éclaircis avant de procéder à l'analyse du corpus. Il a fallu prendre des décisions préalables en ce qui les concerne. Il faut dire, d'ailleurs, que cette tâche inaugurale de résolution s'est révélée la plus longue et la plus pénible de notre travail. On peut classer ces éléments en deux catégories : les problèmes d'encodage d'un côté, et de l'autre les problèmes de marquage.

Problèmes d'encodage

L'encodage des textes n'est pas en soi un préalable à l'analyse. Les textes du CTMG ne sont pas tous encodés, et les textes encodés ne le sont pas toujours convenablement ni sous une forme homogène. En général, il est possible de distinguer, du point de vue de la numérisation :

– les éditions imprimées, critiques ou commentées : presque tous les ouvrages de mathématique grecque font l'objet d'une édition imprimée ;
– les éditions du *Thesaurus Linguae Graecae* (TLG) : la base de données numérique du TLG, la plus ancienne base de données

de textes grecs, contient presque tous les textes importants de l'Antiquité, également pour la mathématique grecque[2].

À partir de ces sources, nous avons commencé à créer une base de données d'ouvrages mathématiques « bien encodés ». Le TLG contient à peu près 80 % des ouvrages mathématiques imprimés, mais beaucoup de problèmes de codage y subsistent, parce que ce dernier se montre parfois aléatoire et pas toujours homogène. Il faut, donc, revoir l'encodage de ces textes et l'homogénéiser afin de le rendre utile dans la perspective du traitement stylométrique : il convient de réintégrer des mots, des abréviations, de ré-encoder les symboles, etc.

La conversion d'une édition imprimée en un texte numérisé et utilisable pour le traitement stylométrique est plus difficile encore à opérer : les procédures d'OCR (*Optical Character Recognition*) pour les textes grecs anciens commettent beaucoup d'erreurs, même si elles ont été constamment améliorées (voir Boschetti *et al.*).

Enfin, la correction (dans le cas du TLG) et la numérisation *via* OCR (dans le cas de textes imprimés) sont des tâches qui nécessitent des ressources humaines considérables, parce qu'elles sont en grande partie réalisées à la main ; en réalité, c'est l'opération qui demande le plus de temps.

Problèmes d'annotation

L'annotation de base est la lemmatisation : concrètement, il faut décider le lemme de chaque forme du texte. Mais il faut effectuer un balisage plus complet, et décider où il convient de s'arrêter. Deux types d'annotations sont nécessaires :

– L'annotation de la macro-structure du CTMG. Il faut introduire la division macroscopique des textes grecs, avec les données suivantes : auteur, ouvrage, livre, préface en forme d'épître, introduction avec définitions, propositions (avec leurs différentes parties : énoncé, exposition,

2. La mathématique grecque inclut des disciplines comme l'astronomie, la musique, la mécanique et d'autres, considérées comme relevant d'elle par les Grecs de l'Antiquité.

détermination, construction, démonstration, conclusion) et les démonstrations alternatives.

– L'annotation des micro-structures, concrètement la structure syntaxique, la structure mathématique et la structure logique.

Ces types de marquage sont encore très réduits, mais nous sommes en train d'accélérer le processus grâce au recours à des outils de marquage automatique (dont le *Natural Language Toolkit* de Python).

Résultats

Nous l'avons signalé, deux processus doivent être réalisés successivement : l'encodage/lemmatisation et l'analyse des textes codifiés.

L'encodage / lemmatisation

Nous avons encodé et lemmatisé intégralement les textes des auteurs suivants : Archimède, Apollonius, Diophante, Domninus, Pappus, Serenus et Théodose. En outre, nous avons encodé les textes qui suivent : les *Éléments* et les *Données* d'Euclide, les *Metrica* d'Héron d'Alexandrie, et les *Prolégomènes à l'Almageste*.

L'état actuel des données de base du corpus est le suivant :

– 733 003 occurrences sur un total de 2,5 millions dans le CTMG,
– 17 618 formes,
– 3 181 lemmes.

Le très petit nombre de lemmes est très remarquable, dans une partie aussi importante du corpus. En outre, il est vraisemblable que le total des lemmes dans le CTMG soit inférieur à 6 000.

