REVUE DE
LINGUISTIQUE
FRANÇAISE
DIACHRONIQUE

2017

DIACHRONIQUES

LES ÉTATS ANCIENS DES LANGUES À L'HEURE DU NUMÉRIQUE

Guadagnini – 979-10-231-2161-2

PUPS

7

DIACHRONIQUES

REVUE DE LINGUISTIQUE FRANÇAISE DIACHRONIQUE

LES ÉTATS ANCIENS DES LANGUES À L'HEURE DU NUMÉRIQUE

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

12 €

http://pups.paris-sorbonne.fr

Maquette : www.stephanemercier.fr

LES ÉTATS ANCIENS DES LANGUES À L'HEURE DU NUMÉRIQUE

Les états anciens des langues à l'heure du numérique

PUPS

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

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

PDF complet – 979-10-231-2155-1

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

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

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

Tél. (33) 01 53 10 57 60

sup@sorbonne-universite.fr

sup.sorbonne-universite.fr

La lexicographie de l'italien médiéval et les corpus de l'OVI : un bilan provisoire et quelques nouvelles perspectives

Elisa Guadagnini*
Istituto Opera del vocabolario italiano
CNR, Firenze / KU Leuven

> Avec les techniques modernes de traitement mécanique de textes sont mis à notre disposition de grandes masses de matériaux bruts, et ceci en peu de temps et à peu de frais. Il revient à celui qui crée ces matériaux de les traiter selon les règles de l'art, afin de collaborer vraiment à l'analyse des sources historiques écrites et à l'augmentation du savoir.
>
> *Déclaration de Heidelberg*, 2001[1]

> *Compilare il presente Vocabolario parve la più alta, e vera maniera, fra tutte l'altre, di benificare questo idioma*.
>
> Préface « A' lettori », *Vocabolario della Crusca*, 1612

* Cette contribution transcrit de manière assez fidèle ma communication orale. S'agissant d'une présentation générale, et somme toute assez sommaire, du passé et des activités présentes de l'OVI, je n'ai ajouté ici que très peu de notes, jugeant inopportun d'alourdir le texte avec des références bibliographiques ponctuelles qui seraient pourtant nécessaires pour accompagner chacun des arguments que je ne fais qu'aborder de manière superficielle : j'ai cru pouvoir conserver à l'écrit le caractère vulgarisateur que présentait mon oral. Je tiens à souligner aussi le fait que je présente ma vision personnelle de l'OVI : le tableau qui en est brossé est donc certainement partiel et partial, et borné par les limites mêmes de mon expérience de travail. À mon expérience propre, par ailleurs, se rattachent les projets *DiVo* et *ReMediA*, que je cite ici parmi d'autres existants, et qui eussent assurément tout autant mérité d'être mentionnés.

1. *L'héritage culturel européen et la lexicologie du XXIe siècle : l'avenir de la lexicographie historique. L'exemple du* Dictionnaire étymologique de l'ancien français, 28-30 juin 2001, Internationales Wissenschaftsforum der Universität Heidelberg.

L'Opera del vocabolario italiano (OVI) est un institut de recherche dédié à la rédaction du vocabulaire historique de l'italien ancien, le *Tesoro della lingua italiana delle origini* (*TLIO*).

L'OVI, qui fait partie du CNR (Consiglio nazionale delle ricerche), est constitué en institut depuis 2001, après avoir fonctionné de 1985 jusqu'à cette date sous le statut de centre d'études. Son activité cependant est plus ancienne[2] : s'il a toujours été financé par le CNR, pendant ses vingt premières années l'activité et l'existence même de l'OVI ont été liées à l'Accademia della Crusca[3], dont l'institut partage aujourd'hui encore le siège (la villa médicéenne de Castello, à Florence). C'est en effet à l'Accademia della Crusca que l'on situe, en 1964, le début officiel des travaux pour la réalisation d'un vocabulaire historique décrivant, selon le projet de l'époque, le lexique italien des origines à nos jours[4].

Dès les années 1960, l'idée fut conçue de se consacrer d'abord à la rédaction d'un *Tesoro delle origini* : au fil des années, la décision fut prise de ne s'intéresser qu'aux origines, et de repousser les textes des époques moderne et contemporaine à une phase de travail ultérieure. Aujourd'hui encore, l'équipe de l'OVI travaille à cette première tranche historique du vocabulaire, tout en gardant pour le futur le projet de rédiger les articles relatifs aux séquences chronologiques suivantes.

De longues discussions ont amené les équipes à considérer comme « période des origines » la phase qui va des premiers

2. Pour une histoire de l'institut de sa fondation à la direction Beltrami (1992), voir Vaccaro (2013). Je renvoie à ce travail fondamental pour toutes les informations concernant l'histoire de l'institut et de ses projets, contenues dans les premières pages de l'ouvrage.
3. L'Accademia della Crusca est l'institution lexicographique italienne par excellence depuis 1612, année de la publication de la première édition du *Vocabolario della Crusca* : quatre rééditions du *Vocabolario* se sont succédées depuis, et la cinquième est restée incomplète, car sa rédaction a été interrompue en 1923, après plus d'un siècle de travail.
4. L'*OVI* fut dirigé depuis sa fondation en 1965 et jusqu'en 1972 par Aldo Duro, puis par Giovanni Nencioni (jusqu'en 1974), et ensuite par d'Arco Silvio Avalle. Quand l'OVI est devenu un centre d'études du CNR, en 1985, la direction a été assurée par Carlo Alberto Mastrelli, qui est demeuré en fonction jusqu'en 1992. Pietro G. Beltrami fut le directeur de l'OVI de 1992 à 2013 ; Paolo Squillacioti a ensuite assuré les fonctions de directeur par intérim jusqu'à l'arrivée, le 1er octobre 2014, de l'actuel directeur, Lino Leonardi.

documents écrits conservés (le plus ancien d'entre eux est l'*Indovinello veronese*, que l'on date du IXe siècle) à la fin du XIVe siècle : il s'agit là d'une périodisation classique pour l'histoire de la littérature et de la langue italienne, qui place au XVe siècle – avec l'Humanisme – le début de l'époque moderne.