Dans le tableau suivant se trouvent les données relatives aux 10 lemmes les plus fréquents. Il est remarquable de constater que près de 55 % des occurrences des textes lemmatisés appartiennent à un lemme de cette liste, c'est-à-dire que plus d'une occurrence sur deux d'un texte mathématique grec s'y trouve. Aucun autre texte en grec ancien ne présente cette concentration d'occurrences en un si petit groupe de lemmes.

Dans cette liste, on trouve l'article défini, avec plus de 20 % des occurrences, les lettres dénotatives, trois conjonctions, trois prépositions, les adjectifs numéraux, et, finalement, le verbe *être*, le seul représentant des catégories qu'on peut considérer comme sémantiquement pleines[3].

Tableau. 1. Les dix lemmes les plus fréquents

Lemme	733 003	%	% cumulé
o/article	158 904	21,68	21,68
lettre dénotative	104 893	14,31	35,99
και/et	29 625	4,04	40,03
ειμι/être	29 381	4,01	44,04
προς/préposition	17 592	2,40	46,44
δε/conjonction	13 959	1,90	48,34
nombre	13 534	1,85	50,19
αρα/par conséquent	11 916	1,63	51,81
αρο/préposition	11 003	1,50	53,32
υπο/préposition	10 626	1,45	54,77

Analyse des textes

Afin de parvenir à une analyse plus approfondie, nous avons procédé à une annotation plus précise pour l'œuvre intégrale de quelques auteurs et quelques livres isolés. Les auteurs traités intégralement sont Archimède et Apollonius, et les ouvrages isolés sont les *Éléments* et les *Données* d'Euclide ainsi que les *Metrica* d'Héron d'Alexandrie. Les données de base de ce corpus sont les suivantes :

– il contient 369 485 occurrences,
– le nombre de formes s'élève à 8 208,
– le nombre de lemmes est de 1 318.

Le balisage complémentaire inclut la catégorie grammaticale (presque 50 % des lemmes de cette partie du corpus ont déjà été marqués avec leur catégorie grammaticale) et les parties du texte

3. Il faut dire aussi que la fréquence de ce verbe est très élevée : elle représente plus de 4 % des occurrences (par ex. la fréquence de ce verbe dans Platon, l'auteur ancien qui utilise le plus le verbe être, est à peu près de 3 %). Tous les accents diacritiques en grec ont été éliminés dans notre texte et dans les tables pour faciliter l'édition.

(jusqu'à Proposition). Grâce à ce marquage, nous pouvons procéder à plusieurs types d'analyses : description statistique du lexique du corpus, analyses comparatives entre parties du corpus, analyse structurelle (syntaxique, logique et des unités mathématiques).

Nous montrerons maintenant quelques résultats intéressants parmi ceux obtenus grâce à ce traitement.

La langue d'Archimède vs. *la langue commune dans le corpus marqué*

Treize livres d'Archimède ont survécu et nous les avons tous numérisés, lemmatisés et annotés. Nous avons procédé à une comparaison de la langue d'Archimède avec celle de tous les textes du corpus encodé, lemmatisé et annoté (soit 36 livres). Pour faire une première comparaison, nous choisissons les lemmes communs de la langue d'Archimède et ceux de la langue du corpus déjà encodé.

Il y a 27 lemmes communs dans le lexique d'Archimède. Cette catégorie représente 3,41 % de l'ensemble des lemmes qu'il contient (761 lemmes au total), mais seulement 1,79 % des lemmes de la langue du corpus encodé sont communs à tout le corpus (1 226 lemmes). Ces lemmes représentent respectivement 62,71 % des occurrences chez Archimède (97 876 occurrences) et 64,20 % des occurrences dans le corpus (234 897 occurrences). Il faut noter que la proportion d'occurrences des lemmes communs aux œuvres d'Archimède est presque égale au nombre d'occurrences des lemmes communs aux ouvrages du corpus.