Dès le début du projet, la décision fut prise de réunir dans la partie ancienne du vocabulaire l'ensemble des variétés italo-romanes dont sont restés des vestiges écrits, et non la seule variété toscane, ou plus spécifiquement florentine ancienne, qui est à la base de l'italien contemporain[5]. Il s'agissait dans les années 1960 d'un choix lexicographique révolutionnaire, mais à vrai dire le tableau diatopique restitué par les dictionnaires allait s'élargir sous peu – et indépendamment des travaux de l'OVI : bien avant la publication du premier article du *TLIO*, il faut rappeler que le *Grande dizionario dell'italiano*, fondé par Salvatore Battaglia et complété sous la direction de Giorgio Barberi Squarotti, cite parmi les textes médiévaux des œuvres provenant de l'Italie du Nord et de l'Italie médiane, offrant un aperçu géographique plus vaste que celui retenu par les grands dictionnaires italiens qui l'avaient précédé, c'est-à-dire les cinq éditions du *Vocabolario della Crusca* et le *Dizionario della lingua italiana* de Niccolò Tommaseo et Bernardo Bellini[6]. Mais le premier essai d'une lexicographie de l'ancien italien qui ne fût pas le rassemblement des glossaires des éditions critiques – une option de description du lexique des « origines », qui, d'ailleurs, fut prise en compte, à un moment donné, à l'*OVI* aussi[7] – fut l'expérience du *Glossario degli antichi volgari italiani* (*GAVI*), que Giorgio Colussi a rédigé à partir des années 1980 et jusqu'à sa mort[8] : pour la première fois, l'on tentait une description lexicographique qui utilisait

5. On peut lire une excellente réflexion sur le concept d' « italien ancien » selon les critères chronologiques et diatopiques dans Tomasin (2013).
6. Le *Vocabolario della Crusca*, tout comme le Tommaseo-Bellini, cite en majorité absolue des textes toscans (à l'exception près des poèmes de Iacopone de Todi).
7. Le projet d'un *Glossario dei glossari* (« Glossaire des glossaires ») fut entrepris dans les années 1970 et poursuivi pendant quelques années, avant d'être abandonné en faveur d'une autre idée de dictionnaire : *voir* Vaccaro (2013), p. 363 et *passim*.
8. Le premier volume du *GAVI* date de 1983, le dernier (XX/2) de 2006 : les volumes publiés couvrent les lettres A, B, C, D, S, U, V et Z.

directement les textes médiévaux (en principe jusqu'à 1321, année de la mort de Dante, mais sont cités aussi des textes ultérieurs offrant même quelques ouvertures sur l'époque moderne), visant à restituer pour le vocabulaire des « origines » un tableau lexicologique, et non purement glossographique, en s'appuyant sur l'entière documentation disponible.

Le *Tesoro della lingua italiana delle origini* (*TLIO*), le vocabulaire que l'OVI est en train de rédiger, couvre donc l'ensemble des variétés italo-romanes dans lesquelles sont conservés des textes écrits avant l'an 1400.

Dès les années 1960, le choix fut fait de ne pas s'appuyer sur la lexicographie précédente mais de rédiger un vocabulaire de première main utilisant comme source directe les textes, en entendant par là les éditions modernes existantes, et non les manuscrits ni les *editiones principes*. La décision ayant été prise d'accomplir un dépouillement exhaustif des sources médiévales publiées, un premier problème apparut, avec le recensement des textes et des éditions. À la suite notamment d'un voyage à Nancy, où s'élaborait un autre « Trésor », celui « de la langue française », le directeur du projet, Aldo Duro, créa l'*Ufficio filologico* (le « Bureau philologique »), auquel fut confiée la tâche de mener à bien l'élaboration de la table des textes à citer ainsi que le repérage et l'évaluation des éditions existantes. Quand, le premier janvier 1965, Domenico De Robertis devint le directeur du bureau philologique nouvellement né, il entreprit avec ses collaborateurs et ses élèves un immense travail de recherche et d'étude, mais aussi de contrôle et de révision des textes : l'équipe du bureau philologique ne se borna pas, en effet, à recenser les éditions et à choisir les meilleures d'entre elles, mais travailla beaucoup à l'amélioration des textes critiques. Pour ne donner qu'un exemple des activités du bureau, pour tous les documents à tradition unique l'on procéda à la collation intégrale de l'édition avec le manuscrit ; il en fut de même pour toutes les éditions que l'on déclarait fondées sur un manuscrit déterminé.

Quand d'Arco Silvio Avalle, en 1974, prit la direction de l'*OVI*, il vit dans l'activité du bureau philologique un risque

réel de paralysie du vocabulaire : dix ans plus tard, en 1983, le CNR stigmatisa – avec les mots de Scevola Mariotti – le « philologisme » de l'*Opera del vocabolario*, qui paraissait nocif pour la réalisation du vocabulaire[9]. Ce contraste entre la position du CNR, qui voulait une assurance quant aux délais et exigeait que les progrès du projet soient visibles, et celle de l'Accademia della Crusca, qui défendait sa méthode philologique, perdura plusieurs années et créa de multiples tensions. Pendant les trois premières décennies de son activité, l'équipe de l'OVI ne rédigea pas un seul article du vocabulaire, mais réalisa un énorme travail de préparation des textes qui allaient constituer la base de données pour la rédaction : la documentation relative à ce travail est aujourd'hui librement consultable en ligne, grâce à un projet mené à terme par Pär Larson et Zeno Verlato[10].

Entre-temps, par ailleurs, l'informatique a vite évolué. Dès 1965, à l'OVI, l'équipe oeuvra à la constitution d'un archivage des sources dépouillées de façon électronique et lemmatisées : cet archivage devait être – dans la vision portée par Aldo Duro – un outil autonome, disponible à la consultation pour l'ensemble des chercheurs. À la suite des échanges associant Duro et le père Roberto Busa, qui, avec le support technique d'IBM, était en train de publier à Gallarate l'*Index Thomisticus*, pendant plusieurs années l'OVI confia l'informatisation de ses sources à IBM : à l'époque le support utilisé était la carte perforée. Ensuite, avec l'essor de logiciels capables de gérer des textes longs et non seulement une courte chaîne de caractères, le projet d'une archive de contextes évolua naturellement vers le projet d'une archive digitale de textes entiers : en 1988, l'OVI, depuis trois ans centre d'études du CNR, adopta un logiciel – écrit par Eugenio Picchi, de l'Istituto di linguistica computazionale (CNR, Pise) – qui répondit ensuite à l'appellation « DBT » et se trouva assez répandu en

9. Se référer à Vaccaro (2013), p. 371.
10. Le projet *LIVS – Lingua italiana e Vocabolario storico : Metodi antichi e moderni*, financé par la région Toscane, a existé du 9 février 2011 au 8 septembre 2013. Les archives des matériaux inventoriés sont disponibles en ligne : http://tlio.ovi.cnr.it/LIVS/livsdoc/.

Italie dans les années 1990[11]. Les données informatisées pendant les années précédentes furent récupérées et converties dans le format DBT : le corpus ainsi constitué par récupération de matériaux plus anciens et par ajout de nouvelles acquisitions atteignit une dimension convenable pour la rédaction en 1995[12]. Ce premier corpus était partiellement lemmatisé, car DBT permettait une lemmatisation manuelle, réalisée texte par texte. En 1998, une version simplifiée du corpus (sans la lemmatisation) fut publiée sur le réseau grâce au consortium Italnet[13]. La même année devint opérationnel le logiciel GATTO, dédié, conçu et écrit spécifiquement pour la base de données de l'OVI par Domenico Iorio-Fili, un informaticien travaillant à l'époque à l'institut. Parmi les avantages que présentait l'adoption de ce logiciel, il y avait le fait que l'on pouvait dès lors procéder à une lemmatisation qui s'effectuait sur le corpus entier[14], ce qui permettait un contrôle optimal des matériaux qui jusqu'à cette date ne pouvaient être analysés que contexte par contexte (à l'époque des cartes perforées) ou lemmatisés forme par forme à l'intérieur d'un seul texte (DBT). En 2005 fut publié en ligne le corpus lemmatisé, qui offre aux usagers des possibilités de recherche très complexes grâce au logiciel GattoWeb, écrit lui aussi par Domenico Iorio-Fili.

Aujourd'hui, l'OVI teste Gatto4, un nouveau logiciel écrit par Domenico Iorio-Fili juste avant sa retraite, survenue en août 2014.

---

11. Il s'agit notamment du logiciel de la *LIZ – Letteratura Italiana Zanichelli*, de Pasquale Stoppelli et Eugenio Picchi, dont la première édition date de 1993.
12. La récupération des matériaux codifiés dans des formats précédant le format DBT a été réalisée par Rosalba Cigliana et Valentina Pollidori. La première version du *corpus* est due au travail des informaticiens Eugenio Picchi et Elisabetta Marinai et de la chercheuse Valentina Pollidori, qui a géré le *corpus OVI* jusqu'à sa mort, en 2004. Pour la première phase des travaux qui ont conduit à la réalisation du *corpus*, voir Avalle (1979), De Robertis (1985) et Duro (1985).
13. La base de données consultable sur le site *Italnet* a été constituée par Theodore J. Cachey, Mark Olsen et Christian Dupont (en ligne : http://artfl-project.uchicago.edu/content/ovi). Cette base de données a eu le grand mérite de faire connaître le *corpus OVI* à la communauté scientifique – elle a été en effet beaucoup utilisée par les chercheurs –, mais est aujourd'hui obsolète, sa dernière mise à jour datant de 2005.
14. La lemmatisation du *corpus OVI* a été effectuée jusqu'en 2006 par Roberta Cella et ensuite par Elena Artale. Pour la méthode de lemmatisation, *se référer à* Esperti (1979). Elena Artale est en train d'accomplir la révision et la mise à jour des normes de lemmatisation : un premier résultat de ce travail, daté de 2013 et non publié, est réservé à l'usage interne.

Ce logiciel, outre qu'il présente de nouvelles fonctions de gestion des textes, constitue la tentative de répondre au piège de la « complexité » informatique contre lequel met en garde Robert Martin[15] : alors que, jusqu'à maintenant, le corpus OVI utilisait une évolution du « vieux » codage DBT, Gatto4 adopte le marquage international XML (format TEI), ce qui va rendre possible – dans un futur proche – l'interopérabilité des données du corpus au moyen d'un langage informatique largement partagé.

Quant au vocabulaire, une esquisse d'article lexicographique, qui anticipe le modèle adopté par le *TLIO*, fut élaborée en 1989[16] : l'on choisit – notamment et définitivement – d'utiliser de façon exclusive les ressources et le support informatiques, que ce soit pour la rédaction ou pour la publication du vocabulaire.

L'élaboration d'un modèle d'article pour le *TLIO*, tout comme la conception – ou peut-être la *vision* – de ce que devait être cette œuvre lexicographique (et de ce qu'elle est effectivement), après maintes années de théories et de réflexions préliminaires, date cependant des années 1990 et toutes deux sont dues à Pietro Beltrami, qui prit la direction de l'OVI en 1992 et la conserva jusqu'à 2013. Le premier article du *TLIO*, rédigé en 1996 par Beltrami lui-même, fut celui correspondant au lemme *abaco* ; en novembre 1996 fut publié en ligne un premier ensemble regroupant 122 articles, qui atteignit le seuil des 1 000 articles à la fin de 1998 : depuis lors, la rédaction progresse en moyenne de 2 000 articles par an.

Fidèle à sa conception originaire, le *TLIO* se compose d'articles rédigés à partir du dépouillement direct du corpus, qui est la

15. Je me réfère à la communication que Robert Martin a donnée à l'occasion de ce congrès : « Les réussites (et les pièges) de la lexicographie électronique ». Dans cette communication, il a souligné l'extrême complexité du langage informatique, qui rend souvent difficile le passage des données à de nouvelles versions du *software*, tout comme le passage de consignes entre l'informaticien qui a géré le logiciel jusqu'alors et l'informaticien qui va devoir en prendre la relève ; voir *supra*, p. 19-20.