Si on analyse le type et les fonctions du lexique commun dans chaque cas, on peut noter que :

– le pourcentage d'éléments anaphoriques/déterminants (articles, pronoms, etc.) est très semblable dans la langue commune d'Archimède et dans la langue commune du corpus codifié (à peu près 40 %) ;
– les éléments de formation de *formulae*[4] mathématiques sont à la hauteur de 5,23 % dans la langue d'Archimède et de 5,57 %

4. Au sens du terme *formula* introduit par Reviel Netz (dans Netz, 2003, chapitre IV).

dans la langue commune du corpus encodé, c'est-à-dire une différence de 6,11 % ;

– il y a 9,03 % d'éléments logiques dans la langue d'Archimède et 10,31 % dans la langue commune, soit une différence de 12,45 % ;
– il y a dans la langue d'Archimède davantage d'éléments sémantiques pleins que dans la langue commune. On compte dans la langue commune d'Archimède quatre verbes (*être, avoir, conduire, couper*), deux tours relationnels (*égal à* et *plus grand que*), un substantif (*droite*), un adjectif (*ce qui reste*), alors que, dans la langue commune du corpus, il y a deux verbes (*être et avoir*) et deux tours relationnels (*être égal à* et *être plus grand que*).

Avec ces données, il semble évident que la langue d'Archimède a une articulation logique plus faible (moins d'éléments logiques) et que son style est moins formulaire (moins d'éléments de formation de *formulae* mathématiques et davantage d'éléments sémantiques pleins). Si on a lu Archimède et Euclide, ces conclusions semblent plausibles, mais il serait impossible de les obtenir au moyen d'un relevé manuel.

Quelques analyses comparatives plus complexes

Des outils existent qui autorisent des analyses comparatives plus approfondies basées sur la recherche des mots clefs (*keywords*), le calcul de la *keyness*, et les mesures de proximité lexicale.

La *keyness* d'un lemme, nommé *keyword*, dans une partie du corpus mesure l'importance de ce lemme dans cette partie du corpus par rapport à son importance dans une autre partie du corpus. La table qui suit présente par exemple la *positive keyness* des lemmes présents chez Archimède par rapport à ceux contenus dans les *Éléments* d'Euclide[5].

5. Nous n'avons pas intégré de diacritiques dans les mots employés pour énoncer le lemme, parce qu'il est plus facile de les traiter informatiquement ainsi et, aussi, parce que de cette manière il est très aisé de les distinguer des différentes formes et occurrences.

Tableau 2. Positive keyness

Freq	Keyness	Lemme
1063	1293.356	κωνος
580	951.424	επιφανεια
956	912.546	τμημα
1566	891.203	εχω
364	687.368	βαρος
513	684.329	αξων
468	615.457	τομη
481	570.49	σχημα

Le terme κωνος, « cône », est le lemme le plus caractéristique chez Archimède par rapport aux *Éléments* d'Euclide. Dans la liste des huit lemmes les plus caractéristiques d'Archimède par rapport aux *Éléments* d'Euclide, sept sont des substantifs : επιφανεια (« surface »), τμημα (« segment »), βαρος (« gravité »), αξων (« axe »), τομη (« section ») et σχημα (« figure »). Y figure aussi un verbe, εχω (« avoir »), très caractéristique d'Archimède. Dans la table suivante nous identifions au contraire les lemmes plus caractéristiques des *Éléments* d'Euclide, par rapport au corpus archimédien :

Tableau 3. Negative keyness

Freq	Keyness	Lemme
11906	1619.83	lettre dénotative
644	1267.767	αρα
11	683.935	μετρεω
6	658.03	συμμετρος
2	486.951	ασυμμετρος
118	433.117	αριθμος
161	408.984	γωνια
3352	379.373	ειμι

Euclide utilise beaucoup plus de lettres dénotatives qu'Archimède, et c'est le lemme le plus caractéristique d'Euclide par rapport à Archimède. En second rang nous trouvons la conjonction conclusive αρα, « par conséquent » (ce qui signifie que les propositions d'Archimède sont moins conclusives que les propositions d'Euclide, parce que cette conjonction est utilisée très souvent dans la conclusion d'une proposition mathématique). Il y a, aussi, deux verbes (μετρεω, « mesurer » et ειμι, « être »), deux adjectifs (συμμετρος, « commensurable » et ασυμμετρος, « incommensurable ») et deux substantifs (αριθμος, « nombre » et γωνια, « angle »), qui sont très caractéristiques d'Euclide par rapport à Archimède.

Avec la recherche de *keywords* et de leur *keyness*, on peut définir les grands axes thématiques (verbes et substantifs préférés), préciser l'usage de diverses particules logiques (très importantes dans les textes mathématiques), etc.

Mais on peut calculer d'autres mesures de proximité entre les textes en recourant aux outils stylométriques, car ils nous donnent une idée de la proximité stylistique entre parties du corpus. Différents types d'analyse peuvent être réalisés : *cluster analysis*, *multidimensional scaling* ou *principal component analysis*. Le but principal de ces techniques est de grouper les ouvrages, ou des parties de ces ouvrages, selon leur proximité lexicale, à partir de certains paramètres. Par exemple, le graphique suivant montre une *cluster analysis* des données lemmatiques tirées de quelques ouvrages de quatre auteurs grecs : Euclide (en bleu), Archimède (en vert), Apollonius (en rouge) et Héron d'Alexandrie (en noir).

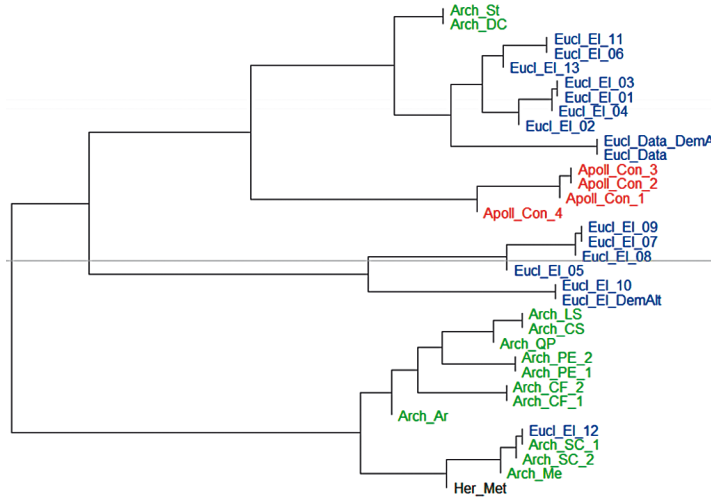

Fig. 1. *Cluster analysis*

Ce graphique montre un regroupement « objectif » de ces textes à partir des données lexicales. En cette phase de notre recherche, nous sommes en train d'évaluer ces outils : il s'agit de savoir si les résultats qu'ils nous offrent coïncident avec ce que nous savons de la mathématique grecque. Le graphique montre bien que les auteurs et les ouvrages constituent des regroupements de forme assez cohérente : tous les textes d'Apollonius sont contigus, presque tous ceux d'Archimède sont dans un même groupe, et aussi presque tous ceux d'Euclide. D'ailleurs, les sous-groupes de textes par auteur qui apparaissent dans le graphique sont aussi très cohérents. Ces faits nous montrent que, à première vue, la classification stylométrique nous offre une information pertinente.

On peut aller un peu plus loin : on peut chercher pourquoi des éléments apparaissent « déplacés » dans le graphique. Par exemple, Eucl_El_12 (Livre XII des *Éléments* d'Euclide) jouxte Arch_SC1, Arch_SC2 (Livres I et II de *Sur la sphère et le cylindre* d'Archimède). On peut expliquer ce fait du point de vue du contenu : ces livres se rapportent à des thématiques semblables et différentes de celles développées dans les autres livres des

Éléments ; en fait, certains chercheurs vont jusqu'à dire que ces œuvres d'Archimède sont la continuation naturelle du Livre XII des *Éléments*. De la même manière, Arch_St et Arch_DC (*Stomachion* et *Dimensio Circuli* d'Archimède) sont très excentriques par rapport aux autres textes d'Archimède, et effectivement, pour des raisons différentes, on les considère comme des ouvrages particuliers, et très différents du corpus central d'Archimède.