16. Il s'agit de *COVIRED* : *se référer à* Ceccoli, Lorenzi et Pollidori (1989), *COVIRED : una procedura automatica come ausilio per la redazione del Tesoro della lingua italiana delle origini*, Firenze, Centro di Studi CNR Opera del vocabolario italiano – texte à usage interne, non publié mais refondu en partie dans Ceccoli, Lorenzi et Pollidori (1991), p. 67-84.

source principale (et en principe unique) de la documentation : le rédacteur, qui travaille seul[17], rédige chaque article après avoir lu et interprété toutes les occurrences du lexème présentes dans le corpus, qu'il peut repérer facilement puisque la lemmatisation fournit la liste des formes graphiques que revêt le lexème dans la base de données. Ce parti pris de rédaction implique une analyse *a posteriori* du matériel : en théorie, le rédacteur travaille en mettant de côté sa compétence de locuteur italien et en extrapolant les acceptions du lexème à partir de l'interprétation de chaque contexte[18]. Il organise ensuite ces acceptions dans une structure purement sémantique – les aspects grammaticaux, notamment, ne constituent pas un principe de structuration du matériel : s'il y a coïncidence sémantique, il est possible de réunir sous une même définition « cumulative », par exemple, l'adjectif et le substantif ou l'adverbe, ou – pour les verbes – l'emploi transitif et celui intransitif ou impersonnel.

Chaque article présente, en tête, des informations générales sur le lexème : la liste de ses formes graphiques présentes dans le corpus, l'étymon, la première attestation absolue et la

17. En 2005 a été instituée et confiée à Rossella Mosti la « prérédaction », une sorte d'étape intermédiaire entre la collecte des données résultant du *corpus* et la rédaction proprement dite : *voir* Mosti (2012). L'un des principaux résultats de cette activité a été la compilation de la liste des articles prévus pour le *TLIO*, qui a été complétée en 2014 (*cf. infra*).

18. Le fait que la rédaction soit effectuée *a posteriori* à partir du *corpus* ne signifie cependant pas qu'elle adopte une procédure inductive inspirée des méthodes de la linguistique statistique. Comme l'a si bien exprimé Diego Dotto : « *Qualsiasi proposta di descrizione di una varietà linguistica antica, una grammatica o un vocabolario, presuppone l'esistenza di un corpus. […] Di volta in volta la discriminazione tra ciò che non si trova [nel corpus] per una lacuna casuale e ciò che non si trova per una regola della lingua è un'operazione di estrema delicatezza. L'evidenza del principio si scontra infatti con la competenza inevitabilmente parziale di chi si propone di ordinare, descrivere o addirittura spiegare una costruzione o un lessema di una varietà linguistica antica, per la quale manca la possibilità di elicitare le regole grazie all'introspezione o all'interrogazione diretta di un parlante nativo. Il margine d'errore è sensibilmente più ampio. Un atteggiamento prudente è quindi giustificabile, ma la cautela non deve sfociare nella rinuncia programmatica a formulare ipotesi ; è vero invece che l'ipotesi deve essere evidente in sé e soprattutto sottoponibile in qualsiasi momento alla verifica su (nuovi) dati empirici o su (nuove) ipotesi che la potranno confermare o falsificare. Da altro punto di vista, importa rimarcare che la limitazione all'attestato senza alcuna distinzione tra frequenza e regola rimane essa stessa un'ipotesi al pari della ricostruzione, da valutare, di nuovo, caso per caso, alla luce della sua maggiore o minore plausibilità* » (Dotto 2012, p. 344-345).

« distribution géographique » du lexème, qui elle-même indique la première attestation pour chaque variété diatopique[19].

Le *TLIO* vise à décrire un état de langue, et non à constituer un glossaire des sources. Destiné à reconstruire l'architecture du lexique italien ancien, il se tient depuis toujours à la règle imposant de citer de préférence des textes documentaires plutôt que des œuvres littéraires : cela est dû à la conviction que ce genre de texte restitue un usage des mots autant que possible dénotatif et spontané, à l'opposé de l'usage littéraire. Le *TLIO* donne en outre la priorité à la citation d'exemples ayant valeur de glose, qui sont d'ailleurs marqués par un sigle spécifique dans le vocabulaire (Gl : « glose »).

Qui plus est, toujours parce qu'il vise à décrire un état de langue, le *TLIO* s'impose de « regarder au-delà de la donnée matérielle[20] », qui n'est qu'un reflet de ce que l'on cherche à voir et à décrire. Tout en ayant muni chaque texte d'une fiche bibliographique qui fournit les informations philologiques essentielles, et en attribuant une place privilégiée au témoignage des textes documentaires, pour lesquels l'œuvre coïncide souvent avec le témoin, le *TLIO* se fonde sur trois présomptions philologiques fondamentales : 1- que l'édition restitue l'œuvre ; 2- que la datation correspond à celle (connue ou présumée) de la composition de l'œuvre ; 3- que la localisation linguistique dépend, elle, de l'histoire de la tradition du texte. La localisation décrit donc la couleur linguistique du texte tel qu'il est restitué par l'édition, conformément à la distinction philologique classique entre critique des leçons et critique des formes.

Né dans les années 1960 italiennes et aujourd'hui encore fidèle à cette tradition et à cette école, l'OVI fonde sa mission scientifique sur l'idée que, comme l'a écrit Gianfranco Contini, la « réalité » du document ne constitue pas, en soi, une « vérité »

19. Pour une description sommaire d'un article du *TLI*, se référer à « Avvertenze per la consultazione », disponible en ligne sur la page : http://tlio.ovi.cnr.it/TLIO/.

20. Voir Beltrami (2010), p. 245 : « *per la datazione del lessico non si può cessare di guardare al di là del dato materiale* ».

textuelle[21]. Considérant qu'il est essentiel d'éviter tout fétichisme des données matérielles qui amènerait à faire de la leçon attestée un absolu, entre deux abstractions possibles – trouver le texte dans des éditions ou dans des manuscrits, étant tous deux des hypothèses de travail –, le *TLIO* a choisi de se fonder sur des éditions, tout en sachant qu'une édition n'exprime qu'une tentative d'approcher la vérité du texte, tentative qui ne saurait donner des résultats absolus ni définitifs. L'application de ce principe informe les données d'attestation du lexème et la gestion de la documentation. Dans les articles, les premières attestations qui dérivent de corrections ou de conjectures de la part de l'éditeur critique sont signalées en tant qu'interventions éditoriales, mais sont en principe acceptées. L'application de ce principe implique aussi le fait que le *corpus* n'inclut que le texte critique, et ne dépouille pas les apparats.

Aujourd'hui le *corpus OVI dell'italiano antico*, librement consultable sur la toile[22], recueille 2 318 textes, pour un total de 23 173 538 occurrences de 467 548 formes graphiques (*tokens*) ; dans le *corpus* sont présents 116 596 lemmes, pour un total de 3 654 946 occurrences lemmatisées[23].

Le *corpus OVI* constitue aujourd'hui une ressource essentielle pour toute étude qui s'intéresse à l'ancien italien ou aux textes italiens médiévaux, compte tenu de son ampleur et du prestige philologique dont il jouit – le nom de Domenico De Robertis en est, en quelque sorte, le garant. Le *corpus OVI* est de fait devenu dès sa publication sur le réseau un outil de travail fondamental pour tout historien de la langue italienne, bien que sa conception originelle ne le vouât qu'à servir de base à la rédaction du *TLIO* : la promotion du *corpus* à source autorisée, quasiment

21. Il s'agit d'une idée continienne bien connue : qu'il suffise ici de citer l'article « Filologia » de l'*Enciclopedia del Novecento*, 1977, Roma, Istituto della Enciclopedia Italiana fondata da Giovanni Treccani (en ligne : http://www.treccani.it/enciclopedia/filologia(Enciclopedia-del-Novecento)/), récemment republié avec un commentaire de Lino Leonardi (voir Contini, 2014).
22. En ligne : http://gattoweb.ovi.cnr.it.
23. Les données se réfèrent au *corpus* mis à jour le 5 décembre 2014. Le *corpus OVI* est géré depuis 2006 par Elena Artale et Pär Larson ; pour une description et un bilan récent, *voir* Artale et Larson (2012).

incontournable, de données linguistiques ou – plus encore – la considération du *corpus* comme véritable autorité en matière de données linguistiques ne sont pas exemptes de risques.

Un *corpus*, comme n'importe quel outil, est construit avec une finalité spécifique : sa qualité et sa valeur se mesurent par rapport à son efficacité à répondre au besoin pour lequel il a été conçu et réalisé. Or, le *corpus OVI* a été construit pour être la source d'un vocabulaire de l'ancien italien : il se porte garant, pour le dire ainsi, de la fiabilité lexicale de ses données (et, là encore, il requiert tout de même de ses usagers une certaine finesse d'interprétation). Dans la base de données sont présents des textes pratiques tout comme des textes littéraires, des éditions d'autographes comme des éditions de textes transmis par des copies (par un seul témoin ou par plusieurs témoins, manuscrits ou imprimés), des éditions critiques récentes comme des éditions du XVIII^e^ ou XIX^e^ siècle, des *testi di lingua* (c'est-à-dire des textes représentatifs d'une variété linguistique déterminée) et des textes qui témoignent d'une langue peu caractérisée du point de vue diatopique. En généralisant et en simplifiant beaucoup, il est possible de dire que l'on trouve dans le *corpus* trois situations ecdotiques typiques, qui sont restées telles quelles depuis les années 1960[24] :

1) la poésie lyrique, spécialement celle du XIII^e^ siècle, jouit depuis toujours d'une position de prestige et a donc profité de l'attention de l'avant-garde de la philologie italienne – qu'il suffise de rappeler ici la publication des *Poeti del Duecento* de Gianfranco Contini (1960) –, ce qui fait qu'elle se lit dans des éditions critiques souvent récentes et très soignées d'un point de vue ecdotique (« néolachmannisme » italien) ;

24. À l'époque, tout comme aujourd'hui, une documentation conséquente était disponible pour l'ancien italien, mais de qualité très différente et surtout éditée selon des critères et pour des finalités très différents. Giorgio Pasquali, en 1941, évaluant la disponibilité de textes en vue du projet d'un vocabulaire, écrivait : « *Certo, parecchi testi della nostra letteratura, perfino di quelli del periodo più antico di essa, sono, per nostra vergogna, inediti. Questo è un caso eccezionale ; molto più frequentemente non si possono leggere se non in edizioni insufficienti…* » (*voir* Pasquali, 1941).

2) pour ce qui concerne les statuts et les documents pratiques, l'on dispose de beaucoup d'éditions qui sortent de la grande école italienne d'histoire de la langue (les *Testi fiorentini del Dugento e dei primi del Trecento* de Alfredo Schiaffini datent de 1926 ; les *Nuovi testi fiorentini del Dugento* d'Arrigo Castellani datent de 1951-1952 ; les *Testi veneziani* de Alfredo Stussi datent de 1965) : les *testi di lingua* sont jugés comme des témoins fiables d'une variété particulière et diatopique, et leurs éditions constituent la base documentaire sur laquelle s'appuie la description linguistique (notamment phonétique) de cette variété. Les éditions sont scrupuleuses et très conservatives, allant chez l'école Castellani jusqu'à maintenir, dans certains cas, la segmentation des mots dans la chaîne graphique telle que la présente le manuscrit (qui est bien sûr, souvent, un original et un autographe) ;

3) à côté, pour ainsi dire, de ces deux traditions ecdotiques, scientifiques et contemporaines se trouvent beaucoup de textes notamment en prose (des textes littéraires, des traductions, des encyclopédies et des traités) : ces textes sont bel et bien édités, mais ils l'ont été en majeure partie pendant le XIX^e^ siècle[25]. À cette époque, en effet, l'on assiste en Italie à une véritable course à l'édition de textes du « bon siècle », c'est-à-dire du XIV^e^ : celle-ci est l'une des voies qu'emprunte l'engagement politique du *Risorgimento*, le combat nationaliste pour l'indépendance italienne. L'existence d'une langue « nationale » qui connut pendant le XIV^e^ siècle sa période de splendeur est l'un des mythes fondateurs de l'unité italienne, et c'est dans ce sens que beaucoup de textes furent édités par divers érudits et savants engagés. Bien sûr, les critères ecdotiques sont préscientifiques : habituellement ces éditions suivent le témoignage d'un