On constate, donc, que la classification stylistique résultant de ce type de graphique se reflète plus ou moins dans les caractéristiques du contenu et corrobore certaines intuitions des chercheurs quant aux ouvrages du corpus. En fait, ces résultats étaient prédictibles, parce que, nous l'avons dit, la mathématique grecque est un genre littéraire ancien, et par conséquent le contenu et le style doivent être très liés. On peut donc affirmer qu'en principe ces analyses sont susceptibles de donner des résultats intéressants et suggestifs, qui nous permettront d'étendre les recherches portant sur la mathématique grecque.

Pour finir, signalons que ces analyses ont été réalisées au moyen des logiciels suivants :

– pour la lemmatisation, les logiciels AntConc et AntWordProfiler,
– pour la recherche des *keywords* et de leur *keyness*, le logiciel AntConc également,
– pour les analyses stylistiques, avec *Cluster Analysis*, *Multidimensional Scaling* et *Principal Component Analysis* ; on a utilisé le programme statistique R et RStudio, et plus concrètement le *script* StyloR. On a recouru aussi au logiciel Gephi pour la visualisation et traitement de graphes.

L'écrit mathématique en grec ancien est un genre littéraire de l'Antiquité : on peut donc le définir à partir de ses caractéristiques stylistiques. Mais on peut supposer aussi que les sous-genres de ce genre littéraire doivent présenter des caractéristiques stylistiques communes, et que par ailleurs chaque auteur peut avoir ses particularités stylistiques. Pour ces raisons, il est très intéressant de traiter numériquement le Corpus des textes mathématiques grecs (CTMG), afin d'obtenir les données

stylistiques des ouvrages et les comparer, ce qui nous permettra de mieux connaître la mathématique grecque.

Par ailleurs, le CTMG est un corpus que l'on peut traiter du point de vue numérique : le nombre d'ouvrages est restreint, le nombre de lemmes et d'occurrences est également de faible importance, le corpus ne comporte presque pas d'ambigüités et les structures syntaxiques de base y sont réduites.

Les résultats de nos premières analyses nous confirment que le CTMG est un corpus au lexique très réduit : nous recensons désormais 733 003 occurrences, et seulement 3 181 lemmes. Le lexique est lui aussi très concentré : 7 lemmes concentrent 50 % de toutes les occurrences. Ces caractéristiques ne sont communes avec aucun autre genre littéraire de l'Antiquité.

Les analyses comparatives des textes, utilisant différents outils (*keywords/keyness*, *cluster analysis*, *multidimensional scaling* et *principal component analysis*) nous confirment que des caractéristiques spécifiques peuvent être rattachées à chaque auteur/texte, et que les groupements d'ouvrages obtenus en utilisant ces outils sont cohérents avec ce que nous savons déjà des ouvrages et des auteurs. En outre, on constate bien une connexion entre les caractéristiques du contenu des ouvrages et leur style. Cette confirmation nous permet de conclure que l'utilisation de ces analyses nous permettra de mettre au jour des relations significatives entre textes, ou entre auteurs, inconnues jusqu'à présent.

Références bibliographiques

Textes

Acerbi, Fabio, « I codici stilistici della matematica greca: dimostrazioni, procedure, algoritmi », *Quaderni Urbinati di Cultura Classica*, n° 101, 2012, p. 167-214.

Aujac, Germaine, « Le langage formulaire dans la géométrie grecque », *Revue d'histoire des sciences*, n° 3, 1984/2, p. 97-109.

Bakker, Stéphanie J., *The Noun Phrase in Ancient Greek*, Brill, Leiden, 2009.

Boschetti, Federico *et al.*, « Improving OCR Accuracy For Classical Critical Editions », dans *Research and Advanced Technology for Digital Libraries. 13th European Conference*, ECDL 2009, Berlin/Heidelberg, Springer-Verlag GmbH, 2009, p. 156-167.

Crane, Gregory, « Generating and Parsing Classical Greek », *Literary and Linguistic Computing*, n° 6, 1991/4, p. 243-245.

Deodati, Sara et Kindt, Bastien, « La lemmatisation automatisée des sources en grec ancien : présentation de ressources linguistiques et d'outils de traitement », dans Corino, Elisa, Marello, Caria et Onesti, Cristina (dir.), *Atti del XII Congresso Internazionale di Lessicografia*, Alessandria, Edizioni dell'Orso, 2006, t. II, p. 1137-1143.