25. Parmi eux bien sûr figure l'édition établie par Michele Barbi de la *Vita nova* de Dante (révision parue en 1932 d'une première édition datant de 1907), qui est considérée comme le point de départ de la philologie italienne, mais cette édition est restée en quelque sorte un cas isolé. Dans les années 1950, la parution des *Volgarizzamenti del Due e Trecento* de Cesare Segre (1953) et de la *Prosa del Duecento* de Cesare Segre et Mario Marti (1959) était censée stimuler la production d'éditions critiques, qui cependant ne se développèrent pas vraiment.

manuscrit, toujours corrigé au nom de l'orthopédie formelle (orthographique, mais aussi morphologique et syntaxique) et parfois corrigé aussi sur le plan de la « pureté » textuelle – les termes triviaux, par exemple, sont quelquefois censurés. Or, ces éditions sont bel et bien présentes dans le *corpus OVI*, parce qu'en général elles restituent fidèlement – à l'exception près de quelques champs lexicaux[26] – les lexèmes attestés par le manuscrit édité, même si elles affichent des interventions très lourdes du point de vue de la forme.

Si cette variété de typologies ecdotiques, donc, n'affecte et n'affaiblit pas pour autant la valeur des données du point de vue lexical, qui est le seul qui intéresse la rédaction du vocabulaire, il faut être conscient qu'elle pourrait compter bien davantage si l'on s'intéressait à d'autres points de vue, et tout spécifiquement à l'histoire de la langue pour la phonétique ou la morphologie. Le repérage de lexèmes et de formes à partir d'une base textuelle informatisée induit facilement un nivellement a-philologique des données, alors qu'une interprétation historique et critique attentive est nécessaire, pour éviter tout malentendu. Pour guider les chercheurs dans la consultation du *corpus*, et tout particulièrement ceux qui s'intéressent à l'histoire de la langue du point de vue phonétique ou morphologique, les textes affichant une couleur linguistique déterminée sont marqués par le sigle « TS » (« texte significatif »), qui permet de construire des sous-*corpus* de textes bien caractérisés du point de vue de la diatopie linguistique.

26. Pour le lexique des *realia* dans les textes de traduction, voir p. ex. Guadagnini (2015) : « *La tendenza a restituire la terminologia "corretta" (dal punto di vista dell'intelligenza del testo latino), alterando le rese originarie attestate nelle traduzioni medievali, interessa generalmente il lessico che, all'interno delle testimonianze medievali, potremmo definire "storico", vale a dire tutti quei vocaboli che fanno riferimento a una realtà del passato non proseguita o radicalmente mutata nella contemporaneità degli scriventi : si tratta insomma di un lessico che è insieme erudito, poco attestato e tendenzialmente concentrato nei testi di traduzione, come appunto gli etnici ma anche i nomi di vesti, monete, misure, cariche pubbliche, ecc. Dal punto di vista lessicografico, per la documentazione di questa tipologia lessicale è presumibilmente assai rilevante l'effetto che ha sortito sui dati l'attività editoriale moderna.* »

Quant au vocabulaire, le *TLIO* compte aujourd'hui presque 30 000 articles publiés sur la toile[27] : grâce à la liste complète des lemmes récemment rédigée par Rossella Mosti, l'on prévoit que le vocabulaire complet en rassemblera environ le double[28]. Le segment A-F est quasiment complet, mais on compte déjà beaucoup d'articles rédigés pour les segments alphabétiques suivants, jusqu'à Z : dans les dernières années surtout, l'équipe des rédacteurs du *TLIO* a eu tendance à adopter une stratégie de rédaction non pas alphabétique, mais plutôt onomasiologique ou étymologique au sens large, choisissant par exemple de rédiger tous les articles qui partagent le même hyper-étymon latin ou le même préfixe.

Depuis le 1er octobre 2014, l'OVI a un nouveau directeur, Lino Leonardi : une nouvelle phase s'ouvre pour l'institut et pour ses projets.

Le progrès du *TLIO*, ou peut-être tout simplement son passage de la théorie préventive à la pratique, a signifié l'abandon de certains aspects de l'article lexicographique qui étaient prévus encore au début des années 2000, comme la description des constructions verbales ou le recueil des synonymies et des antonymies : l'OVI, confiant dans les possibilités de la « lexicographie évolutive », espère pouvoir compléter ses articles avec ces informations dans le futur.

Bien qu'il ait quelque peu simplifié ses objectifs, il reste vrai que le *TLIO* est toujours ouvert à de nouvelles acquisitions ou améliorations qui modifieraient les articles publiés en ligne. Pour déjouer le « piège de l'instabilité[29] », le *TLIO*, doté d'un numéro ISSN qui lui est propre en tant que publication périodique, associe à chaque article deux dates : la date de la rédaction de la première version publiée, qui se trouve au point 0.8 de l'en-tête et qui se réfère à l'entrée de l'article dans le *TLIO*, et la date de la

27. En ligne : http://tlio.ovi.cnr.it/tlio/. Il s'agit de 29 422 articles exactement (mise à jour du 17 février 2015). Pour un bilan assez récent du vocabulaire, on peut se reporter à Beltrami (2009) ou à Squillacioti (2012).
28. La liste complète des lemmes est consultable en ligne : http://reddyweb.ovi.cnr.it.
29. Je cite à nouveau les mots de Robert Martin.

dernière mise à jour, qui se trouve à la fin de l'article et identifie la version que l'on est en train de consulter.

Le progrès du *TLIO* emmenant l'approfondissement des connaissances lexicologiques, au sein de l'OVI a mûri le désir de développer certains aspects du travail : les normes de rédaction étant désormais fixées et stables, de nouvelles voies s'ouvrent pour la documentation.

Je mentionnerai brièvement deux projets actuellement en cours de réalisation et qui développent la base de données principale, le *corpus OVI*. Le plus ancien est le projet *DiVo* (Dizionario dei volgarizzamenti, « Dictionnaire des traductions vernaculaires[30] »). Né de l'expérience de rédaction du *TLIO*, le *DiVo* vise à analyser la langue des traductions médiévales des classiques et des ouvrages latins de l'antiquité tardive, qui constituent une partie majeure de la documentation des origines – cet aspect n'a toutefois pas été pris en compte par la lexicographie italienne des deux siècles derniers, bien qu'il ait été discuté dans les quatre premières éditions du *Vocabolario della Crusca*[31]. Comme l'a bien résumé Cosimo Burgassi, « le *DiVo* cherche [...] à mettre en lumière les relations linguistiques et, plus généralement, culturelles, qui sont établies entre le modèle classique et la tradition littéraire médiévale à travers la traduction. [...] Ce plan d'investigation [...] enregistre les connexions culturelles entre la romanité et le Moyen Âge par la lexicologie[32] ». Le projet s'appuie sur deux bases de données qu'il a lui-même créées : le *corpus DiVo*, qui recueille les traductions vernaculaires associées par paragraphes au texte latin traduit, et le *corpus CLaVo*, qui recueille les textes latins traduits associés par paragraphes à la traduction vernaculaire[33]. Financée par

30. Hébergé par l'OVI et la Scuola Normale Superiore de Pise, dirigé par Giulio Vaccaro et moi-même, le projet *DiVo* a officiellement pris naissance en mars 2012 et se développera pendant quatre ans. Pour une description du projet, voir Guadagnini et Vaccaro (2014) et Burgassi (2014).
31. Se référer à Guadagnini (2013).
32. Se référer à Burgassi (2014).
33. Le *corpus DiVo* (en ligne : http://divoweb.ovi.cnr.it) recueille exhaustivement les traductions médiévales de classiques latins – de Cicéron à Boèce, ce dernier étant considéré comme limite conventionnelle – et une sélection plutôt riche de traductions

l'État italien il y a cinq ans, l'équipe du projet a complété les bases de données et travaille actuellement à la deuxième phase du projet, qui se propose d'isoler les particularités de la langue des *volgarizzamenti* par rapport aux autres sous-codes de l'ancien italien[34].

Le second projet, *ReMediA* (*Repertorio di Medicina Antica*, « Répertoire de médecine ancienne »), dirigé par Ilaria Zamuner et Elena Artale, représente quant à lui la première tentative de création d'un *corpus* plurilingue à l'OVI. Le *corpus ReMediA*, dont une première version a été publiée sur la toile le 29 juillet 2014[35], est conçu pour restituer la spécificité de la terminologie technique et en particulier de la langue de la médecine et de la pharmacopée médiévale, dont Elena Artale et Ilaria Zamuner sont spécialistes : cette terminologie est à la fois un sous-ensemble assez clos du lexique vernaculaire et un domaine que l'on comprend mieux dès lors que l'on considère le contexte plurilingue dans une perspective comparative[36]. Le *corpus* va recueillir les principaux traités médicaux latins et romans, qui sont souvent des traductions d'un texte latin préexistant : la possibilité de comparer le latin et les diverses langues vernaculaires va permettre de mieux comprendre les textes du point de vue linguistique (lexical, bien sûr, mais aussi syntaxique) et également du point de vue historique. Le récent travail d'Ilaria Zamuner portant sur les résultats romans du lexème latin *aranea* (*tunica*)[37] fournit un premier exemple des résultats que permet d'obtenir l'application de cette méthode comparative.

---

d'ouvrages grecs (qui ont été traduits à partir d'un texte latin intermédiaire, tel l'*Éthique à Nicomaque* d'Aristote) et des pères de l'Église, composées en n'importe quelle variété de l'italien. À côté de ces traductions, le *corpus DiVo* regroupe également les éléments para-textuels (les gloses marginales et interlinéaires, les commentaires et les glossaires éventuellement associés à la traduction). Le *corpus CLaVo* (en ligne : http://clavoweb.ovi.cnr.it) recueille, comme on l'a exposé, les textes latins, ordonnés du plus ancien au plus récent suivant la chronologie des traductions vernaculaires associées. *Voir* Dotto (2013).

34. Pour les premiers résultats, voir Burgassi et Guadagnini (2014), Dotto (2015) et Guadagnini et Vaccaro (2011).
35. En ligne : http://remediaweb.ovi.cnr.it.
36. Voir par exemple Artale (2014). Pour une description synthétique du projet, consulter : http://www.sifr.it/ricerca/remedia.pdf.
37. Voir Zamuner (2015).

Ici se termine ce tour d'horizon de l'activité passée et présente de l'OVI, qui a permis d'évoquer quelques moments cruciaux et quelques aspects de ses travaux : l'activité de l'OVI est certes centrée sur la rédaction du *TLIO*, mais l'institut a aussi développé et développe, tout autour de son dictionnaire, des outils et des méthodes interconnectés, visant à approfondir toujours plus les connaissances déposées dans son œuvre majeure et les possibilités de recherche mises à la disposition des chercheurs. Pour cela, le pari sur l'informatique – qui, pris il y a un demi-siècle, témoigne d'un esprit véritablement pionnier – s'est révélé un choix des plus heureux.

Références bibliographiques

Artale, Elena et Larson, Pär, « Il punto sui corpora dell'Opera del Vocabolario Italiano », *Dizionari e ricerca filologica. Atti della Giornata di studi in memoria di Valentina Pollidori, Firenze, 26 ottobre 2010 (Supplemento III al Bollettino dell'Opera del Vocabolario Italiano)*, Alessandria, Edizioni dell'Orso, 2012, p. 25-40.

Artale, Elena, « Testi medici antichi e banche dati informatizzate. L'indicizzazione come risorsa ecdotica ed esegetica », dans Garavelli, Enrico et Suomela-Härmä, Elina (dir.), *Atti del XII Congresso SILFI (Helsinki, 18-20 giugno 2012)*, Firenze, Franco Cesati Editore, 2014, p. 43-50.

Avalle d'Arco Silvio, *Al servizio del vocabolario della lingua italiana*, Firenze, Accademia della Crusca, 1979.

Beltrami, Pietro G., « The Lexicography of Early Italian : Its Evolution and Recent Advances », dans Bruti, Silvia, Cella, Roberta et Foschi Albert, Marina (dir.), *Perspectives on Lexicography in Italy and Europe*, Newcastle upon Tyne, Cambridge Scholars Publishing, 2009, p. 27-53.