Eder, Maciej, « Style-markers in Authorship Attribution. A Cross-Language Study of Authorial Fingerprint », *Studies in Polish Linguistics*, n° 6, 2011, p. 99-114.

—, « Mind Your Corpus: Systematic Errors », *Authorship Attribution. Literary and Linguistic Computing*, n° 28, 2013/4, p. 603-614.

—, « Stylometry, Network Analysis and Latin Literature », *Digital Humanities 2014: Book of Abstracts,* EPFL-UNIL, 2014, p. 457-458.

Garside, Roger, Leech, Geoffrey et McEnery, Tony, *Corpus Annotation: Linguistic Information from Computer Text Corpora*, London/New York, Longman, 1997.

Gelbukh, Alexander et Sidorov, Grigori, *Procesamiento automático del español con enfoque en recursos léxicos grandes*, México, Instituto Politécnico Nacional, 2006.

Manning, Christopher et Schütze, Heinrich, *Foundations of Statistical Natural Language Processing*, Cambridge, MIT Press, 1999.

Mugler, Charles, *Dictionnaire historique de la terminologie géométrique des grecs*, Paris, Klincksieck, 1959.

Netz, Reviel, « Proclus' Division of the Mathematical Proposition into Parts: How and why was it formulated? », *Classical Quaterly*, n° 49, 1999/1, p. 282-303.

—, *The Shaping of Deduction in Greek Mathematics*, Cambridge, Cambridge University Press, 2003.

Vitrac, Bernard, *La Transmission des textes mathématiques grecs*. En ligne : https://www.academia.edu/16162595/La_transmission_des_textes_math%C3%A9matiques_grecs.

Pages web

AntConc : http://www.laurenceanthony.net/software/antconc/

AntWordProfiler : http://www.laurenceanthony.net/software/antwordprofiler/

R : https://www.r-project.org/

RStudio : https://www.rstudio.com/

StyloR : https://sites.google.com/site/computationalstylistics/scripts

TLG Perseus Digital library : http://www.tlg.uci.edu/

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.

COMITÉ SCIENTIFIQUE

Hava BAT-ZEEV SHYLDKROT (Université de Tel Aviv)
Françoise BERLAN (Université Paris-Sorbonne)
Mireille HUCHON (Université Paris-Sorbonne)
Peter KOCH (Universität Tübingen)†
Anthony LODGE (Saint Andrews University)
Christiane MARCHELLO-NIZIA (École normale supérieure-LSH, Lyon)
Robert MARTIN (Université Paris-Sorbonne/Académie des inscriptions et belles-lettres)
Georges MOLINIÉ (Université Paris-Sorbonne)†
Claude MULLER (Université Bordeaux Montaigne)
Laurence ROSIER (Université Libre de Bruxelles)
Gilles ROUSSINEAU (Université Paris-Sorbonne)
Claude THOMASSET (Université Paris-Sorbonne)

COMITÉ DE RÉDACTION

Claire BADIOU-MONFERRAN (Université de Lorraine)
Michel BANNIARD (Université Toulouse 2-Le Mirail)
Annie BERTIN (Université Paris Ouest Nanterre La Défense)
Claude BURIDANT (Université Strasbourg 2)
Maria COLOMBO-TIMELLI (Université Paris-Sorbonne)
Bernard COMBETTES (Université de Lorraine)
Frédéric DUVAL (École nationale des chartes)
Pierre-Yves DUFEU (Université Aix-Marseille 3)
Amalia RODRIGUEZ-SOMOLINOS (Universidad Complutense de Madrid)
Philippe SELOSSE (Université Lyon 2)
Christine SILVI (Université Paris-Sorbonne)
André THIBAULT (Université Paris-Sorbonne)

COMITÉ ÉDITORIAL

Olivier SOUTET (Université Paris-Sorbonne), Directeur de la publication
Joëlle DUCOS (Université Paris-Sorbonne-EPHE), Trésorière
Stéphane MARCOTTE (Université Paris-Sorbonne), Secrétaire de rédaction
Thierry PONCHON (Université de Reims Champagne-Ardenne), Secrétaire de rédaction
Antoine GAUTIER (Université Paris-Sorbonne), Diffusion de la revue

Table des matières