—, « Lessicografia e filologia in un dizionario storico dell'italiano antico », dans Ciociola, Claudio (dir.), *Storia della lingua italiana e filologia. Atti del VII Convegno ASLI (Pisa-Firenze, 18-20 dicembre 2008)*, Firenze, Cesati, 2010, p. 235-248.

Burgassi, Cosimo, « Le projet DiVo et ses corpus : une base de données italo-latine de traductions médiévales », *Bulletin du centre d'études médiévales d'Auxerre*, n° 18, 2014/1.
En ligne : http://cem.revues.org/13423.

Burgassi, Cosimo et Guadagnini, Elisa, « Prima dell' "indole", Latinismi latenti dell'italiano », *Studi di Lessicografia Italiana*, n° 31, 2014, p. 1-39.

Ceccoli, A., Lorenzi, F. et Pollidori, V., « Un programma per la redazione del Vocabolario Storico della Lingua Italiana assistita dal calcolatore », dans *Récit et informatique. Actes de la journée d'études, C.R.L.L.I., Université de Paris X - Nanterre, 9 décembre 1989*, Claude Cazalé Bérard (éd.), La Garenne-Colombes, Éditions de l'Espace européen, 1991, p. 85-106.

Contini, Gianfranco, *Filologia*, Bologna, Il Mulino, 2014.

De Robertis, Domenico, « L'ufficio filologico dell'Opera del vocabolario, il suo impianto, il suo lavoro », dans Alfieri, Gabriella *et al.*, *La Crusca nella tradizione letteraria e linguistica italiana*, Firenze, Accademia della Crusca, 1985, p. 443-451.

Dotto, Diego, « Note per la lemmatizzazione del corpus DiVo », *Bollettino dell'Opera del vocabolario italiano*, n° 17, 2012, p. 336-364.

—, « Notizie dal DiVo. Un primo bilancio sulla costituzione del corpus », dans Larson, Pär, Squillacioti, Paolo et Vaccaro Giulio (dir.), « *Diverse voci fanno dolci note », L'Opera del vocabolario italiano per Pietro G. Beltrami*, Alessandria, Edizioni dell'Orso, 2013, p. 71-83.

—, « Esercizi sul contributo del lessico di traduzione in lessicografia : dal TLIO al DiVo », dans Buchi, Éva, Chauveau, Jean-Paul et Pierrel, Jean-Marie (dir.), *Actes du XXVII^e^ Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013)*, Strasbourg, Société de linguistique romane/ ÉliPhi, 2015.

Duro, Aldo, « L'impianto del nuovo vocabolario: profilo storico », dans Alfieri, Gabriella *et al.*, *La Crusca nella tradizione letteraria*

e linguistica italiana, Firenze, Accademia della Crusca, 1985, p. 431-442.

ESPERTI Piero, « Grammatichetta della lingua italiana ad uso del calcolatore », dans AVALLE, d'Arco Silvio (dir.), *Al servizio del vocabolario della lingua italiana*, Firenze, Accademia della Crusca, 1979, p. 123-187.

GUADAGNINI, Elisa, « Notizie dal DiVo. Parole tradotte e lessicografia dell'italiano », dans LARSON, Pär, SQUILLACIOTI, Paolo et VACCARO, Giulio (dir.), « *Diverse voci fanno dolci note », L'Opera del vocabolario italiano per Pietro G. Beltrami*, Alessandria, Edizioni dell'Orso, 2013, p. 59-70.

—, « Variazioni aborigene : note di lessicografia dell'italiano antico », *Bollettino dell'Opera del vocabolario italiano*, 2015.

GUADAGNINI, Elisa et VACCARO, Guilio, « "*Nom de pays : le nom…*" Parole, paesi e popoli nel corpus DiVo », dans LUBELLO, Sergio (dir.), *Volgarizzare, tradurre, interpretare nei secc. XIII-XVI. Atti del Convegno internazionale di studio : Studio, archivio e lessico dei volgarizzamenti italiani (Salerno, 24-25 novembre 2010)*, Strasbourg, Éditions de linguistique et de philologie, 2011, p. 267-281.

GUADAGNINI, Elisa et VACCARO, Guilio, « Un contributo allo studio del "volgarizzare e tradurre" : il progetto DiVo », dans PACCAGNELLA, Ivano et GREGORI, Elisa (dir.), *Lingue, testi, culture : L'eredità di Folena, vent'anni dopo. Atti del XL Convegno Interuniversitario (Bressanone, 12-15 luglio 2012)*, Padova, Esedra editrice, 2014, p. 91-105.

MOSTI, Rossella, « Tra lemma e voce: ruolo della preredazione nel *Tesoro della lingua italiana delle origini* », *Dizionari e ricerca filologica. Atti della Giornata di studi in memoria di Valentina Pollidori, Firenze, 26 ottobre 2010 (Supplemento III al Bollettino dell'Opera del Vocabolario Italiano)*, Alessandria, Edizioni dell'Orso, 2012, p. 85-99.

PASQUALI, Giorgio, « Per un tesoro della lingua italiana », *Atti della R. Accademia d'Italia. Rendiconti della Classe di scienze morali e storiche*, s. 7, II, 1941, p. 490-521.

Squillacioti Paolo, « Uno sguardo al *Tesoro della Lingua Italiana delle Origini*: procedure e prospettive del vocabolario storico dell'italiano antico », *Dizionari e ricerca filologica. Atti della Giornata di studi in memoria di Valentina Pollidori, Firenze, 26 ottobre 2010 (Supplemento III al Bollettino dell'Opera del Vocabolario Italiano)*, Alessandria, Edizioni dell'Orso, 2012, p. 74-84.

Tomasin, Lorenzo, « Qu'est-ce que l'italien ancien ? », *La Lingua Italiana*, n° 9, 2013, p. 1-18.

Vaccaro, Giulio, « Veniamo da molto lontano e andiamo molto lontano. Documenti per la storia dell'Opera del Vocabolario Italiano dalle origini al 1992 », *Bollettino dell'Opera del Vocabolario Italiano*, n° 18, 2013, p. 277-390.

Zamuner, Ilaria, « *'Aranea' tunica* e la lessicografia medico-scientifica romanza », *Cultura Neolatina*, n° 75, 2015/1.

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